University of Nottingham
UK | CHINA | MALAYSIA

Testing a Connectionist Model of Acquired Equivalence

Sara Bru Garcia

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the School of Psychology – University of Nottingham

2020

Abstract

Over the past decades, experimental research with animals has demonstrated that the generalisation between two stimuli is determined not only by their intrinsic properties, but by their associative history. This phenomenon is illustrated with the use of acquired equivalence tasks, which show that stimuli are treated as more similar when they come to elicit the same response as a result of conditioning. Different associative learning theories have been proposed to accommodate extant experimental findings. Mediated conditioning can explain simple forms of acquired equivalence (Honey & Hall, 1989), but is unable to accommodate findings from more complex configural acquired equivalence tasks, where stimuli are equally reinforced and nonreinforced. Pearce's (1994) connectionist model and its extended version (Honey & Watt, 1998) can explain findings from revaluation configural acquired equivalence procedures, but cannot anticipate findings from other forms of configural acquired equivalence. Alternatively, Honey and colleagues (Honey, 2000; Honey et al., 2010) proposed a connectionist model that allows for similar inputs that share a common reinforcer to share hidden units. This model was able to accommodate a wide range of experimental findings that other models failed to explain.

Honey and colleagues claimed that their connectionist model could also accommodate the results from Intra-dimensional/Extra-dimensional shift tasks (IDS/EDS), which consistently find that IDS is easier than EDS, without explicitly invoking the need for attention. In Chapter 2, we tested this claim by assessing the correlation between performance in a configural acquired equivalence task and two attentional set tasks: IDS/EDS and optional-shift. Findings revealed an overall positive correlation between test performance in acquired equivalence and optional-

shift, but no correlation between performance in our acquired equivalence task and IDS/EDS, in what could be seen as a challenge to Honey et al. (2010).

Chapter 3 tested the effects of various outcome manipulations in configural and non-configural acquired equivalence. Experiments in this chapter revealed an enhanced revaluation and acquired equivalence effect in participants who had experienced different outcomes across training and revaluation, compared to participants who had received the same outcomes across stages. However, these group differences disappeared in a second experiment that intermixed configural and non-configural trials during the initial discrimination. A second set of experiments in this chapter failed to replicate findings from Delamater (1998), which reported a faster reversal acquisition in a group of rats that received different outcomes within stimulus modality compared to a group of rats that received the same outcomes within stimulus modality.

Although Honey and colleagues carefully described the characteristics of their network, verbal descriptions could be prone to error. To the aim of qualifying the model, Chapter 4 describes a series of simulations of the experimental data presented in chapters 2 and 3 of this thesis using a formal computer instantiation of Honey's model that was recently published (Robinson et al., 2019). This Hebbian learning network successfully simulated data from our 2-Stages configural acquired equivalence task and confirmed an enhanced acquired equivalence effect in a simulation with different outcomes across training and revaluation. This instantiation of the model was also able to accommodate findings from Delamater (1998), despite our unsuccessful attempts to replicate and extend the generality of the findings to human participants.

Summary of Experiments

Following Honey et al.'s (2010) suggestion of a common mechanism underlying performance in acquired equivalence and attentional set, Experiment 1 in Chapter 2 investigated whether performance following revaluation in a configural acquired equivalence task correlated with performance in two attentional set tasks: IDS/EDS and optional-shift. Participants demonstrated the acquired equivalence effect and showed the anticipated IDS superiority when assessed with the Cambridge Neuropsychological Test Automated Battery task (CANTAB), but not with our pilot intra/extra dimensional set shifting task. Results revealed that test performance in these two tasks was not correlated, with substantial Bayesian support in favour of the null model. Experiment 2 incorporated an optional-shift task based on the experimental design used by Duffaud et al. (2007). This optional-shift task was matched in stimuli, number of trials, experimental design and way of administration as closely as possible to the configural acquired equivalence task to allow for a meaningful comparison. Results revealed a correlation between performance at test in both tasks, with substantial Bayesian support for a positive correlation. Experiment 3 replicated this finding from Experiment 2, revealing a positive correlation between test performance in our configural acquired equivalence and optional-shift tasks, with strong Bayesian support for the alternative model. It also revealed that participants whose eye-gaze was directed to the predictive, over the non-predictive elements of the discrimination during training, demonstrated a more pronounced attentional set difference at test. This observation adds to the existing body of evidence suggesting a preference for the predictive elements of a discrimination (e.g., Haselgrove et al., 2016; Le Pelley et al., 2011) even when outcomes are unchanged across stages. Experiment 4 was conducted as a final

replication of the findings of Experiments 2 and 3, and incorporated a control *N*-back task that sought to control for any non-specific effects in the correlation. Contrary to our expectations, results yielded no correlation in performance between the tasks. However, an overall Bayesian correlation (n = 96) provided strong support for a positive correlation between test performance in both tasks, offering overall partial support for Honey et al.'s (2010) claims.

Chapter 3, tested the effects of outcome manipulations in different forms of acquired equivalence. Experiment 5 investigated differences in performance resulting from presenting either the same or different outcomes across stages in our configural acquired equivalence task. Results showed an enhanced performance during revaluation and test trials in the group that had received a different set of differential outcomes, compared to the group that received the same set of differential outcomes across stages. However, participants in the group with the same outcomes failed to demonstrate the anticipated acquired equivalence. Experiment 6 sought to account for possible effects non-specific to the outcome manipulation in performance (e.g., arousal). The task incorporated non-configural trials during training, which allowed for all possible outcomes to be present from the onset of the task. However, results in this experiment could not be interpreted due to an experimental confound. Experiment 7 rectified and replicated Experiment 6. Participants in group *Same*, which experienced the same stimulus-outcome contingencies across stages, and group *Different*, which experienced different stimulus-outcomes contingencies, showed acquired equivalence. However, no group differences in performance were found.

Experiments 8, 9, and 10 in Chapter 3 investigated whether a reversal stage with different outcomes within stimulus dimension resulted in an enhanced

discrimination learning compared to a reversal stage with the same outcomes within stimulus dimension, as findings in Delamater (1998) suggest. In Experiment 8, participants learned a simple discrimination between stimuli that belonged to two distinct visual stimulus dimensions. In a subsequent stage, performance between a group of participants that received a series of reversals with the same outcomes within stimulus dimension and a second group that received reversals with different outcomes within stimulus dimension was compared. Experiment 9 rectified the counterbalancing of the previous experiment and attempted to make the task more challenging by increasing the number of exemplars presented from four to 12. Experiment 10 was identical to Experiment 9, but attempted a more direct replication of the findings in Delamater (1998) by using stimuli from auditory and visual modalities. These three experiments were unable to replicate the differences in performance between the two groups reported in Delamater (1998).

Chapter 4 presented a series of simulations of the experiments conducted in this thesis to evaluate and help qualify a formal implementation of the 3-layered connectionist network verbally described by Honey and colleagues (Robinson et al., 2019). Simulated data captured the results observed in the configural acquired equivalence tasks used in this thesis, and offered computational support for the enhanced acquired equivalence observed in the group that experienced different outcomes across stages in Experiment 5. The simulations also offered valuable insights about the behaviour of this implementation of the model once non-configural inputs were added. Although our experiments in Chapter 3 failed to find group differences in reversal performance, the network was able to accommodate Delamater's (1998) findings and simulate faster reversal acquisition when different outcomes within modality were used. Overall, this chapter helped qualify the current

instantiation of Honey's network by increasing its generality and offering suggestions for improvements.

Abbreviations

CS: Conditioned Stimulus

US: Unconditioned Stimulus

CANTAB: Cambridge Neuropsychological Test Automated Battery

IDS/EDS: Intra-Dimensional Shift/Extra-Dimensional Shift

p-IDS/EDS: Pilot Intra-Dimensional Shift/Extra-Dimensional Shift

DOE: Differential Outcome Effect

ROI: Region of Interest

CPCA: Conditional Principal Component Analysis

Acknowledgments

I could not have asked for a better supervisor than Dr Jasper Robinson. Thank you for your generous support, guidance and motivation during these past years. You made this process significantly more bearable.

Many thanks to my second supervisor, Dr Ruth Filik, and to my internal examiner, Dr Mark Haselgrove, for the insightful comments during my end of year review. Thank you to my external examiner, Dr Gonzalo Urcelay, for reviewing my thesis. I am also thankful to my funding bodies, the Vice Chancellor's Scholarship for Research Excellence and the School of Psychology at The University of Nottingham. Thank you to the associative learning group for the interesting talks and discussions and to my office/Uni mates for the intellectual and not so intellectual conversations.

Eternal gratitude to my family and friends (Javi, both Gabis, Emma, Izzy, Jamie, Melody…). Very special mention to Melanie and Isa for their relentless support and friendship, no matter how far they may be. And finally, thank you to Ólafur Arnalds, for without his music I am not sure I would have been able to write this thesis.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: GENERAL INTRODUCTION

CHAPTER 2: DISSOCIATION OF TWO MEASURES OF ATTENTIONAL SET WITH CONFIGURAL ACQUIRED EQUIVALENCE

CHAPTER 4: A HEBBIAN LEARNING NETWORK: SIMULATING EMPIRICAL EVIDENCE OF OUTCOME MANIPULATIONS

Chapter 1:

General introduction

A conditioned response that an animal has been trained to perform upon the presentation of a given stimulus can also be elicited, to some extent, by a second stimulus; a phenomenon known as stimulus generalisation. Traditional interpretations of this phenomenon have assumed that the presentation of any given stimulus will excite a number of representational elements. Some of these elements will be uniquely activated by the trained stimulus itself (unique elements), whilst others (common elements) will be activated by a range of stimuli (Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981). For example, a bright white light, a bright yellow light, and a dim red light will each be represented by a finite number of representational elements, some of which will overlap (e.g., the brightness of both the white and yellow lights). Once a conditioned response has been established to one stimulus (e.g., the bright white light), stimuli that are, or animals perceive to be, more similar (e.g., the bright yellow light) will have a greater number of shared representational elements and will allow for greater generalisation. That is, they will elicit a stronger conditioned response. Stimuli that are subjectively or physically more dissimilar (e.g., the dim red light) will have fewer elements in common and will afford a weaker stimulus generalisation. That is, they will elicit a weaker conditioned response.

However, this standard view of stimulus generalisation is challenged by the finding that generalisation between two stimuli is determined not only by their intrinsic properties, but by their associative history. Support for this suggestion comes from experiments on *acquired equivalence*, which demonstrate that generalisation between two stimuli, even if apparently dissimilar, can occur if these stimuli have come to elicit the same response as a result of initial conditioning (Delamater, 1998; Hall et al., 1993; Hodder et al., 2003; Honey & Hall, 1989; Honey

& Ward-Robinson, 2001; Honey & Watt, 1998; Iordanova et al., 2007; Lawrence, 1950; Miller & Dollard, 1941; Robinson & Owens, 2013; Ward-Robinson & Hall, 1999; Ward-Robinson & Honey, 2000). Subsequently, theories of associative learning have begun to surpass the more traditional notions of stimulus generalisation which do not afford the fact that stimuli are treated as more similar when they share a common training history, compared to stimuli that have been trained to signal different associates.

A mediated conditioning interpretation of acquired equivalence (Honey & Hall, 1989) proposes that the associatively activated stimulus representations that stimuli come to activate as a result of conditioning can, in themselves, enter into further associations. The generalisation observed between two stimuli that share a training history will be mediated by the conditioned properties of their common associate. However, more complex acquired equivalence procedures illustrate the inadequacy of mediated conditioning as a sole mechanism to explain acquired equivalence.

In this first chapter I will explain different acquired equivalence procedures, starting from the simplest form building to the more complex configural acquired equivalence tasks, and discuss to what extent various theories can accommodate extant experimental findings. The focus of the remainder chapters will be on acquired equivalence and a 3-layered connectionist network (Honey, 2000; Honey et al., 2010; Honey & Ward-Robinson, 2002). I first outline my tests of specific claims derived from Honey et al. (2010). I then present different outcome manipulations in configural and non-configural acquired equivalence tasks and interpret them from the perspective of a network of the characteristics described by Honey and colleagues. Finally, I present simulations from a formal instantiation of the

connectionist learning network (Robinson et al., 2019) to evaluate and qualify the model.

1.1 Mediated conditioning and acquired equivalence

In the earliest mention to the notion of acquired equivalence, Miller and Dollard (1941) argued that stimuli might become equivalent, even if dissimilar in appearance, when they have been trained to signal the same response. The suggestion that generalisation is not only driven by the intrinsic properties of the stimuli, but also mediated by their shared associative history, prompted a body of experimental work that tended to offer support in animal and human subjects (e.g., Grice & Davis, 1958; Lawrence, 1949). However, this early work produced results that were open to interpretations that did not involve changes in stimulus discriminability as a result of a common training history (see Honey & Hall 1989).

Honey and Hall (1989) produced clearer evidence of the consequences of a common training history. In experiment 3 reported by Honey and Hall (1989), summarised in the upper half of **Table 1**, rats in two separate groups received an auditory discrimination of the form A-, B+, N- or A+, B-, N+, where "+" and "-" indicated access to food and no food, respectively. The key to this design was not whether stimuli had been reinforced during training, but whether they had a shared training history or not. After pairing N with a shock in a second stage, conditioned suppression to A and B was measured. Results showed that rats generalised more readily between A and N, which had signalled the same consequence during training, than between B and N, which had signalled different consequences. That is, rats showed a stronger conditioned suppression when presented with A compared to B after having revalued N with a mild shock. Honey and Hall (1989) interpreted these results in line with Miller and Dollard's (1941) notion of acquired equivalence as follows. In group A+, B-, N+, stimulus A and N would both form an association with food as a consequence of their equivalent training. The presentation of N during the

second stage would activate that mental representation of food. When the footshock occurred in a second stage, this associatively active representation of food would enter into further association with the shock and acquire the capacity to elicit the conditioned response to some extent. Because the presentation of A would also evoke the representation of food at test, and that representation of food had acquired aversive properties, A would also elicit the conditioned response, which resulted in the observed reluctance to consume food. On the other hand, the representation of no food was never associated with shock, therefore B would not elicit the conditioned response. Because what drives generalisation is the matched treatment given to A and N, and not whether they have been reinforced or not, the same interpretation would apply to group A-, B+, N-.

Table 1

Experimental designs for Honey and Hall (1989) and Holland (1981)

Stage 1	Stage 2	Test
a. Honey and Hall (1989)		
A + B – N +	N – Shock	**A ?** **B ?**
A – B + N –		
b. Holland (1981)		
T + L –	T – LiCl L – T – L – LiCl	**Food consumption**

Note. (a) A, B and N represent a tone, clicker and white noise, respectively. (b) T and L represent a tone and a light, respectively. LiCl indicates an injection of lithium chloride. + and – indicate food pellets and no food.

This *mediated conditioning* interpretation of the acquired equivalence effect assumes that the associatively activated representation of an event can substitute for the event itself in the formation of new associations involving that event. This assumption could be evidenced by experimental data on sensory preconditioning (e.g., Rizley & Rescorla, 1972). In a sensory preconditioning task, two neutral stimuli are initially paired (A → B). In a subsequent stage, a conditioned response is established by pairing one of them with a reinforcer (B+). Finally, responding to the other stimulus (A), which has never been paired with the reinforcer, is measured. Conditioned responding to A could be interpreted in terms of the formation of an association between A and B during the initial training. When the second stage establishes the B+ association, stimulus A is also capable of producing a conditioned response by virtue of the association between its evoked mental representation and the reinforcer (i.e., after training, the presentation of stimulus B evokes the mental representation of stimulus A which, in turn, becomes associated with the food). However, sensory preconditioning is amenable to other interpretations and cannot be taken alone as sufficient evidence for mediated conditioning. Sensory preconditioning tasks could instead reflect the formation of an associative chain. In the aforementioned example, training would result in the formation of an excitatory association between A and B. Subsequently, B would be associated with the reinforcer (+) and produce a conditioned response. Therefore, on test, A would also able to activate the representation of the reinforcer and produce the appropriate conditioned response by dint of the A → B → + associative chain.

More robust experimental evidence for the role of mediated conditioning comes from modified sensory preconditioning procedures. Holland (1981) conducted a series of experiments with stimuli following a backwards sensory preconditioning

sequence. In one of his experiments, different groups of rats received a simple audio-visual discrimination. For one subgroup of rats, a tone – but not a light – was followed by food. Following this training, the tone was subsequently paired with lithium chloride (LiCl), an illness inducing chemical. At test, the consumption of food pellets was measured and compared against food consumption in a second subgroup of rats, for which the light – and not the tone – signalled LiCl, as summarised in the bottom half of **Table 1**. Rats in the former group consumed fewer food pellets than rats in the latter group. Holland interpreted these results as evidence for mediated conditioning. For the first group of rats, the presentation of the tone was able to activate the representation of food pellets, which became associated with illness after the injection of LiCl in the second stage, resulting in a decreased food consumption. For the second subgroup of rats, the mental representation of food was never associated with illness, which resulted in an increased food consumption at test. Holland also offered direct evidence against an alternative interpretation. That is, that a backwards excitatory association resulting in a food → tone → illness – and a food → tone → no illness – associative chain. A third subgroup of rats received the delivery of food followed by the presentation of a tone during training, before having the tone paired with LiCl in a subsequent stage. That is, the standard forward sensory preconditioning task. The treatment of this subgroup would have encouraged the formation of an explicit food → tone association during training which, once LiCl was administered after the tone, should have produced aversion to the food. However, rats in this subgroup showed no aversion to the food at test, suggesting that no food → tone → illness association chain had formed.

The mediated conditioning interpretation of acquired equivalence received further support from Ward-Robinson and Hall (1999), who replicated Honey and

Hall's (1989) experiment and included an explicit test of mediated conditioning as follows. Two groups of rats received a discrimination identical to that of Honey and Hall's (1989) paper (i.e., A+, B-, N+/A-, B+, N-) followed by the revaluation of N with a shock. After the acquired equivalence test, Ward-Robinson and Hall gave animals a further test in which they were allowed to press a lever in order to retrieve food pellets. This stage was included as an explicit test of the fact that the associatively active representation of food should have acquired aversive properties in group A+, B-, N+ but not in group A-, B+, N-. Results replicated the acquired equivalence effect, that is, more generalised responding between A and N compared to B and N, but, importantly, they showed that lever press response latencies were reliably higher in group A+, B-, N+. This finding is consistent with the idea that for this group, the mental representation of food had acquired aversive properties through the pairing of N with a shock.

Mediated conditioning has proven to be a successful mechanism to explain generalised conditioned responding in simple, non-configural acquired equivalence tasks (Bonardi et al., 1993; Honey & Hall, 1989; Ward-Robinson & Hall, 1999). However, other experimental procedures are inexplicable in terms of mediated conditioning (e.g., Coutureau et al., 2002; Delamater, 1998; Honey & Watt, 1998; Iordanova et al., 2007; Robinson & Owens, 2013; Ward-Robinson & Honey, 2000). Take as an example the generalisation of conditioned responding observed after revaluation in a configural acquired equivalence task, central to this thesis. After establishing equivalence relationships in an appetitive configural discrimination of the form Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+, stimuli are revalued in a subsequent stage by, for example, pairing A with a mild footshock and B with no shock. At test, findings show greater conditioned fear to stimulus C than to stimulus

D. A mediated conditioning analysis is inadequate to explain this generalisation. If we assume that stimulus A evokes, for example, the representation of food during the revaluation stage, this would require stimulus C to also evoke the representation of food during revaluation. However, because stimuli A, B, C and D were equally likely to signal food or no food in the absence of w and x during training, it is unclear why A should evoke the representation of food any more than the representation of no food. Even if A happened to elicit the mental representation of food, and this representation entered into further association with shock during revaluation, the initial configural training would render stimulus C and D equally likely to activate this now aversive representation of food, and we should not observe reliable greater conditioned fear to C than to D.

1.2 A brief mention to elemental theories: Rescorla and Wagner (1972)

Associative learning theories are based on the assumption that the repeated pairing of two stimuli will result in the development of a connection, or association, between their internal representations. However, theories have differed in their assumptions when compound stimuli are presented for conditioning.

Elemental theories (e.g., Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981) have in common the assumption that when two or more stimuli are presented as a compound for conditioning, each separate element will have the opportunity to enter into an association with the outcome. Take as an example the model produced by Rescorla and Wagner (1972), arguably the most influential model of associative learning. The elemental nature of this model is evidenced by the fact that according to Rescorla and Wagner, repeated presentations of a compound stimulus (CS) and an outcome (US) will result in a change in the strength of the connection between the internal representations of each element of the CS and the US in accordance to **Equation 1.1**.

$$\Delta V = \alpha\beta(\lambda - V_T) \qquad \textbf{Equation 1.1}$$

In this equation, the change in the associative strength ($\triangle V$) between a given stimulus and an outcome is determined by the discrepancy between the asymptote of conditioning supported by the outcome (λ) and the combined associative strength of all stimuli present on that trial (V_T). The magnitude of this change is influenced by

two learning rate parameters – with fixed values between 0 and 1 – corresponding to the unconditioned salience of the stimulus (α) and the outcome (β).

The Rescorla-Wagner (1972) model is successful in explaining phenomena that are crucial to associative learning such as blocking. In brief, blocking refers to the finding that learning about the relationship between a stimulus (e.g., B) and an outcome (e.g., +) will be impaired if that stimulus is presented in conjunction with a second stimulus (A) that has previously been trained to signal the outcome. That is, establishing stimulus A+ as a reliable predictor of the outcome during training will block subsequent learning about A**B**+. Learning about stimulus B, however, will occur if the training stage is omitted (Kamin, 1968). The model developed by Rescorla and Wagner can explain these findings because by the end of training, the association between stimulus A and the outcome will be at asymptote. That is, stimulus A will be a perfect predictor of the outcome. Therefore, when stimulus B is subsequently presented in compound with A, there will be no discrepancy between the maximum level of conditioning supported by the outcome and the associative strength of all cues present $(\lambda - V_T)$ and no learning will occur.

However, elemental theories like the Rescorla-Wagner (1972) model are unable to explain how animals solve certain discriminations central to this thesis. For example, configural discriminations of the form Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+ are a challenge to the Rescorla-Wagner model because the capacity of a compound to elicit a response simply reflects the combined associative strength of the elements that constitute that compound. Because the associative strength of the elements presented in a trial is identical to that of the elements presented in any other trial, there are no grounds to solve the discrimination. Aware of this issue, Rescorla and Wagner proposed that whenever a compound of two or more elements is

presented, a configural representation could develop and take part in conditioning, which would equip it to explain how a configural discrimination is acquired. However, the finding that generalisation of conditioned responding occurs after revaluation in a configural acquired equivalence task constitutes a challenge to even this modified Rescorla-Wagner model, which has no scope for the prior associative history of stimuli to influence subsequent learning.

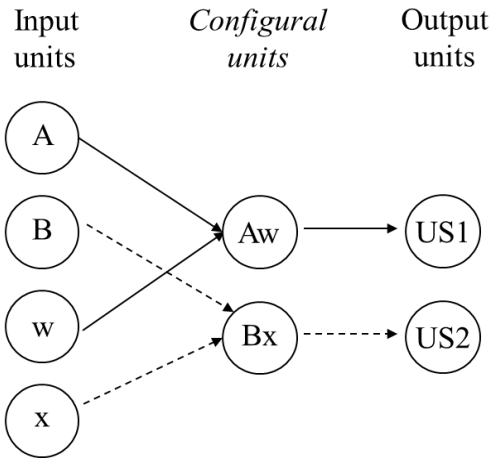
1.3 Pearce (1987, 1994) configural theory and acquired equivalence

Mediated conditioning can successfully account for simple forms of acquired equivalence. However, as outlined earlier, mediated conditioning cannot anticipate the findings of more complex configural acquired equivalence tasks. Here, I discuss the influential configural theory of Pearce (Pearce, 1987, 1994) and its interpretation of configural acquired equivalence.

In contrast to elemental accounts of leaning (e.g., Pearce & Hall, 1980; Wagner, 1981; Rescorla & Wagner, 1972), which assume that the capacity of a compound stimulus to elicit responding reflects the combined associative strength of each of the elements presented in any one trial, configural accounts propose that the combined presentation of two stimuli results in a unique, or configural, representation that enters into association with the outcome of a particular trial (e.g., food or the absence of food). If the same pattern of stimulation is presented in a later trial, it will activate the previously formed configural representation fully. If a different pattern of stimulation is presented, it will activate the configural representation to an extent that is directly determined by the similarity between the two patterns. Pearce (1994), proposed a 3-layered connectionist network that incorporated a layer of configural, or hidden, units between the input and output layers, as illustrated in **Figure 1**, to formalise the essence of his early theory (Pearce, 1987).

Figure 1

Pearce (1994) Connectionist Network



Note. For the sake of clarity only trials Aw+ and Bx- are represented.

Pearce (1994) assumed that any pattern of stimulation would result in activation to the corresponding input units. For example, when pattern Aw is presented, input units A and w will be excited from their resting value (0) to their activated value (1), and they will quickly become connected and activate a single configural unit (*Aw*) to its maximal level. Once an outcome occurs (e.g., food), a connection will also develop between the maximally activated configural unit and the output unit. The strength in the connection between the *Aw* configural unit and the output unit, $V_{Aw,}$ will develop gradually as described in **Equation 2.1**, where λ is the asymptote of conditioning – set at 1 – and β is a learning rate parameter – set between 0 and 1 – that is determined by the properties of the reinforcer.

$$\Delta V_{Aw} = \beta(\lambda - V_{Aw}) \qquad \textbf{Equation 2.1}$$

This connection between the configural unit and the output unit will lead to a conditioned response. According to Pearce (1994), the strength of the conditioned response that will follow after the presentation of the US will be determined by the level of activation to the configural unit *Aw* multiplied by the strength of the connection between *Aw* and the output unit. Thus, if pattern Aw is presented again, the configural unit *Aw* will be activated maximally and a strong conditioned response will follow. If only one stimulus is presented, for instance A, the excitation to configural unit *Aw* and the activation to the corresponding output unit will be less, and the subsequent conditioned response will also be reduced. The associative strength of pattern Aw will generalise to pattern A in accordance with **Equation 2.2**. Here, the strength of the conditioned response to stimulus A (E_A) is determined by the similarity between A and Aw ($_AS_{Aw}$) multiplied by the level of activation of configural unit *Aw* (V_{Aw}). Similarity between A and Aw, which is directly related to the proportion of common elements shared by the two patterns and the elements unique to each pattern, is derived from **Equation 2.3**. Here, N_C represent the number of input units common to both patterns (one, in this example), and N_A and N_{Aw} are the input units activated by each pattern. Whenever a pattern of stimulation fails to excite any existing configural unit to its maximum level, it will be regarded as novel, and eventually a new configural unit will be recruited.

$$E_A = {}_AS_{Aw} * V_{Aw} \qquad \textbf{Equation 2.2}$$

$$_AS_{Aw} = N_C/N_A * N_C/N_{Aw} \qquad \textbf{Equation 2.3}$$

Pearce's (1994) model provides an account for a broad range of findings that elemental theories are unable to accommodate (see Pearce, 2002). However, it is ill-equipped to explain configural acquired equivalence. Consider once again a discrimination of the form Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+, where A and B are later revalued with a mild footshock and no shock, respectively. Pearce's (1994) connectionist model can readily anticipate the acquisition of the initial biconditional discrimination. According to this model, the biconditional training will result in the development of connections between the configural representations of Aw, Bx, Cw, and Dx and the representation of food. Conversely, the configural representations of Ax, Bw, Cx, and Dw will become associated with the representation of no food (see **Figure 1**). However, there are no grounds to anticipate an increased generalisation between A and C – or B and D – as a result of their common training history. On the one hand, the model cannot anticipate generalisation in terms of similarity. The presentation of A might produce some activity in the configural units to which it is similar (*Aw* and *Ax*). However, the nature of the initial discrimination means that these representations are equally similar to the representations involving B (*Bw, Bx*), C (*Cw, Cx*) and D (*Dw, Dx*) so no differences in generalisation should be observed; indeed there is evidence showing that any association between A, B, C, and D and w and x are equivalent following biconditional discrimination training (see Experiment 2 Honey & Watt, 1998). Instead, Pearce predicts that the presentation of A will result in the recruitment of an additional configural unit once it enters into association with the representation of shock. In addition, because configural units are only activated by

specific conjunctions of stimuli, and stimulus A was never presented in combination with stimulus C, it will not be possible for C to ever excite a configural unit sensitive to the presentation of A. This is not to say that Pearce's (1994) model account of biconditional discrimination learning is without merit. A study by Coutureau et al. (2002) found a dissociation between conditional learning and acquired equivalence. In their study, rats with a lesion to the entorhinal cortex learned an initial biconditional discrimination of the form previously described but failed to show an increased generalisation between stimuli that had been initially trained as equivalent after a revaluation stage. This suggests that, under some conditions, biconditional discrimination learning proceeds just as anticipated by Pearce (1994).

1.3.1 Extended configural theory and acquired equivalence

Pearce's (1994) theory is unable to account for the fact that in discriminations where some stimuli (e.g., A and C) have accompanied some relationships (e.g., w+, x-) and other stimuli (e.g., B and D) have accompanied complementary relationships (e.g., w-, x+), animals develop an acquired equivalence to stimuli that have signalled the same associate. However, it is possible to extend this theory, whilst preserving its configural nature, by assuming that configural units might not only represent the stimuli presented in a specific trial, but also encode some component of the trial outcome (Honey & Watt, 1998). This extension to Pearce's (1994) model allows it to account for the results observed after revaluation in a configural acquired equivalence task, central to this thesis. According to the extended configural model, presenting the biconditional discrimination previously described would result in the recruitment of configural units *Aw+*, *Ax-*, *Bw-*, *Bx+*, *Cw+*, *Cx-*, *Dw-*, *Dx+*. Because similarity is determined by the proportion of unique and common shared elements,

once the outcome of a given trial also forms part of the representation, the configural units activated by A and C (e.g., *Aw+* and *Cw+*) would be more similar than those activated by, for example, A and D (e.g., *Aw+* and *Dx+*), and generalisation should be more marked.

However, this extended configural theory is still inadequate to account for other configural acquired equivalence procedures, such as the finding that a discrimination proceeds more readily if it is contextually congruent than incongruent (e.g., Hodder et al., 2003; Honey & Ward-Robinson, 2001; Robinson & Owens, 2013). For example, Honey and Ward-Robinson (2001) (see **Table 2**) presented rats with an initial biconditional discrimination identical to the one already described, in which some contexts (A and C) signalled the delivery of food when presented in conjunction with a tone (w) and some contexts (B and D) signalled the delivery of food when presented with a clicker (x) (Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+). In a subsequent stage, rats received a new discrimination in which the same set of contexts were now presented in conjunction with new visual stimuli. For a subset of rats, this new discrimination was contextually *congruent* with the initial auditory discrimination (i.e., Av+, Ay-, Bv-, By+, Cv+, Cy-, Dv-, Dy+, where v and y represent a constant or pulsated light, respectively). That is, the equivalent relationships that had been established during training (e.g., A and C) remained relevant during this second discrimination. A second subset of rats received a contextually *incongruent* discrimination (i.e., Av+, Ay-, Bv-, By+, Cv-, Cy+, Dv+, Dy-), where initially equivalent relationships were no longer preserved. Results showed that rats acquired the discrimination more readily when presented with a congruent discrimination.

Table 2

Experimental design for Experiment 2 in Honey and Ward-Robinson (2001)

Training	Congruent Discrimination	Incongruent Discrimination
Aw+ Ax-	Av+ Ay-	Av+ Ay-
Bw- Bx+	Bv- By+	Bv- By+
Cw+ Cx-	Cv+ Cy-	Cv- Cy+
Dw- Dx+	Dv- Dy+	Dv+ Dy-

Note. A, B, C and D represent different contexts (A and C warm or cool floors; B and D dotted or checked wallpaper). w and x denote a tone and a clicker and v and y a constant or pulsed light. + and – indicate the delivery of food and no food, respectively.

The extended configural account has no grounds to predict a difference in the ease at which the congruent and incongruent discriminations are acquired. After the recruitment of the initial configural units: *Aw+*, *Ax-*, *Bw-*, *Bx+*, *Cw+*, *Cx-*, *Dw-* and *Dx+*, the presentation of the congruent discrimination should result in the formation of configural units *Av+*, *Ay-*, *Bv-*, *By+*, *Cv+*, *Cy-*, *Dv-* and *Dy+*. Similarly, the incongruent discrimination should lead to the formation of configural units *Av+*, *Ay-*, *Bv-*, *By+*, *Cv-*, *Cy+*, *Dv+*, *Dy-*. Here, it is important note that with regards to the initial discrimination, all the configural units formed during the congruent and incongruent discriminations reflect the fact that (i) the auditory elements (w and x) have been replaced with a visual one (v or y), (ii) each initial context (e.g., A) will partially activate two configural units (*Av* and *Ay*), and (iii) each context will be equally likely to signal food or its absence. Thus, the extended configural theory has no grounds to anticipate a faster acquisition of the congruent over the incongruent context discrimination. In order to account for these findings, a theory would have to allow for the configural units to somehow encode the fact that different contexts

(e.g., A and C) have received an equivalent training that is preserved in the congruent, but not the incongruent, subsequent discrimination.

1.4 Honey's connectionist network and acquired equivalence (Honey, 2000; Honey et al., 2010)

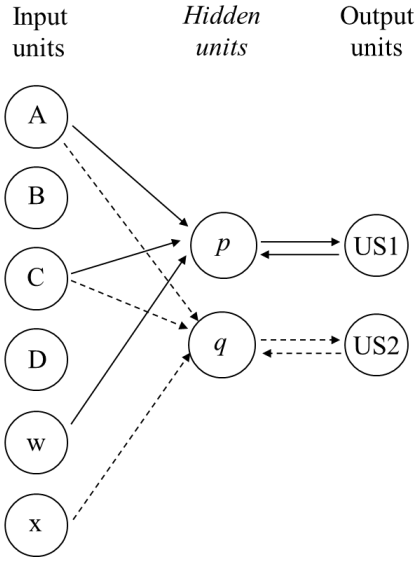
The finding that generalisation between two stimuli that have shared a common training history in a biconditional discrimination procedure increases after revaluation to one of the stimulus is beyond the scope of an important configural theory like Pearce's (1987, 1994). Even when this theory is extended to allow for some representational element of the outcome to be captured by each configural unit (e.g., Honey & Watt, 1998), it fails to account for the faster acquisition of a congruent versus an incongruent contextual discrimination. These more complex acquired equivalence procedures illustrate the need for a theory to allow for stimuli that are otherwise equally similar to be grouped together on the basis of the associations they have signalled.

Honey and colleagues (Honey, 2000; Honey et al., 2010; Honey & Ward-Robinson, 2002) proposed a connectionist network in which the presentation of one stimulus (e.g., C) is capable of generating activity in a hidden unit that is also activated by other stimuli (e.g., A and w), as exemplified in **Figure 2**. This type of analysis assumes that a hidden unit (e.g., *p*) will mediate the association between the input (e.g., Aw) and output (e.g., food) units. However, it assumes that this same hidden unit can be activated by the presentation of other stimuli (e.g., Cw). In general terms, the input and output layers of the network will become active upon presentation of specific stimuli (e.g., a specific light and food pellets). Initially, the available hidden units will not become active upon presentation of specific stimuli. Instead, it is assumed that the connections between inputs and hidden units will be weak and random at the onset of training, but these links will change as a consequence of training in the following way. As training proceeds, hidden units will

become tuned to respond to specific patterns of stimulation through the enhancement of the input-to-hidden layer connections (e.g., A → *p*, w → *p*). The inhibitory links between hidden units (Honey, 2000) will ensure that once a random hidden unit 'wins' the competition, it will be the only one fully active upon presentation of a given input pattern. At the same time, once a specific hidden unit is recruited, it will develop a connection to the corresponding output unit (e.g., *p* → food). The changes in the weight of the connections across network layers will be governed by Hebbian and anti-Hebbian learning principles (Hebb, 1949). For example, if hidden unit *q* is active when the output unit is also active, the weight of the connection between the two will increase. If, however, hidden unit *q* is inactive when a given output unit is active, the strength in the connection between the two will decrease.

Figure 2

Honey's (Honey, 2000; Honey et al., 2010) Connectionist Network



Note. For the sake of clarity only trials Aw+, Cw+, Ax- and Cx- are represented. Another two hidden units would encapsulate inputs Bw-, Dw- (hidden unit *r*) and Bx+, Dx+ (hidden unit *s*).

Up until this point, the network described is essentially like Pearce's (1994). However, the critical feature which sets this network apart from other connectionist networks is the bidirectional link between hidden and output units, which results in the formation of an excitatory forward connection between a hidden unit and the outcome trial, and an excitatory backward connection from the outcome back to the corresponding hidden unit. This feature plays and essential role in determining why the presentation of another compound (e.g., Cw) comes to also elicit activation in hidden unit *p* over any other hidden unit (e.g., *q*). According to Honey et al. (2010), the probability that a specific hidden unit will be selected depends on the activity that it receives from the input units and the output unit that the pattern activates. Suppose that, as part of the biconditional discrimination described in preceding paragraphs, trials Cw+ and Bw- occur after the Aw+ → *p* connection has been established. Both compounds might have an initial tendency to activate hidden unit *p* by dint of the w → *p* connection. However, once the outcome occurs, hidden unit *p* will likely be selected upon presentation of Cw+ by virtue of the additional source of activation from the food → *p* backward connection previously established. This will result in the strengthening of the connections across layers (Cw → *p*, *p* → food and food → *p*). This will not be the case when Bw followed by no food occurs. Even if hidden unit *p* is initially selected, there will be no changes in the reciprocal connections between *p* → food (because no food will be delivered) and as a result, hidden unit *p* will be less likely to be selected the next time Bw- is presented. Instead, through training, Bw- and similar compounds will become linked to a different hidden unit (e.g., *q*), which will anticipate the delivery of no food.

Evidence demonstrating the critical role of the reciprocal hidden-to-output and output-to-hidden layers in selecting the appropriate hidden unit comes from

experiments by Honey and Ward-Robinson (2002), in which they deliberately attempted to activate specific hidden units to control performance upon presentation of a series of novel stimulus compounds. To that end, rats were given an initial biconditional discrimination in which contexts A and B signalled food when presented together with auditory stimulus x, but not with y, and context C and D signalled the delivery of food when presented with y, but not with x (see the upper half of **Table 3**). According to a connectionist network of the characteristics just described, this discrimination should result in the formation of four hidden units: *p* (Ax+, Bx+), *q* (Ay-, By-), *r* (Cx-, Dx-) and *s* (Cy+, Dy+). Once rats had acquired the initial discrimination, they were placed in undecorated chambers and they were primed with food and no food manipulations. In the food priming manipulation, rats received a period with no presentation of food followed by a period in which several food pellets were delivered. This manipulation ensured that the food output unit was active immediately prior to the presentation of test compounds AB and AD. In the no food priming manipulation, rats received several food pellets before a period with no food delivered right before the presentation of test compounds CD and CB. The bottom half of **Table 3** illustrates the number of sources of activation to each hidden unit that should have resulted after this manipulation, assuming that priming with food and no food had the ability to activate specific hidden units.

Table 3

Experimental design for Experiment 2 in Honey and Ward-Robinson (2002) and Sources of Activation to Hidden Units during the Test.

Training	Test			
Ax+ Ay- Bx+ By- Cx- Cy+ Dx- Dy+	No food --> food; AB vs. AD Food --> No food; CD vs. CB			
Hidden unit (and corresponding output unit)	Food		No food	
	AB	AD	CD	CB
p (food)	3	2	0	1
q (no food)	2	1	1	2
r (no food)	0	1	3	2
s (food)	1	2	2	1

Note. A, B, C and D denote contexts (A and C; checked or dotted walls, respectively, and B and D; warm or cool floors). x and y represent a tone and a clicker. + and – indicate the delivery of food or no food, respectively. *p, q, r,* and *s* represent notional hidden units resulting from a connectionist network of the kind described by Honey and colleagues. *p* reflects that A, B and x will become linked to the food output unit after training. *q* reflects that A, B and y will become linked to the no food output unit. *r* shows that C, D and x will become linked to the no food output unit after training. *s* shows that C, D and y will become linked to the food output unit. The numerical values indicate the number of sources of activation that each hidden unit will received upon presentation of novel compounds AB vs. AD (primed by food) and CD vs. CB (primed by no food).

For example, the presentation of AB in the food priming condition should elicit activation in hidden units *p* and *s*, which would have established links to the representation of food during training. Moreover, inputs A and B would act as two additional sources of activation to hidden unit *p* and hidden unit *q*, through their corresponding A → *p/q* and B → *p/q* connections. Similarly, presenting AD in the food priming condition would elicit activation to the *p* and *s* hidden units. However,

this novel compound should now add one extra source of activation to each hidden unit, through the connections from A → *p/q* and D → *r/s*. Overall, compound AB would add three sources of activation to a hidden unit and should control behaviour over compound AD, which would add a maximum of two sources of activation to any given hidden unit. When this same analysis is applied to compounds CD and CB in the no food priming condition, CD should control performance over compound CB. Results in Honey and Ward-Robinson (2002) showed that rats responded more vigorously to AB than to AD after having been primed with food, and that they responded more vigorously to CD than to CB in the no food prime condition. These findings are important because they demonstrate that the presentation of an outcome can influence the process of selecting which hidden unit becomes active upon presentation of specific inputs, and provide direct support for the connectionist architecture described by Honey and colleagues (Honey, 2000; Honey et al., 2010).

The ability of this network to allow for the sharing of hidden units by compounds that have not been presented together in a given trial is a critical distinction from Pearce's (1994) connectionist network, and equips it to accommodate findings from complex configural acquired equivalence procedures beyond the scope of even extended configural accounts. First, consider the set of hidden units that will be recruited as a result of a conditional discrimination of the form Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+. By the end of training, we should expect four hidden units to capture the fact that A and C signal food when presented with w (hidden unit *p*) but not with x (hidden unit *q*) and that B and D signal food when presented with x (hidden unit *r*) but not with w (hidden unit *s*). This grouping of A and C, and B and D, will mediate the generalisation of responding observed when A and B are revalued in a subsequent stage. When A is presented and revalued

with a footshock, it will partially activate hidden units *p/q*, which will result in them becoming linked to shock. Presenting C will also partially activate the now fear-eliciting *p/q* hidden units. On the other hand, the presentation of B, which will partially activate hidden units *r/s*, will result in them becoming linked to the representation of no shock, and no generalised fear to D. Consider now how the finding that a congruent contextual discrimination is more readily acquired than an incongruent one, which challenged the extended configural account, can also be explained in terms of this connectionist network (e.g., Experiment 2 in Honey & Ward-Robinson, 2001). By the end of training, the state of the network will be identical to that just described, with ACw+ and ACx- activating hidden units *p* and *q*, respectively, and BDw- and BDx+ activating hidden units *r* and *s*. For rats in group congruent, the presentation of Av+, Ay-, Bv-, By+, Cv+, Cy-, Dv- and Dy+ will map into the existing network's representations and aid learning. For example, the presentation of Av+ will activate hidden unit *p* by virtue of the A → *p* and bidirectional food → *p* connections. When Cv+ is presented, it will correctly excite hidden unit *p* as well, through the C → *p* and bidirectional *p* → food associations. The same will be true for the rest of the input patterns and hidden units *q, r,* and *s.* However, rats in group incongruent, which received Av+, Ay-, Bv-, By+, Cv-, Cy+, Dv+ and Dy- will not benefit from the existing representations and learning should be hindered. For instance, the presentation of Av+ will excite hidden unit *p* by dint of the A → *p* and bidirectional food → *p* links. However, this pattern of activity will not facilitate learning about Cv-, which now signals the opposite reinforcing contingency.

1.5 A formal implementation of Honey's connectionist network (Robinson et al., 2019)

Honey and colleagues have carefully described this 3-layered connectionist network (Honey, 2000; Honey et al., 2010; Honey & Ward-Robinson, 2002). However, in a recent article published during the course of my PhD, Robinson et al. (2019) presented a model of discrimination learning, directly informed by the analysis of various forms of acquired equivalence, which provided one way of formalising the ideas described in the preceding paragraphs as a computational implementation.

This Hebbian learning network consists of a layer of input units, which represent discrete experimental stimuli (e.g., a tone), a layer of hidden units, and a layer of output units, which represent the outcome of a trial (e.g. food). The activity in the network is propagated trough feed-forward (input-to-hidden and hidden-to-output layer) and feedback (output-to-hidden layer) connections. In brief, Robinson et al. (2019) assume that when a trial is presented (e.g., Aw+), the corresponding input and output units will be excited from their resting value (0) to their activated value (1). Activation from the corresponding input and output units will propagate to a hidden unit through the input-to-hidden and the output-to-hidden forward and backward projections, respectively. The activation to any selected hidden unit will be determined by **Equation 3.1**. That is, the activation to hidden unit *p* (y_p) will be determined by the activation of input unit A (x_A) and the weight of the connection between input A and hidden unit p (w_{Ap}) plus the activation to the corresponding output unit (z_{us1}) and the weight of the connection between the output unit and hidden unit *p* (w_{us1p}). Connection weights are determined randomly from a range of 0 to 1 at the beginning of each simulation.

$$y_p = \sum x_A w_{Ap} + \sum z_{us1} w_{us1p} \qquad \textbf{Equation 3.1}$$

In order to enhance the contrast between levels of activation at the hidden unit layer, Robinson et al. (2019) applied a *winner-takes-all* (WTA) mechanism to ensure that activity in the selected hidden unit was proportional to the most active unit within the layer, and that the maximal level of activation that any hidden unit could afford was equal to 1. This WTA mechanism ensures that once the network is trained with an initial discrimination, a single hidden unit becomes fully active upon presentation of the corresponding input and output patterns whilst the rest of the hidden units within the layer receive minimal activity. Finally, weight changes are adjusted across adjacent layers in the network following a conditional principal component analysis (CPCA) Hebbian learning rule. The CPCA determines the probability that a sending unit, from either the input or the output layer, is active, given that the receiving unit, from the hidden or output layer, is active. Once the CPCA is applied to the network, weight changes across adjacent layers are governed by **Equation 3.2**. Here, Δw_{Ap} denotes the change in the weight of the connection between sending input A and receiving hidden unit *p*. ε is a fixed learning rate parameter from 0 to 1. y_p is the activity to hidden unit *p*, x_A indicates the activity to input unit A and w_{Ap} the connection between input A and hidden unit *p*. According to this Hebbian learning algorithm, the changes in the connection weights across adjacent layers will move in the direction of the *sending* unit. That is, if the sending unit (e.g., input unit A) is active when the receiving unit (e.g., hidden unit *p*) is

active, the weight in the connection between the two will increase. If the sending unit is inactive when the receiving unit is active, the weight in the connection between the two will decrease. If the receiving unit is inactive, no changes in the connection weights will occur.

$$\Delta w_{Ap} = \varepsilon\left[y_p\left(x_A - w_{Ap}\right)\right] \qquad \textbf{Equation 3.2}$$

This formal implementation is important for several reasons. On the one hand, it serves as a means to corroborate that a connectionist network with the characteristics of the ones verbally described by Honey and colleagues is indeed equipped to accommodate extant experimental data. Specifically, Robinson et al. (2019) presented a series of successful simulations of an increased generalisation between contexts that had initially been trained as equivalent in a biconditional discrimination (e.g., Honey & Watt, 1998; Iordanova et al., 2007; Ward-Robinson & Honey, 2000), differences in responding to congruent and incongruent context combinations (e.g., Honey & Ward-Robinson, 2002), the more readily acquisition of a congruent, versus an incongruent, discrimination (e.g., Honey & Ward-Robinson, 2001) and, with some parameter dependency, the fact that after an initial configural discrimination, acquisition of a whole reversal proceeds more readily than a partial reversal (e.g., Robinson & Owens, 2013). Additionally, Robinson et al. (2019) reported that the network was also capable of successfully simulating data from a simple acquired equivalence procedure of the kind reported by Honey and Hall (1989). This suggests that the network cannot only accommodate complex

biconditional discriminations, but also simpler non-configural ones. On the other hand, Robinson and colleagues offered computational support for the essential role of the bidirectional links between the hidden and output layers. That is, conducting these simulations after removing the feedback link from the output to the hidden layer abolished acquired equivalence.

1.6 Structure of thesis

A 3-layered connectionist network that allows for the outcome of a trial to influence the hidden unit selection process (Honey, 2000; Honey et al., 2010; Honey & Ward-Robinson, 2002) appears to provide the most comprehensive account for the results observed in the different forms of acquired equivalence described in this introductory section because it reconciles the fact that stimuli that signal the same associate will be treated as equivalent. Nevertheless, there are procedures that could help refute or qualify this model.

In their description of the applications of the model, Honey et al. (2010) argued that a connectionist analysis could also apply to the analysis of intra-dimensional and extra-dimensional shift (IDS/EDS), a procedure traditionally considered to illustrate the influence of predictive validity on attention (e.g., Mackintosh & Little, 1969; Owen et al., 1991; Roberts et al., 1988). Support for this claim came from the observation that performance in both configural acquired equivalence and attentional set tasks has been found to be selectively impaired in healthy older adults (e.g., Owen et al., 1991; Robinson & Owens, 2013) and selectively affected by brain lesions (e.g., Coutureau et al., 2002; Oswald et al., 2001). Therefore, in Chapter 2 we sought to test this claim by assessing participants' performance in configural acquired equivalence and in two different attentional set tasks: IDS/EDS and optional-shift. We also measured eye gaze in one experiment to explore the relationship between a cue's predictive validity and participants' eye gaze further.

Chapter 3 had a dual purpose. First, we asked whether experiencing the same or different outcomes across training and revaluation would have any effect in the observed acquired equivalence effect. This question was theoretically motivated by

the observation that performance in discrimination learning tasks is improved by the use of different outcomes, as opposed to a single outcome across reinforced responses (Trapold, 1970), and experimentally motivated by the fact that some configural acquired equivalence tasks have used the same outcomes across training and revaluation (Coutureau et al., 2002; Iordanova et al., 2007) and others have used different outcomes across stages (Honey & Watt, 1998; Ward-Robinson & Honey, 2000). Second, we attempted to replicate Delamater's (1998) findings that a reversal stage with different outcomes within stimulus modality is acquired more readily than a reversal with the same outcomes within stimulus modality. This is of interest to this thesis because despite being a simple, non-configural, acquired equivalence task, it is not amenable to a mediated conditioning interpretation. Instead, we interpreted the results in connectionist terms.

Finally, although Honey and colleagues provided a careful and detailed description of their network, it should be noted that informal descriptions of a learning network could be prone to unforeseen errors. Chapter 4 focused on the recent publication of one mathematical implementation of this 3-layered connectionist network (Robinson et al., 2019). We sought to qualify the model by assessing the extent to which it could accommodate experimental findings from previous chapters.

Chapter 2:

Dissociation of two measures of attentional set with configural acquired equivalence

Attentional set refers to an organism's preference to attend to relevant information and ignore irrelevant information in its environment. A prominent paradigm used to assess attentional set is the intra-dimensional/extra-dimensional shift (IDS/EDS) task. In an example IDS/EDS, Mackintosh and Little (1969) presented pigeons with visual stimuli on keylights that differed across two dimensions: line orientation and colour. Subjects were trained on a discrimination in which only one of those dimensions (e.g., colour) was a reliable predictor of an appetitive outcome: **A**w+, **A**x+, **B**w-, **B**x-, where A/B represent two exemplars of the relevant dimension (e.g., "red" and "yellow" from the dimension colour), w/x represent two exemplars belonging to the irrelevant dimension (e.g., "vertical" and "horizontal" from the dimension line orientation), and "+" and "-" indicate food and no food, respectively. After the attentional set to the predictive dimension had been established, an attentional shift took place. The shift required two subgroups of pigeons to respond to either new exemplars of the same predictive dimension (i.e., colour): **C**y+, **C**z+, **D**y-, **D**z- (i.e., perform an IDS), or to new exemplars of the previously irrelevant dimension (i.e., line orientation): C**y**+, C**z**-, D**y**+, D**z**- (i.e., perform an EDS). The group that received an IDS mastered the discrimination more readily than the group that received an EDS, suggesting subjects had developed either an attentional bias toward the relevant stimulus dimension, an attentional bias away from the irrelevant dimension, or both. This IDS superiority has been reported in human (Owen et al., 1991; Roberts et al., 1988; Sahakian & Owen, 1992), and non-human subjects alike (Garner et al., 2006; Roberts et al., 1988).

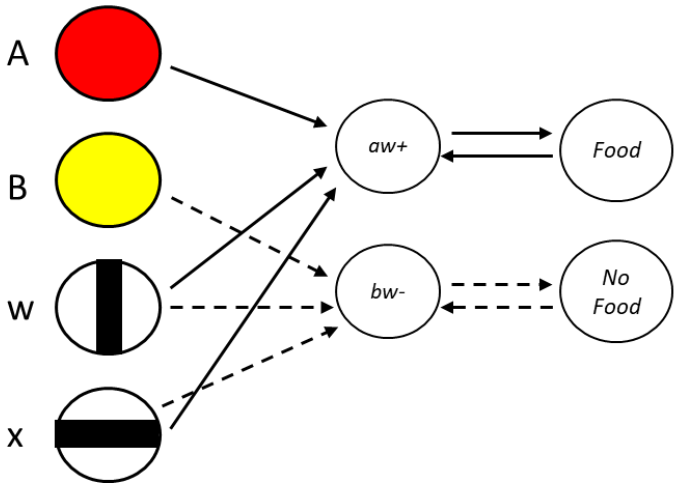
Whilst IDS/EDS has been said to evidence the influence of attention on learning (Le Pelley, 2004; Mackintosh, 1974), the relationship between how predictive a stimulus is and how much is learned about it is open to non-attentional

interpretations. Honey and colleagues (Honey et al., 2010; Honey & Ward-Robinson, 2002) have proposed a three-layered connectionist network to account for the observed IDS superiority with no explicit notion of attention. In brief, the network operates on an input, hidden and output layer. The crucial feature of this network relies on bidirectional links between the hidden and output units, as shown in **Figure 3**. According to this account, any given trial will result in a pattern of activation in the input and output layers of the network, which will be mediated by a hidden unit. For example, during the initial discrimination learning in IDS/EDS, the network will receive activation from a certain pattern of inputs. Let us use trial Aw+ as an example, shown in **Figure 3**, where A refers to the relevant stimulus dimension (colour), w denotes the irrelevant stimulus dimension (line orientation) and "+" signals the outcome trial (food). This trial will result in the activation of input units *a* and *w*. Unlike the activation of input units, the activation of hidden units will be weak and random at first. A *winner-take-all* mechanism will operate in the hidden unit layer, and determine the winner by selecting the single most active hidden-unit upon presentation of a pattern of input activation (e.g., hidden unit *aw+*). The selection of the winning hidden unit will, in turn, reduce activity in all the other hidden units available. Once a hidden unit has been selected, activation to the corresponding output unit (e.g., *Food)* will follow and will, in turn, feedback to the hidden unit. These reciprocal connections between hidden and output units allow for the sharing of hidden units between similar inputs that signal the same outcomes. The next time a similar pattern is presented (e.g., Ax+, where A represents the same colour and x a different line orientation), it will likely activate hidden unit *aw+* by virtue of its shared similarity to Aw and the mediation of the shared outcome. The network will therefore "blend" the internal representation of these two similar trials

into hidden unit *aw+***.** Suppose now that trial Bw-, also depicted in **Figure 3**, occurs. Hidden unit *aw+* could be selected by dint of the common element w. However, when the actual outcome occurs (i.e., *No food*), *aw+* will not receive any activation from the output unit, which will decrease its chance of being selected as the winning hidden unit next time Bw- is presented. Instead, through training, input Bw- and similar inputs will become linked to a hidden unit distinct from *aw+*, one that will map onto the corresponding correct trial outcome.

Figure 3

Depiction of a Connectionist Network Analysis of IDS/EDS



Note. Trials Aw+, Ax+ (red (dark gray) with horizontal or vertical lines) and Bw-, Bx- (yellow (white) with horizontal or vertical lines) are exemplified. The network is composed of: input units, which represent the individual sensory components of each trial (e.g., red colour and horizontal line orientation); hidden units, which encode the combination of stimuli on a trial; and output units, which represent the outcome of a particular trial. Weight changes between the connections change through learning. The ouput → hidden connection gives feedback about the outcome of the trial, and allows for similar inputs to share hidden units.

By the end of the training stage in Mackintosh and Little's, (1969) experiment, the connection between inputs Aw/Ax and Bw/Bx and their respective outcomes (food and no food) will be mediated by different hidden units (e.g., *aw+* and *bw-*). Hebbian and anti-Hebbian learning processes (Honey et al., 2010) will govern the changes in the connections between the input units from the relevant and irrelevant dimensions and their respective hidden units. Because inputs from the relevant dimension unambiguously signal the outcome (A → + and B → -), their connection to the hidden units will be double the strength than between those from the irrelevant dimension and any of the hidden units (w → +/- and x → +/-). The ability for the network to anticipate differences between the IDS and EDS stages hinges on these enhanced input-to-hidden connections formed during training. When new compounds Cy, Cz, Dy, Dz are introduced, the previously relevant dimension will be responsible for the majority of activation to the corresponding hidden unit. The group that requires subjects to perform an IDS will benefit from a greater stimulus generalisation between previous stimuli and new exemplars Cy/Cz and Dy/Dz because the already established hidden units and their connections to the corresponding outputs will be used with no additional modification. The group that is required to perform an EDS will have to restructure the set of connections established during training, which will slow discrimination learning. Let us assume a Cy+ trial occurs in the IDS group, where new stimuli from the previously relevant dimension (e.g., "blue" and "green") are reliable predictors of the trial outcome. The existing Aw/Ax → *aw+* association might correctly generalize to trial Cy+ through immediate positive transfer, allowing the new exemplar to benefit from an existing hidden unit that reliably predicts the occurrence of the trial outcome (food). Of course, the existing Bw/Bx → *bw-* association could also generalize to trial Cy+

through immediate negative transfer. On such trials, the *bw*- hidden unit will be incorrectly selected upon presentation of Cy+ and the incorrect outcome trial will be anticipated. However, the occurrence of the actual trial outcome (food) will result in the crucial bidirectional links between *bw*- and the output unit not sustaining any activity, and the capacity of a Cy+ trial to activate the *bw*- hidden unit will diminish over time. Because stimuli from the previously irrelevant dimension will have little or no ability to activate hidden units, the EDS group, in which new stimuli from the previously irrelevant dimension (e.g., "left diagonal" and "right diagonal") are now predictive of the outcome, will not benefit from stimulus generalisation to the same extent and the discrimination will not be as easily solved by the network. With the functioning of this simple connectionist network in mind, the observed IDS discrimination superiority could be the reflection of these Hebbian and anti-Hebbian processes taking place. An IDS superiority would reflect the cost associated with the EDS group having to restructure the internal representations between different inputs and outcomes. Under this connectionist interpretation, the predictiveness of a stimulus dimension would simply influence its ability to activate hidden units, which mediate learning about the stimulus and its outcome, without a need to invoke explicit changes in attention (Honey et al., 2010).

Attentional set tasks provide examples of learning in which subjects need to identify and learn to respond to a relevant dimension. However, other forms of discrimination learning require subjects to learn about the conditional relationship between two or more dimensions to correctly solve a discrimination. An example of conditional discrimination learning are configural acquired equivalence tasks, in which generalisation between stimuli is enhanced as a result of a common training history.

During a configural acquired equivalence procedure, subjects need to learn about the *configuration* of a series of stimulus dimensions to predict outcomes correctly. I have already offered a detailed explanation of different forms of configural acquired equivalence using nonhuman animals as subjects. However, these tasks can also be adapted for human participants. For example, Hodder, George, Killcross, and Honey, (2003) gave participants an initial discrimination of the form Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+ on an allergy prediction task. Cues A-D represented meat products and cues w/x represented different vegetables. "+" and "-" denoted patient's Mr. *X* allergic reaction and no allergic reaction, respectively. Note how, when meals Aw+ or Cw+ were presented, Mr. *X* suffered an allergic reaction, but Mr. *X* was fine with meals Ax- or Cx-. The opposite arrangements were true for meals Bw-, Dw-, Bx+ and Dx+, so A/C and B/D signalled equivalent stimulus-outcome contingencies when presented with w and x. Thus, it was not enough for participants to learn about a single stimulus dimension (e.g., meat products A-D) because, unlike in attentional set tasks, no single cue uniquely signalled either outcome. Participants had to learn about the configuration of both stimulus dimensions to successfully predict the occurrence of an allergic reaction. In a subsequent stage, participants were asked to learn a new congruent or incongruent discrimination in which the same meat products (A-D) were paired with new vegetables v and y. For participants in the congruent condition, the new discrimination kept the equivalent stimulus-outcome relationship for cues A/C and B/D when presented with v and y (Av+, Ay-, Cv+, Cy- and Bv-, By+, Dv-, Dy+). Participants in the incongruent condition, however, were presented with a new discrimination in which A/C and B/D were no longer treated as equivalent (Av+, Ay-, Cv-, Cy+ and Bv-, By+, Dv+, Dy-). Hodder et al. (2003) found that participants

performed more proficiently in the congruent discrimination, suggesting that the initial configural discrimination had resulted in representational equivalence between cues A/C and B/D, which had subsequently facilitated the acquisition of the congruent, but not the incongruent new discrimination.

Honey et al.'s (2010) argued that their connectionist network could not only accommodate IDS superiority, but could also readily explain the results observed in configural acquired equivalence tasks. Briefly, after the end of the initial discrimination, the state of the network would be analogous to the one at the end of the IDS/EDS: similar patterns (e.g., Aw+ and Cw+) will tend to activate a common hidden unit (*acw*+) which will, in turn, reliably predict the outcome trial (e.g., allergic reaction). Similar patterns that have been trained to predict a different outcome trial (e.g., Ax- and Cx-) will tend to activate a different common hidden unit (*acx*-), which will signal the opposite outcome (no allergic reaction). The subsequent congruent discrimination will benefit from the existing network's structure. For example, trials Av+ and Cv+ will tend to activate hidden unit *acw*+ by dint of the A, C and + → *acw*+ connections, which will anticipate the correct outcome trial. However, Just like in an EDS discrimination, some trials in the incongruent discrimination will require the network to readjust its connections, resulting in the observed decrement in performance.

Honey et al.'s (2010) claims of a common mechanism accounting for attentional set and configural acquired equivalence receive indirect support from the finding that healthy older adult participants show impairments specific to acquired equivalence and IDS/EDS (e.g., Owen et al., 1991; Robinson & Owens, 2013; Simon & Gluck, 2013). Using a configural acquired equivalence procedure,

Robinson and Owens (2013) reported a decreased performance during test trials for a group of older participants compared to younger controls, but no age differences during the initial discrimination learning. That is, both young and older adults learned the initial configural discrimination, but older adults seemed to experience difficulties forming an acquired equivalence set. Similarly, Owen et al. (1991) found that healthy older adults were selectively impaired when performing an EDS, but not an IDS, as compared to younger controls. Recent computational evidence has suggested that these selective impairments could be the result of a reduction in the critical bidirectional connections between output and hidden units (Robinson et al., 2019). The specific impairments in performance observed in experiments with older participants (analogous to the impairments in performance observed when reducing connections between the output and hidden units in computer simulations) could be the result of developmental changes in the older adults that have not yet manifested in the younger volunteers.

In more general terms, individual differences in the way the network changes the weights in the connections between output and hidden units are to be expected, and even young participants should show variations in their performance in acquired equivalence and attentional set tasks. If configural acquired equivalence and attentional set can be amenable to a single mechanism, we reasoned that individual differences in the way the network forms its connections should affect performance in both tasks and performance should correlate. However, to the best of our knowledge, studies have not assessed the relationship between performance in configural acquired equivalence and attentional set tasks in a single, comparable experiment; a necessary step in the assessment of individual differences in performance.

The experiments reported in this chapter address this issue by conducting a direct comparison between performance in a configural acquired equivalence task and two different attentional set tasks: IDS/EDS and optional-shift. Based on previous evidence and under the premise that the network described by Honey and colleagues could accommodate both processes, we reasoned that performance in both tasks should be expected to correlate positively. Experiment 1 provided within-subjects demonstrations of the acquired equivalence effect and IDS superiority, but failed to detect a positive correlation in performance between the two. Experiment 2 found a positive correlation between configural acquired equivalence performance and a different attentional set task: optional-shift. Experiment 3 replicated the findings from Experiment 2 and incorporated eye-tracking to assess the relationship between predictiveness and learning. Finally, Experiment 4 aimed to provide conclusive support for a positive relationship between configural acquired equivalence and optional-shift and incorporated a control *N*-back.

2.1 Experiment 1

Experiment 1 intended to demonstrate the acquired equivalence and attentional set effects in our cohort of participants, and to directly assess the relationship between performance in both tasks. The acquired equivalence task was adapted from previous biconditional configural discrimination preparations with rats (e.g., Ward-Robinson & Honey, 2000), and is summarised in **Table 4(a)**. Stage 1 presented participants with a configural discrimination of the form Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+. During Stage 2, we presented stimuli A-D in isolation and revalued A and B. Revaluation trials A+ and B- continued to reliably signal an outcome trial during Stage 2. Test trials C and D, however, were never followed by explicit feedback. Revaluation and test trials during Stage 2 were intermixed. The initial configural discrimination was designed to render stimuli A and C, which signalled the same relationships when paired with w(+) and x(-), and stimuli B and D, which also signalled the same relationships when paired with w(-) and x(+), as equivalent. Thus, we anticipated participants would generalise their responses from stimulus A to stimulus C, and from stimulus B to stimulus D despite the absence of explicit feedback; the acquired equivalence effect.

Specifically, participants were asked to put themselves in the role of a marine tour guide and try to determine the threat that different types of octopuses posed to the tourists taking the tour. Letters A-D represented different types of eyes and w/x represent different tentacles. + and – represented the different outcomes that could follow each octopus (i.e., *bite* or *sting*). During Stage 1, all individual cues (A, B, C, D, w, x) were equally paired with outcomes + and -. Therefore, participants had to learn about the *configuration* of the different cues in order to solve the discrimination correctly. During Stage 2, cues A-D were presented but only

revaluation cues A and B were followed by explicit feedback. This stage was ambiguous because no single cue uniquely predicted either outcome during Stage 1. However, we expected participants to transfer responding to C and D based on the initial equivalent training.

Table 4

Experimental Design for the Configural Acquired Equivalence and Pilot Intra/Extra Dimensional Set-Shifting Tasks in Experiment 1

Stage 1	Stage 2
a. Acquired Equivalence	
Aw + Ax –	A +
Bw – Bx +	B –
Cw + Cx –	**C ?**
Dw – Dx +	**D ?**
b. Pilot Intra/Extra Dimensional set-shifting	
Ap +	**C**r +
Aq +	**C**s +
Bp *	**D**r *
Bq *	**D**s *
Wα +	Y**λ** +
Wβ +	Y**δ** *
Xα *	Z**λ** +
Xβ *	Z**δ** *

Note. In the acquired equivalence task(a) letters A-D represent different eyes and w/x represent different types of tentacles. Participants had to learn about specific combinations of eyes and tentacles to respond correctly. + and – represent outcomes bite and sting, respectively. ? indicates the absence of feedback. Revaluation and test trials were intermixed during Stage 2. This experimental design was used in all experiments reported. In the Intra/Extra dimensional shift task, letters A, B, C and D represent different cell walls. Letters P, Q, R and S represent different cell organelles. Letters W, X, Y and Z represent different molecule shapes and Greek letters α, β, λ and δ represent differently coloured molecule bounds. + and * indicate different outcomes (i.e., dinosaur or lizard). Intra and Extra dimensional shift trials were intermixed during Stage 2.

Experiment 1 used the IDS/EDS task from the Cambridge Neuropsychological Test Automated Battery (CANTAB - Cambridge Cognition, Cambridge, UK), an attentional set task that assesses participants' ability to perform IDS and EDS at various points in time with the use of compound stimuli from two dimensions (i.e., lines and pink geometrical shapes).

IDS/EDS is a well-established task, widely used in clinical settings (e.g., Bünger et al., 2019; Lawrence et al., 1996; Shamay-Tsoory et al., 2007), which has contributed to our understanding of attention and learning. However, it faces some problems. In this task, the experimenter is only able to counterbalance the order in which the dimensions (lines and pink geometrical shapes) are first presented to the participant. Without full counterbalancing, the differences in IDS and EDS performance could be the results of differential stimulus generalisation, rather than a specific attentional set effect. Additionally, because IDS and EDS are assessed at different points in time, and the IDS always takes place before the EDS, the diminished performance observed during the EDS could, at least in part, reflect temporal effects. To the aim of accounting for these issues, Experiment 1 incorporated an additional pilot IDS/EDS (herein p-IDS/EDS) that allowed for complete orthogonal counterbalancing and assessed performance in IDS and EDS at, on average, identical points in time, as summarised in **Table 4(b).**

During Stage 1 of our p-IDS/EDS, participants took part in a predictive task where they were asked to learn the outcomes that followed the presentation of compound stimuli. Participants were asked to put themselves in the role of a laboratory technician and try to discern which samples belonged to a common lizard and which belonged to a dinosaur. Stimuli took the form of cells or molecules, each made from two dimensions: letters A-D represented different cell walls and letters P-

S different cell organelles, together they formed our cell stimuli. Letters W-Z represented different geometric shapes and Greek letters α, β, λ and δ different coloured bounds, together they formed our molecule stimuli. + and * denote the two possible outcomes (i.e., common lizard or dinosaur). Critically, during Stage 1, only one of the stimulus dimension (e.g., cell wall) was relevant to solve the discrimination. Accordingly, to solve trials Ap+ and Aq+, a participant could simply learn that cell wall A reliably indicated +, and disregard the second stimulus dimension (cell organelle). During Stage 2, new exemplars from the same dimensions were presented. Participants received trials that required participants to perform an IDS, where new exemplars from the previously relevant dimension (e.g., cell walls) were still relevant to solve the discrimination, intermixed with EDS trials, where new exemplars from the previously irrelevant dimension (e.g., cell organelle) were now relevant. All participants completed the acquired equivalence task before the IDS/EDS. After a 5 min break, participants completed the p-IDS/EDS.

Stage 2 results in Experiment 1 demonstrated the acquired equivalence effect in our participants. Participants also showed the expected IDS superiority in the CANTAB IDS/EDS. However, we failed to obtain the same IDS superiority in our p-IDS/EDS. Contrary to what we anticipated, we did not find a positive correlation between test performance in these two tasks.

2.1.1 Method

2.1.1.1 Participants

32 students from the University of Nottingham participated (11 men, 20 women and a person who preferred not to disclose their gender (M_{age} = 25.56, *SD* = 3.03, range: 21-35). Participants received course credits or a small monetary reward.

Participants were informed about the task and debriefed after, all agreed to participate. The Research Ethics Committee from the University of Nottingham approved the experiment. The sample size in the present and all subsequent experiments was determined using G*power (version 3.1.9.2) (Erdfelder et al., 1996). An a priori power analysis was conducted to test the difference between an experimental group (mean of approximately .65 in test trials of our configural acquired equivalence task, based on prior pilot tests, not in thesis) and a constant (.50 chance level), with an effect size of .50 and an alpha level of .05. Results showed that a sample of 27 was required to achieve the recommended power of .80 (Cohen, 1992). For counterbalancing purposes, the sample size was rounded up to the next multiple of eight.

2.1.1.2 Apparatus & Materials

This and all subsequent experiments, unless otherwise stated, were conducted in a small quiet room in the Psychology building at the University of Nottingham. Participants were tested individually, sitting at approximately 50 cm from a monitor 52 (width) x 38 (height) cm in size. These and all subsequent tasks run in a desktop or laptop were programmed in PsychoPy (Peirce, 2007).

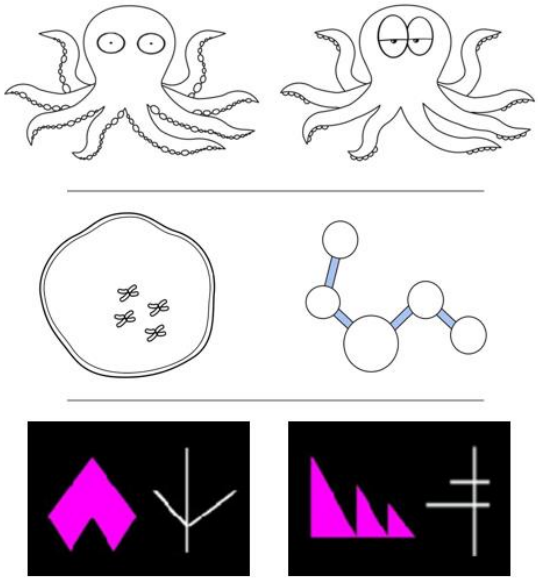
The stimuli used for the acquired equivalence task were eight black and white images of octopuses 10 cm (width) x 8 cm (height) presented on a grey background. The images had a common, body-shape outline, but combined four different sets of eyes (A-D: black eyes, angry looking eyes, sleepy looking eyes and alert looking eyes), and tentacles (w and x: tentacles with a full length of suckers and tentacles with suckers only on their tips), as seen in the uppermost panel of **Figure 4**. Stimuli were followed by the outcomes *bite* or *sting*.

Stimuli in the IDS/EDS consisted of white lines and pink geometric shapes approximately 1.8 cm (width) and 1.6 cm (height) in size. Stimuli were presented inside white rectangles approximately 4.5 (width) x 3.3 (height) cm in size on a black background. Four rectangles appeared to the left, right, above or below the centre of the screen. Specifically, the rectangles presented above and below the centre of the screen were 1.5 cm apart, and the rectangles presented to the left and right were 7.8 cm apart. The selected rectangle turned green on a correct trial and red on an incorrect trial. Only two rectangles were populated at any given trial. Example stimuli are shown in the down most panel of **Figure 4**. The task was administered using an iPad (3rd generation, Apple iOS version 9.3.5) with a 24 (height) x 17 (width) cm screen held by participants.

The stimuli used for the p-IDS/EDS task were eight images of cells and molecules 10 cm (width) x 8 cm (height) presented on a grey background. The cell stimuli combined different sets of cell walls (A-D) and cell organelles (p-s). The molecule stimuli combined different sets of molecule shapes (W-Z) and coloured bounds (α, β, λ, δ). They were followed by the outcomes *dinosaur* or *lizard*. Example stimuli are shown in the central panel of **Figure 4**.

Figure 4

Example stimuli presented during Experiment 1



Note. Example of compound stimuli presented during the acquired equivalence task (upper panel), p-IDS/EDS task (middle panel) and CANTAB IDS/EDS.

2.1.1.3 Procedure

Participants completed the acquired equivalence, IDS/EDS and p-IDS/EDS tasks in that order. All participants read an instruction sheet that emphasized the participants' right to terminate the task at any time. The experimenter left the room after ensuring participants had understood the tasks and returned to set the iPad in preparation for the IDS/EDS task and to start the p-IDS/EDS after a short break.

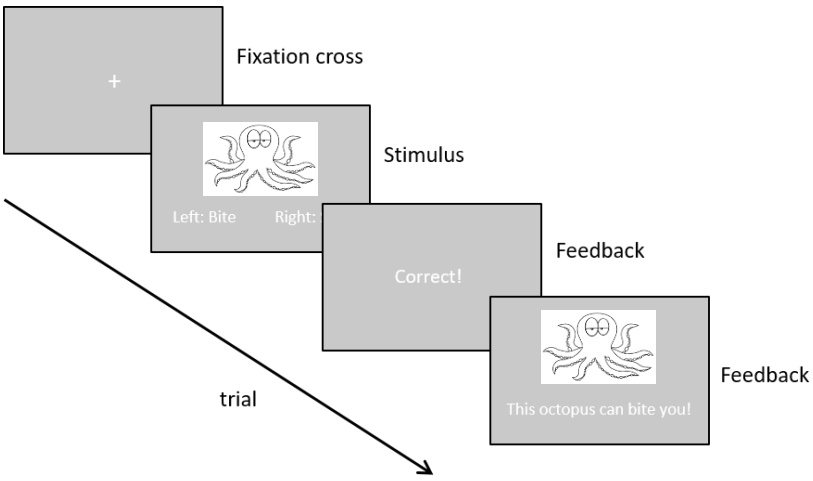
2.1.1.3.1 Acquired equivalence.

Prior to the start of the task, participants read a set of instructions asking them to *"Imagine yourself in the role of a marine tour guide. It is your job to keep*

tourists safe from all dangerous animals. Your boat is about to enter an area densely populated by octopuses that are known to be dangerous to humans". The instructions indicated that it was participants' task to *"look at the octopuses and learn which ones can bite you → press the Left key and which ones can sting you → press the Right key".* Stage 1 comprised the presentation of trial types Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, and Dx+ once per block over 12 blocks (96 trials). Each trial began with a fixation cross centrally located for 0.5 s. The picture of a stimulus was then presented in the centre of the screen for 5 s with the text *Left: Bite, Right: Sting* displayed below the image. After the participant's response, the feedback *Correct!* Or *Ooops! That was wrong* appeared in the centre of the screen for 1 s, followed by the picture of the stimulus and the text *This octopus can bite you!* (for participants for whom the octopus signalled the outcome bite) for 2 s. An example trial is depicted in **Figure 5**. The trial types were block randomised.

Figure 5

Temporal Layout of a Trial during the Acquired Equivalence Task



Stage 2 proceeded similarly. Trials A+, B-, C, and D were presented once per block over 12 blocks and block randomised (48 trials). During Stage 2 the stimuli retained their different eye features, but displayed neutral, white tentacles with no suckers. Revaluation trials A+/B- were followed by the corresponding feedback but no feedback was provided for test trials C/D. Instead, the text *The octopus escaped! No feedback…* appeared in the centre of the screen for 2 s. These trials were used to assess generalisation from A+ and B-. Participants received no indication that they had advanced to Stage 2 trials. Stimulus-outcome contingencies were counterbalanced to create eight counterbalancing sub-groups.

2.1.1.3.2 IDS/EDS.

At the start of the IDS/EDS, a computerised female voice delivered the set of instructions. Participants were told that *"This task will take around 7 minutes to complete. You can see two patterns. A rule exists telling you which one is correct. You need to try and discover this rule. At first, there is nothing to tell you which pattern will be correct. You have to guess and learn from the feedback. We will tell you whether the pattern you selected was the correct or the incorrect one. Try and use the feedback to help you discover the rule. Once it is clear that you know the rule it will be changed, but this will not happen very often. After it is changed, you will have to learn the new rule to continue being correct".* Participants were instructed to touch the rectangle they believed contained the target stimulus. Stimuli remained on the screen until the participant made a response. After six correct consecutive responses, the task moved to the following stage. The task was automatically terminated if participants failed to make six consecutive correct responses within 50 trials. Performance was expressed as errors-to-criterion. The cues were

counterbalanced so half of the participants received lines and the other half received shapes as the relevant dimension.

The IDS/EDS involved nine stages: (1) a simple discrimination (SD), in which participants had to choose from two stimuli of the same dimension (e.g., shape). (2) A simple reversal (SR), in which the target stimulus was reversed with regards to the previous stage. (3) A compound discrimination (C_D), in which the second stimulus dimension (e.g., line) was introduced and presented side by side. (4) A second compound discrimination (CD), in which stimuli from both dimensions were superimposed. (5) A compound reversal (CR), in which the target stimulus was reversed. (6) An intra-dimensional shift (IDS) during which a new exemplar from the previously relevant dimension became the target. (7) An intra-dimensional reversal (IDR), in which the target stimulus was reversed with regards to the previous stage. (8) An extra-dimensional shift (EDS), in which the target stimulus became a new exemplar from the previously irrelevant dimension and (9) an extra-dimensional reversal (EDR), in which the target stimulus was once again reversed.

2.1.1.3.3 p-IDS/EDS

Prior to the start of the task, participants read a set of instructions asking them to *"Imagine yourself in the role of a microbiologist in a lab. The person in charge of handling the samples made a mistake and mixed them all up".* The instructions indicated that it was participants' task to *"examine the cellular and molecular samples and decide whether they come from dinosaurs or common lizards".* Participants were instructed to press *q* for dinosaur and *z* for lizard samples, and they were asked to respond as accurately as possible. Stage 1 comprised the presentation of trial types Ap+, Aq+, Bp*, Bq*, Wα+, Wβ+, Xα* and Xβ* once per

block over six blocks (48 trials), with trials block randomised. The structure of trials was identical to that of the acquired equivalence task.

During Stage 2, participants were presented with stimuli Cr, Cs, Dr, Ds, Yλ, Yδ, Zλ, Zδ once per block over six blocks (48 trials). Each stimulus consisted of a new exemplar from the same dimensions presented during training. For half of the new compounds, the dimension that had been relevant during training continued to be relevant during Stage 2 (i.e., participants were required to perform an IDS). For the other half of the new compounds, the stimulus dimension that had been previously irrelevant became now informative (i.e., participants were required to perform an EDS).

For half of the participants, stimuli A, B, W and X were relevant during Stage 1. For the other half of participants, stimuli p, q, α and β were initially relevant. Each stimulus signalled a dinosaur or lizard trials depending on the counterbalancing sub-group. Stimuli and outcomes in Stage 2 were counterbalanced so all stimuli underwent and IDS or EDS shift depending on the counterbalancing subgroup, cancelling any possible differences in stimulus generalisation. In total, orthogonal counterbalancing resulted in 16 counterbalancing sub-groups.

2.1.1.4 Data Treatment and Analysis

Data for the IDS/EDS are reported separately for pre-shift (i.e., simple discrimination, simple discrimination reversal, simple compound discrimination, compound discrimination and compound discrimination reversal) and shift stages (i.e., IDS, EDS and the corresponding reversals). Analyses were conducted on the number of trials required to reach the criterion of six correct consecutive trials at each stage. In accordance with previous IDS/EDS studies (e.g., Jazbec et al., 2007;

Kempton et al., 1999), analyses were conducted conditionally; that is, only participants who completed a stage were included in the analysis of that stage. Participants who failed at any stage were excluded from the analyses of all subsequent stages.

The proportion of correct trials per block was computed for each stage of the acquired equivalence task. Acquired equivalence relies upon participants learning the initial stimulus-outcomes contingencies. To determine whether participants had learned the initial discrimination, the proportion of correct responses over the second half of training (blocks 7 to 12) was averaged to ensure responding was reliably above chance (.50). Although revaluation and test trials were intermixed, they were treated separately for analysis purposes. For a test trial to be correct, it meant participants had transferred their response to C and D based on the revaluation provided to the stimulus that had been trained as equivalent during Stage 1 of the task.

The proportion of correct trials per block was computed for each stage of the p-IDS/EDS task. After establishing no differences in learning to the different stimuli during Stage 1, we analysed the proportion of correct IDS vs. EDS trials.

The correlation between performance in the acquired equivalence and IDS/EDS tasks required one datum per participant for both tasks. In the acquired equivalence task, critical test trials were averaged to obtain a single datum per participant. Because the IDS/EDS provides separate measures of intra-dimensional and extra-dimensional performance per participant, a datum reflecting participants' IDS superiority was calculated as the difference in number of errors-to-criterion between the EDS and IDS stages. Higher numbers indicate a greater number of EDS than IDS errors; the anticipated IDS superiority. Zero indicates no difference in

number of errors-to-criterion between the two stages, and a negative number indicates a greater number of IDS than EDS errors. Data from one participant were removed from the correlation analysis of the acquired equivalence and IDS/EDS tasks, as they failed to progress to the EDS stage (see section 2.1.2.3).

Data were analysed in RStudio (RStudio Team, 2016) using analysis of variance (ANOVA), one-sample *t*-tests (against .5 chance) and Pearson's correlation. In this and all subsequent experiments, a criterion of statistical significance of *p* less than .05 was adopted. All correlations were one-tailed, since we anticipated a positive correlation. Effect sizes for ANOVAs and one-sample *t*-test are reported as partial eta squared and Cohen's d respectively. 90% confidence intervals (CI) are reported along effect sizes. When needed, degrees of freedom were adjusted using Greenhouse-geisser estimates. When reported, Bayes factors were calculated using the statistical software JASP (JASP Team, 2019) with the default priors. The Bayes factor (BF) indicates how much more likely the data are under the alternative model. A BF above three is considered to support the alternative model. That is, the data are three times more likely to occur under the alternative model than under the null model. Any BF above eight is considered to be substantial support for the alternative model (e.g., Jarosz & Wiley, 2014; Jeffreys, 1961; Quintana & Williams, 2018).

2.1.2 Results and Discussion

2.1.2.1 p-IDS/EDS

Our pilot IDS/EDS task failed to show an overall superiority in IDS compared to EDS trials. Nevertheless, we report the results from this task and briefly discuss some of the limitations that might have resulted in a failure to observe the anticipated difference in performance. Although completed last by participants, the

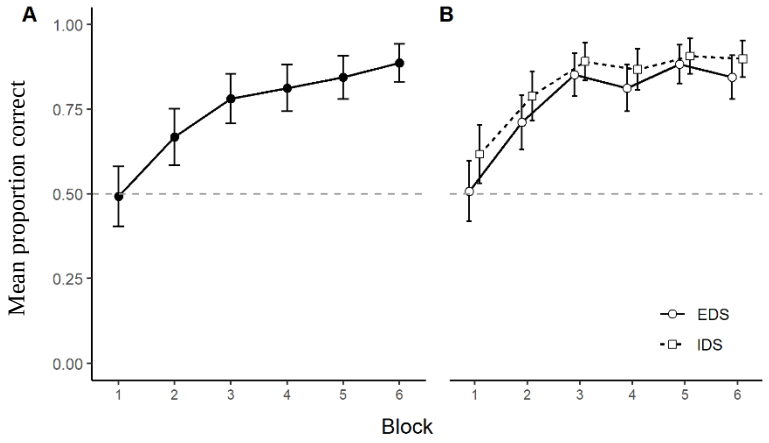
pilot IDS/EDS task is reported first as a standalone task, before reporting the results from the configural acquired equivalence and CANTAB IDS/EDS tasks and the correlation between the two. The reader may choose to skip this section.

Stage 1 learning data are summarised in **Figure 6(a)**. The data suggest participants learned the initial discrimination progressively, reaching and maintaining levels close to asymptote from the second half of training. An ANOVA with the within-subjects factors of outcome and block and the between-subjects factor of relevant dimension (cell walls/molecule bonds or cell organelles/molecule shape) confirmed this observation. The analysis yielded a main effect of block, $F(3.55, 213) = 33.72$, $p < .001$, $\eta_\rho^2 = .36$, 90% CI [0.27, 0.42] but no main effect of outcome, $F(1, 60) = 0.23$, $p = .632$, $\eta_\rho^2 = .004$, 90% CI [0.00, 0.06] or, importantly, relevant dimension, $F(3, 60) = 2.36$, $p = .080$, $\eta_\rho^2 = .11$, 90% CI [0.00, 0.20]. These results suggest that learning proceeded similarly across all dimensions during Stage 1. Any differences observed during Stage 2 could therefore not be attributed to a preference for any stimulus dimension during training. None of the interactions were significant (smallest $p = .051$ for the interaction between block and relevant dimension). In light of the Stage 1 results, data were collapsed over outcome for all subsequent analyses.

Critical Stage 2 data are summarised in **Figure 6(b)**. Inspection of the data suggests that mean performance was numerically better in IDS than EDS trials, albeit marginally. However, the mixed ANOVA with within-subjects factors of shift (IDS vs. EDS) and block and the between-subjects factor of relevant dimension revealed no reliable main effect of shift, $F(1, 140) = 2.32$, $p = .139$, $\eta_\rho^2 = .07$, 90% CI [0.00, 0.07], suggesting participants did not perform differently during IDS and EDS trials.

Figure 6

Collapsed Mean Performance for Stage 1 and IDS and EDS Trials during Stage 2 of our Pilot IDS/EDS

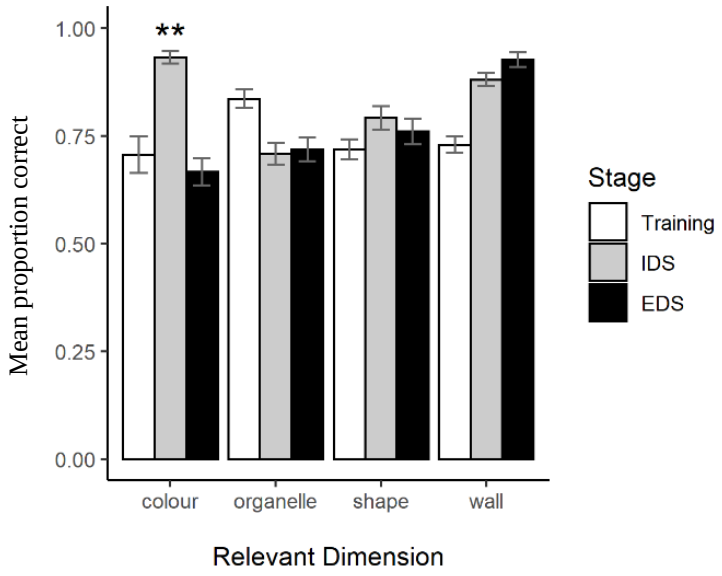


Note. (A) Each block comprised the presentation of compound cues AP, AQ, BP, BQ, Wα, Wβ, Xα and Xβ followed by feedback. (B) Each block comprised revaluation trials A and B followed by feedback. (C) Each block comprised trials CR, CS, DR, DS, Yλ, Yδ, Zλ, and Zδ followed by feedback. Half of these trials required participants to perform an IDS. The other half an EDS. Vertical bars represent standard errors of the mean. The horizontal dashed line represents chance performance. IDS and EDS trials intermixed during Stage 2 of the task.

The analysis revealed a significant interaction between shift and dimension, which is illustrated in **Figure 7**. The source of this 2-way interaction was examined using simple main-effects analysis with separate error terms at each level of the relevant dimension. This revealed that performance in the IDS was reliably better than in EDS trials only when the relevant dimension was colour, $F(1,10) = 11.16$, $p = .007$. The main effect of shift was not reliable in any of the other dimensions (smallest $p = .111$ for the relevant dimension of molecule shape). It is worth noting that the superiority in performance in our task cannot be attributed to a preference for the colour dimension in general. At least in as far as the acquisition of the discrimination proceeded similarly for all dimensions during Stage 1.

Figure 7

Mean Performance for All Stages of the Pilot IDS/EDS per Relevant Dimension



Note. Error bars represent SEM. ** indicates a p < .01.

The analysis showed a main effect of block, $F(3.03, 84.79) = 27.66$, $p < .001$, $\eta_\rho^2 = .50$, 90% CI [0.35, 0.58], reflecting participants' progressive learning. It also showed a main effect of relevant dimension, $F(3, 28) = 3.74$, $p = .022$, $\eta_\rho^2 = .29$, 90% CI [0.03, 0.43]. This reflected a better overall performance in the cell wall dimension ($M = .90$) compared to the cell organelle ($M = .71$) and molecule shape ($M = .70$) dimensions ($p = .003$ and .041, respectively), which, again, cannot be attributed to any preferences during Stage 1. No other interactions were significant (smallest $p = .337$ for the interaction between block and relevant dimension).

We failed to replicate the anticipated IDS superiority using a pilot IDS/EDS task that addressed some of the problems of the CANTAB IDS/EDS. Specifically, our task used a fully counterbalanced design and assessed performance in IDS and EDS, overall, at the same point in time. Whilst these results seem to, at least

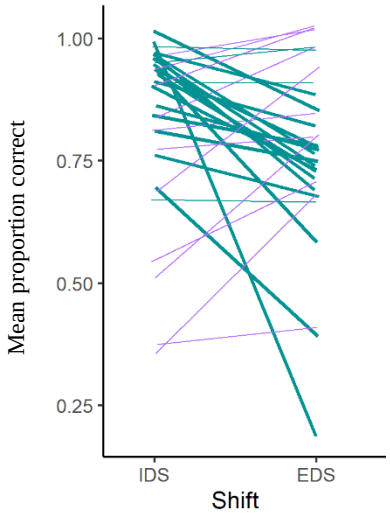
partially, challenge previous results in IDS/EDS, it is necessary to mention some of the problems that might have resulted in this failure to observe differences. From a quick look back at **Figure 7**, it is evident that participants performed as anticipated only when the relevant dimension was colour. This suggests that our pilot task was sensitive enough to detect these potential differences in performance. Evidence suggests that colour is easier to learn and discriminate than other stimulus dimensions like shape or orientation (e.g., Baxter & Gaffan, 2007; Mackintosh & Little, 1969). In an IDS/EDS experiment with rhesus monkeys, Baxter and Gaffan (2007) reported an interaction between the stimulus dimension that was relevant to solve the discrimination during training and the stimulus dimension that became relevant during IDS/EDS, despite no evident preference for any particular stimulus dimension during training. Similar dimension asymmetries could have occurred in our pilot task, with colour showing an enhanced performance compared to the other three, black and white, stimuli dimensions, even if this was not evident during training.

Additionally, the amount of training to learn the initial, relatively simple, discrimination might have been excessive. It is worth noting at this point that other IDS/EDS tasks, like the CANTAB IDS/EDS, require participants to make only six consecutive correct responses before moving to the following stage. Further inspection of participants' individual performance, illustrated in **Figure 8**, showed that out of the 32 participants, 17 participants performed in the expected direction (i.e., IDS > EDS), 11 participants performed in the opposite direction (i.e., EDS > IDS) and four participants performed identically in both tasks (i.e., IDS = EDS). For some of these participants, the initial extensive training might have allowed them to learn similarly about both stimulus dimensions, rather than learn about the relevant

and ignore the irrelevant dimension, influencing subsequent performance during IDS and EDS trials. In light of these results, the focus turned toward the configural acquired equivalence task and the CANTAB IDS/EDS.

Figure 8

Illustration of participants' individual performance during IDS and EDS trials



Note. Thicker green lines represent the 17 participants that perform in the expected direction (i.e., IDS > EDS). Thinner purple lines represent the 15 participants that perform either in the opposite direction (i.e., EDS > IDS) or identically in both shifts (i.e., IDS = EDS).

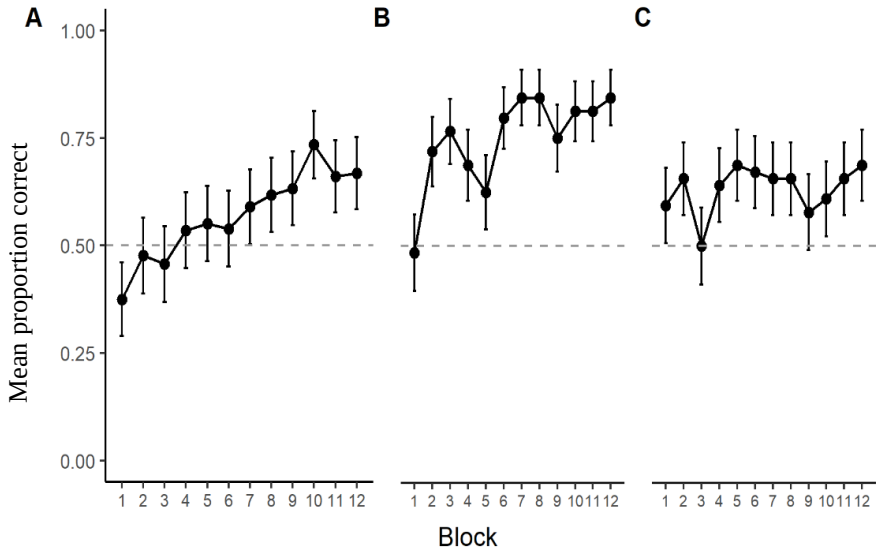
2.1.2.2 Acquired equivalence

Experiment's 1 Stage 1 data are summarised in **Figure 9(a).** The data indicate accuracy was rather low during the first half of training. However, participants' performance was reliably above chance during the second half of Stage 1, $t(31) = 4.55$, $p < .001$, $d = 0.80$, 90% CI [0.46, 1.13], suggesting participants simply took some time to reliably learn the stimulus-outcome contingencies. An

ANOVA with the factors of outcome and block confirmed this observation, yielding only a main effect of block, $F(6.93, 214.83) = 11.39$, $p < .001$, $\eta_\rho^2 = .24$, 90% CI [0.17, 0.31].

Figure 9

Collapsed Mean Performance for Stage 1 and Revaluation and Test trials of the Acquired Equivalence Task



Note. (A) Each block comprised the presentation of compound cues Aw, Ax, Bw, Bx, Cw, Cx, Dw and Dx followed by feedback. (B) Each block comprised revaluation trials A and B followed by feedback. (C) Each block comprised test trials C and D, which were not followed by feedback. Vertical bars represent standard errors of the mean. The horizontal dashed line represents chance performance. Revaluation and test trials were intermixed during Stage 2 of the task.

Data from Stage 2 were collapsed over outcome and split between revaluation (A and B) and test trials (C and D). Collapsed revaluation trials are shown in **Figure 9(b).** Participants showed an initial decline in performance upon presentation of single cues A and B, but accuracy recovered and reached a good level of discrimination. An ANOVA revealed a main effect of block during these trials, $F(5.72, 177.32) = 4.96$, $p < .001$, $\eta_\rho^2 = .14$, 90% CI [0.05, 0.19].

The data of central importance, those reflecting average performance during test trials C and D, are summarised in **Figure 9(c).** Participants demonstrated the acquired equivalence effect, by transferring responding from A to C and from B to D with no explicit feedback. Participants' accuracy started at a good level and declined in the first few blocks before recovering and maintaining good levels of discrimination again. A one-sample *t*-test confirmed that participants' overall discrimination was reliably above chance despite the absence of explicit feedback, $t(31) = 3.71$, $p < .001$, $d = 0.66$, 90% CI [0.33, 0.97]. That is, participants showed the acquired equivalence effect.

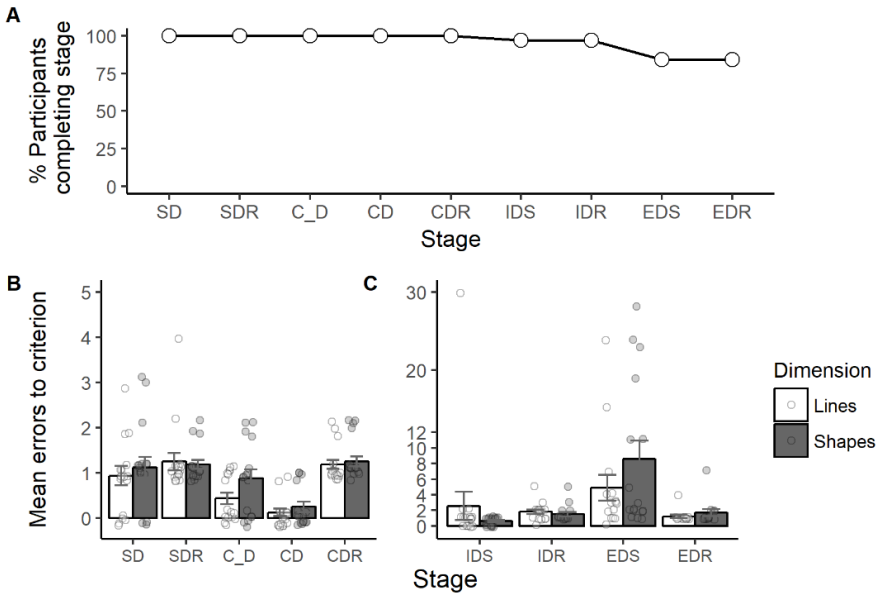
2.1.2.3 IDS/EDS

The percentage of participants that completed each stage of the IDS/EDS procedure is presented in **Figure 10(a)**. One participant failed the IDS and four participants failed the EDS, which resulted in their not progressing to any further stage of the procedure.

Error rates from the pre-dimensional-shift stages of the task per relevant dimension, completed by all 32 participants, are summarised in **Figure 10(b)**. The data show that participants mastered the discrimination in these stages rapidly, with a mean number of errors-to-criterion close to one. An ANOVA with a within-subjects factor of stage and between-subjects factor of relevant dimension revealed a reliable main effect of stage, $F(3.04, 91.20) = 16.31$, $p < .001$, $\eta_p^2 = .35$, 90% CI [0.21, 0.45]. This reflected that participants made even fewer errors in the compound discrimination than in any of the other stages. The analysis revealed no other main effects or interactions (smallest $p = .145$ for the main effect of dimension).

Figure 10

Percentage of Participants Completing the Task, Pre and Post-Shift Stages



Note. (A) Percentage of participants completing each stage of the IDS/EDS task. (B) Mean number of errors-to-criterion in each pre-shift stage. (C) Mean number of errors-to-criterion in each shift stage (SD = simple discrimination, SDR = simple discrimination reversal, C_D = compound discrimination with dimensions side by side, CD = compound discrimination dimension superimposed, CDR = compound discrimination reversal, IDS = Intra dimensional shift, IDR = intra dimensional reversal, EDS = extra dimensional shift, EDR = extra dimensional reversal). Semi-transparent circles show participants' individual performance. Error bars represent SEM.

Crucial data from the dimensional shift-stages, in **Figure 10(c),** show that participants demonstrated the anticipated IDS superiority. Inspection of the data reveals a general increase in the mean number of errors-to-criterion, particularly evident in the EDS. This observation was confirmed by an ANOVA with a within-subjects factor of stage and between-subjects factor of relevant dimension, which yielded a main effect of stage, $F(1.26, 31.5) = 9.84$, $p < .001$, $\eta_\rho^2 = .28$, 90% CI [0.07, 0.45]. Further examination of these data revealed that participants had a higher

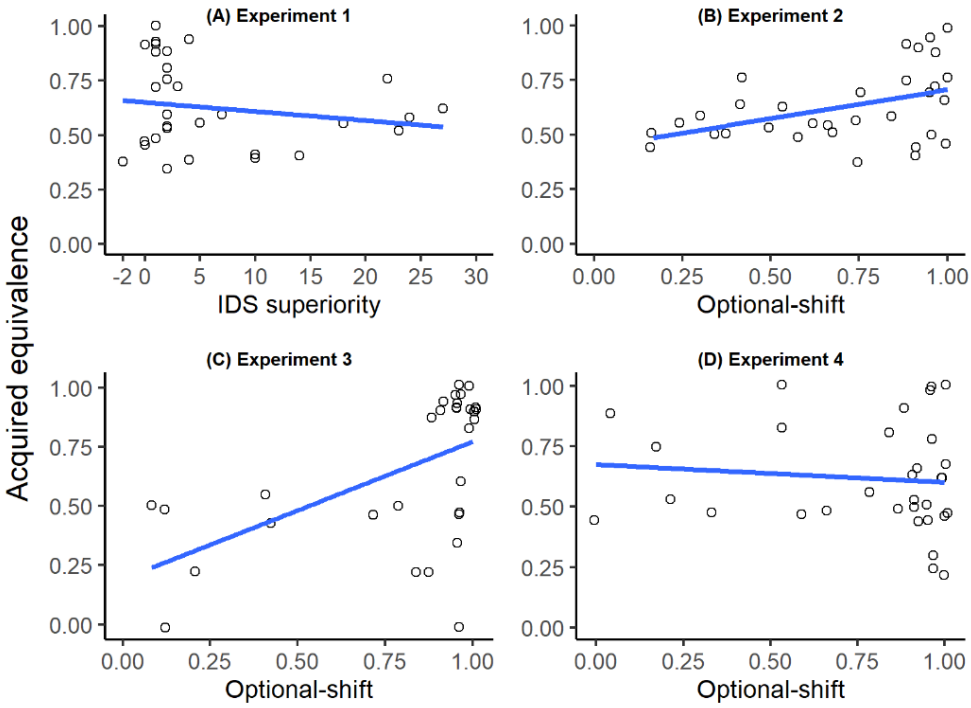
number of errors-to-criterion in the EDS than in any other stage. Importantly, significantly more errors-to-criterion in the EDS than in the IDS, $t(30) = 4.12$, $p < .001$, $d = 0.74$, 90% CI [0.40, 1.07], reflecting the expected IDS superiority effect. No other main effect or interactions were reliable, (smallest $p = .561$ for the main effect of dimension).

Following Honey et al.'s claims that a single three-layered network could account for the discrimination observed in configural acquired equivalence and attentional set, we looked at the relationship between participants' individual overall test performance in both tasks of Experiment 1.

The datum for each participant for the acquired equivalence task reflects participants' mean accuracy during critical test trials C and D. The acquired equivalence effect was correlated with the IDS superiority effect (EDS error rate minus IDS error rate). Data from one participant were removed from both tasks, as they failed to progress to the EDS stage. These data are summarised in **Figure 11(a)**.

Figure 11

Correlation between Configural Acquired Equivalence and Attentional Set for Experiment 1, Experiment 2, Experiment 3 and Experiment 4



Note. (A) Scatterplot illustrating the relationship between performance in acquired equivalence and IDS superiority in Experiment 1, where higher numbers in the IDS superiority indicate a greater number of EDS than IDS errors. Data from one participant were removed due to a failure to proceed to the EDS. (B) and (C) illustrate the positive relationship between performance in acquired equivalence and optional-shift in Experiment 2 and Experiment 3, respectively. (D) illustrates the lack of a positive correlation between performance in both tasks in Experiment 4.

We anticipated participants who performed more accurately in the acquired equivalence task would show a greater IDS superiority. However, the correlation between these two measures failed to reveal the expected positive relationship, Pearson's $r(29) = -.17$, $p = .821$. Because traditional null-hypothesis significance testing does not allow us to discern whether the non-significance is due to data insensitivity or to an actual lack of relationship between configural acquired equivalence and IDS/EDS, we run a Bayesian correlation analysis to obtain the

Bayes factor (BF) associated with the null model. The analysis returned BF = 8.02 in favour of the null model, suggesting the data substantially supported the absence of a correlation between both tasks.

This pattern of results seems not to accord well with Honey et al.'s (2010) claims. However, it is noticeable that whilst the configural acquired equivalence task provided a direct measure of the effect for each individual, the IDS superiority was derived from the differences between trials to criterion in IDS and EDS. Additionally, it is worth remembering the issues with IDS/EDS previously mentioned, such as the IDS and EDS shifts being measured at different time points. It is possible that the lack of correlation reflects the differences between the way acquired equivalence and attentional set were measured, rather than any intrinsic differences. The next experiment aimed to investigate this possibility.

2.2 Experiment 2

The findings from Experiment 1 showed that we could demonstrate the acquired equivalence and IDS superiority in our cohort of participants. The novelty in these findings involved the completion of these tasks in a single session and the direct comparison between test performance in both. However, there are some limitations associated with the differences in measuring configural acquired equivalence and attentional set using IDS/EDS. Specifically, deriving a single measure of attentional set from two separate measures (errors-to-criterion in IDS and EDS). Experiment 2 was intended to solve this problem by testing the relationship between configural acquired equivalence and a different measure of attentional set; optional-shift. The optional-shift task, summarised in **Table 5**, was based on Duffaud et al. (2007). In this task, participants were also asked to predict the outcome that followed each compound stimulus (Aw+, Ax+, Bw-, Bx-). Just like in IDS/EDS, one dimension of the compound consistently signalled a given outcome (e.g., A+ and B-); the other stimulus dimension (w and x) signalled either outcome equally. Participants could, therefore, learn that stimuli varied in their predictive accuracy. During Stage 2, new compound stimuli Cy+, Dz-, Cz and Dy were presented in an intermixed fashion. Critically, in this stage, all individual stimuli were equally predictive of the outcome. However, the optional-shift effect was evidenced by the presentation of test compounds Cz and Dy with no explicit feedback. We expected participants to respond to these compounds based on the stimulus dimension that had been predictive during Stage 1 (e.g., **A**w+, **A**x+, **B**w-, **B**x- → Cy+, Dz-, **C**z+, **D**y-).

Table 5

Experimental designs for the acquired equivalence and optional-shift tasks in Experiment 2, Experiment 3 and Experiment 4

Acquired equivalence			Optional-shift	
Stage 1		Stage 2 Revaluation and test	Stage 1	Stage 2
Aw +	Ax –	A +	Aw +	Cy +
Bw –	Bx +	B –	Ax +	Dz –
Cw +	Cx –	**C ?**	Bw –	**Cz ?**
Dw –	Dx +	**D ?**	Bx –	**Dy ?**

Note. Each task had a visual and audio-visual version. In the visual version of Experiment 2, letters A-D represent different snake tails and w/x represent different skin patterns. + and – represent outcomes poisonous and harmless. In the audio-visual version, letters A-D represented different computerized tones and w/x represented different cartoon robots. + and – represent dangerous and friendly, respectively. ? indicates the absence of feedback. Trials during Stage 2 were intermixed. Although with a different set of stimuli, the acquired equivalence task follows the same design than in Experiment 1.

We modelled the optional-shift task as closely as possible to the configural acquired equivalence task to improve the comparison of attentional set and acquired equivalence over that of Experiment 1. Unlike in Experiment 1, where configural cartoon animals were used to measure acquired equivalence and lines and shapes to measure IDS superiority, stimuli were counterbalanced so that both tasks used the same set of stimuli (i.e., snake cartoons and robot cartoons and tones). This allowed us to control for any possible intrinsic differences between stimuli. The number of stages in the tasks were matched, with participants receiving an identical number of Stage 2 trials in both tasks. The order in which the tasks were presented was also counterbalanced and, crucially, the measure of attentional set in the optional-shift task was not confounded with temporal factors: because revaluation and test trials in Stage 2 were intermixed, test trials were, on average, presented at identical points in time. Unlike the IDS/EDS task in Experiment 1, it provided us with a single measure

of attentional set per participant, allowing for a direct comparison between performance in both tasks. Experiment's 2 critical data replicated the acquired equivalence effect and demonstrated an attentional set effect with optional-shift. Having addressed some important experimental inconsistencies between the tasks, the data revealed the anticipated correlation in performance between both tasks.

2.2.1 Method

2.2.1.1 Participants

32 students from the University of Nottingham participated (10 men and 22 women, M_{age} = 26.06, *SD* = 4.08, range: 21–34). Students received module credits or an allowance for their participation. Participants were recruited using posters and the School of Psychology online booking system.

2.2.1.2 Apparatus & Materials

Stimuli consisted of images of snakes with different combinations of tails (A-D: fork looking tail, pointy tail, axe looking tail and rattle tail) and skin patterns (w and x: spotty skin and stripe skin) for the visual versions of the tasks. For example, trial Aw represents a snake with a rattle tail and a spotty skin pattern. In the audio-visual version, stimuli consisted of different computerized tones (A-D) and images of robots (w and x). Both the acquired equivalence and optional-shift tasks had visual and audio-visual version. Images were 10 (width) x 8 (height) cm presented on a grey background. Tones were generated using version 2.3.0 of Audacity (2019) and differed in their amplitude and frequency (tone A: linear chirp effect, 450 Hz to 800 Hz. Tone B: Paulstretch effect 900 Hz. Tone C: Wahwah effect, 300 Hz. Tone

D: Wahwah effect, 1000 Hz). Participants wore a pair of headphones (Panasonic RP-HT225) during the audio-visual version of the tasks only.

Each of the possible outcomes was presented in text boxes on the same grey background. The outcomes were *poisonous* and *harmless* for the visual and *friendly* or *evil* for the audio-visual version of each task. The contingencies between the different stimuli and outcomes as well as the task versions were counterbalanced orthogonally to create different counterbalancing subgroups.

2.2.1.3 Procedure

Participants completed the acquired equivalence and optional-shift tasks. 16 participants performed the acquired equivalence before the optional-shift; the remainder performed the tasks in the alternative sequence. Each participant completed a different version of each task. For example, the visual version of the acquired equivalence task followed by the audio-visual version of the optional-shift task. This ensured stimuli and outcomes were different across tasks for every participant. All participants read a standard instruction sheet that emphasized the participants' right to terminate the task at any time. The experimenter left the room after ensuring participants had understood the tasks and returned only to set up the second task, before leaving again until the end of the experiment. During the visual version of the tasks, participants were presented with on-screen instructions asking them to *"Imagine yourself in the role of a rainforest tour guide. It is your job to make sure tourists are safe during the duration of the tour. You are about to enter an area densely populated by snakes, some of which are known to be dangerous to humans. It is you task to look at the snakes and learn which ones are poisonous → press the Left key and which snakes are harmless → press the Right key".* In the

audio-visual version of the task, participants were told *"It is the year 2250 and robots have risen against humanity! Fortunately, not all robots present a risk to humans. You will be presented with some robots and robot noises simultaneously. It is your task to learn which robots are dangerous → press the Q key and which ones are friendly → press the Z key".*

In the optional-shift task, Stage 1 comprised the presentation of trial types Aw+, Ax+, Bw-, and Bx- once per block over 12 blocks (48 trials). Trials proceeded as in the acquired equivalence task: stimuli appeared in the centre of the screen for 5 s with the text *Q: Poisonous, Z: Harmless* displayed below the image (for participants assigned to the visual version of the task). After the participant's response, the feedback *Correct!* Or *Ooops! That was wrong* appeared in the centre of the screen for 1 s, followed by the picture of the stimulus and the text *This snake is poisonous* (for participants for whom the snake was poisonous) for 2 s. Trials in the audio-visual version were identical but the visual and auditory stimuli were presented simultaneously during 5 s. The trial types were block randomised. Keyboard responses in the optional-shift and acquired equivalence tasks were spatially orthogonal (left/right vs. q/z). Stage 2 comprised the presentation of new compounds Cy+, Dz-, Cz, and Dy once per block over 12 blocks block randomised (48 trials). Test trials Cz and Dy were not followed by feedback. All unspecified details are identical to those in Experiment 1.

2.2.1.4 Data Treatment and Analysis

The proportion of correct trials per block was computed for each stage of the acquired equivalence and optional-shift tasks. Acquired equivalence and optional-shift rely upon participants learning the initial stimulus-outcomes contingencies. To

determine whether participants had learned the initial discrimination, the proportion of correct responses over the second half of training (blocks 7 to 12) was averaged; to ensure responding was reliably above chance (.50) towards the end of training. Although they were intermixed during the task, revaluation trials in Stage 2 were analysed separately from test trials with no feedback. For a test trial to be correct in the optional-shift task, it meant participants had demonstrated a bias for the dimension established as relevant during Stage 1. Test trials from both tasks were averaged to obtain a single datum per participant and correlated to determine the relationship between performance in both tasks.

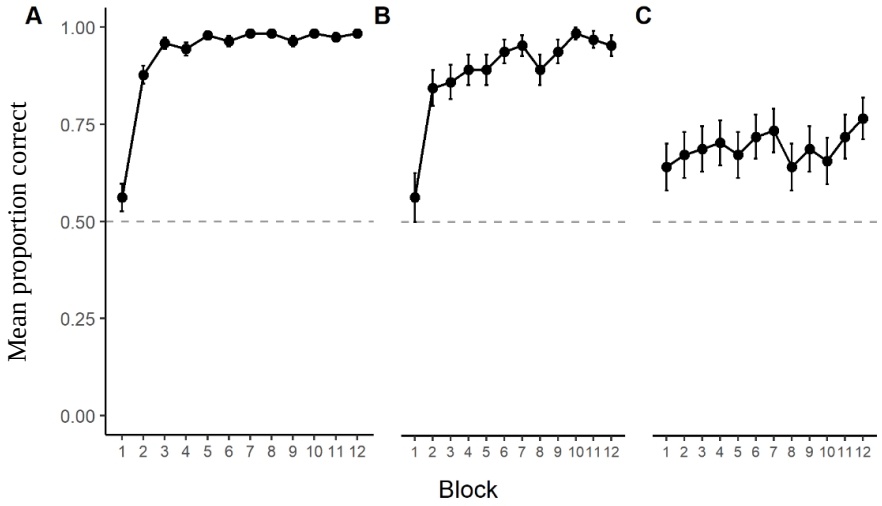
2.2.2 Results and Discussion

2.2.2.1 Optional-shift

Examination of the Stage 1 data, summarised in **Figure 12(a)**, indicates that participants were quick to learn, needing only two blocks to master the initial discrimination. A one-sample *t*-test confirmed participants' reliable performance above chance during the second half of training, $t(31) = 100.85$, $p < .001$, $d = 17.83$, 90% CI [13.95, 21.49]. This observation was supported by an ANOVA with the factors of version (visual vs audio-visual) and block, which yielded only a main effect of block, $F(5.36, 196.14) = 14.93$, $p < .001$, $\eta_p^2 = .31$, 90% CI [0.21, 0.36].

Figure 12

Collapsed Mean Performance for Stage 1 and Revaluation and Test trials of the Optional-Shift Task



Note. (A) Each block comprised the presentation of compound cues Aw, Ax, Bw and Bx followed by feedback. (B) Each block comprised the presentation of new trials Cy and Dz followed by feedback. (C) Each block comprised test trials Cz and Dy, which were not followed by feedback. Vertical bars represent standard errors of the mean. The horizontal dashed line represents chance performance. Revaluation and test trials were intermixed during the task.

Data from Stage 2 were collapsed over version of the task and outcome and split between new compound trials (Cy+ and Dz-) and test trials (Cz and Dy). **Figure 12(b)** shows the discrimination of the new compounds during Stage 2. The presentation of novel compound stimuli resulted in a decline in performance during the first block of Stage 2, but performance recovered quickly. An ANOVA confirmed this observation, revealing a main effect of block, $F(5.94, 184.14) = 9.93$, $p < .001$, $\eta_p^2 = .24$, 90% CI [0.13, 0.30].

Trials Cz and Dy, which denote the crucial measure of attentional set and are summarised in **Figure 12(c),** confirm participants' bias for the dimension established as relevant during Stage 1 despite the absence of any explicit feedback. A one-

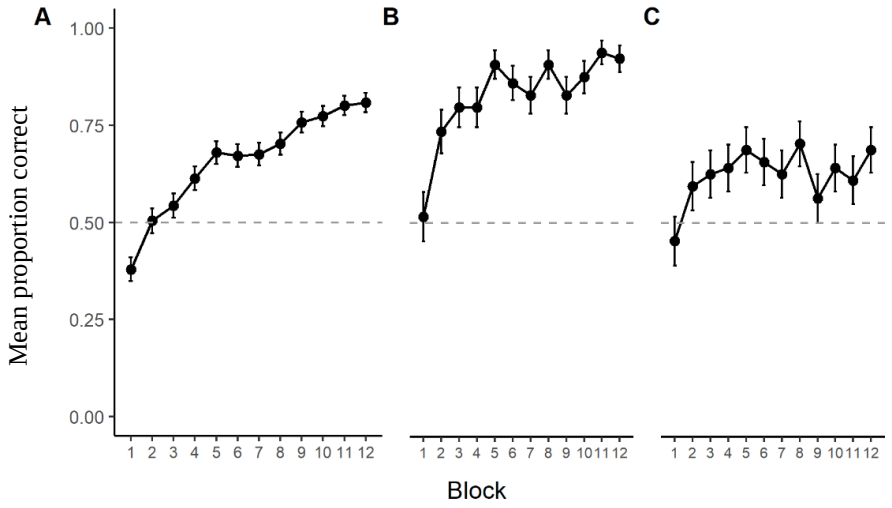
sample *t*-test against chance level performance supported this observation, $t(31) = 3.82$, $p < .001$, $d = 0.68$, 90% CI [0.35, 0.99].

2.2.2.2 Acquired equivalence

Initial examination of the data for Stage 1, summarised in **Figure 13(a)**, shows a noticeable improvement in initial performance as compared to Experiment 1. Participants acquired the initial discrimination progressively and were performing reliably above chance from the second half of Stage 1. A one-sample *t*-test confirmed this observation, $t(31) = 9.74$, $p < .001$, $d = 1.72$, 90% CI [1.25, 2.17]. An ANOVA with the factors of version of the task and block confirmed this observation, yielded an effect of block, $F(6.94, 214.14) = 12.78$, $p < .001$, $\eta_p^2 = .29$, 90% CI [0.21, 0.33].

Figure 13

Collapsed Mean Performance for Stage 1 and Revaluation and Test trials of the Acquired Equivalence Task



Note. (A) Each block comprised the presentation of compound cues Aw, Ax, Bw, Bx, Cw, Cx, Dw and Dx followed by feedback. (B) Each block comprised revaluation trials A and B followed by feedback. (C) Each block comprised test trials C and D, which were not followed by feedback. Vertical bars represent standard errors of the mean. The horizontal dashed line represents chance performance. Revaluation and test trials were intermixed during the task.

Data from Stage 2 were collapsed over version of the task and outcome and split between revaluation (A+ and B-) and test trials (C and D). **Figure 13(b)** shows revaluation trials during Stage 2. It is clear that the presentation of only one stimulus of the compound resulted in an initial decline in performance. However, performance recovered quickly, reaching and maintaining levels close to asymptote. This description of the data was supported by an ANOVA, which yielded a main effect of block, $F(6.38, 191.40) = 8.11$, $p < .001$, $\eta_p^2 = .21$, 90% CI [0.11, 0.27].

Data from test trials, summarised in **Figure 13(c)**, confirms that participants replicated the acquired equivalence effect observed in Experiment 1. Just like during

revaluation trials, performance started at chance level but quickly increased and maintained good levels of discrimination despite the absence of any explicit feedback. A one-sample *t*-test confirmed participants' overall discrimination was reliably above chance during test trials, $t(31) = 4.25$, $p < .001$, $d = 0.75$, 90% CI [0.42, 1.08].

Following Honey et al.'s claims, we looked at the relationship between participants' individual overall performance during test trials in both tasks of Experiment 2, summarised in **Figure 11(b)**. The correlation between these measures was reliable, Pearson's $r(30) = .43$, $p = .013$. To obtain further support for this positive correlation we conducted a Bayesian correlation. The correlation yielded a BF = 8.17 in favour of the alternative model, providing substantial support for a positive correlation between critical performance in our acquired equivalence and optional-shift tasks. A Fisher's z transformation confirmed that the correlations between configural acquired equivalence and IDS/EDS in Experiment 1 and configural acquired equivalence and optional-shift in Experiment 2 were reliably different $z = -2.38$, $p = .017$.

Unlike results from Experiment 1, these findings accord with Honey et al.'s (2010) claims and are theoretically anticipated by Robinson et al.'s (2019) formal instantiation of the model, according to which individual differences in the connections between output and hidden units would affect both processes. The notable experimental differences between the IDS/EDS and the optional-shift tasks could potentially explain the discrepancy in the relationship between configural acquired equivalence and attentional set. For example, the use of similar sets of stimuli across tasks might have accounted for any differences in how engaging participants found stimuli to be, or how easy to discriminate. By measuring

performance to stimuli that were presented, on average, at the same time, we might have eliminated any temporal confounds. In any case, the positive correlation, as suggested by the BF, and the confirmation that the correlation coefficients were reliably different in both experiments strongly suggest a positive relationship between performance in configural acquired equivalence and attentional set when tasks are closely matched to allow for direct comparison. Experiment 3 aimed to confirm these findings.

2.3 Experiment 3

Experiment 2 demonstrated configural acquired equivalence and attentional set using an optional-shift task in human participants. The results demonstrated a positive relationship between performance in these two tasks, a finding supportive of Honey et al.'s (2010) assertion that a common mechanism could govern both forms of learning. Experiment 3 intended to replicate results from Experiment 2. Experiment 3 also measured participants' gaze with the use of an eye-tracker to explore the role of predictiveness in overt attention in our attentional set task.

In human contingency learning tasks, such as an IDS/EDS or optional-shift, participants usually learn more rapidly about predictive stimuli, those that reliably signal the outcome of a trial, than about nonpredictive stimuli. Although these tasks reflect how predictiveness increases the rate of learning about a particular stimulus, they do not allow us to study the relationship between predictiveness and overt attention.

A broadly accepted way of capturing changes in overt attention is by analysing eye movements, which are tightly coupled with shifts in attention (e.g., Deubel & Schneider, 1996; Just & Carpenter, 1980; Rayner, 1998). Greater eye gaze dwell times to predictive over nonpredictive stimuli have been consistently found (e.g., Haselgrove et al., 2016; Le Pelley et al., 2011; Le Pelley et al., 2013). This research has typically focused on contingency learning tasks that require participants to learn about the occurrence of new outcomes. For example, Le Pelley et al. (2011) measured participants' eye gaze in a learned predictiveness task. In an initial stage, participants had to learn the contingencies between a series of compound stimuli-outcomes in which one dimension of the compound was predictive of the outcome and a second dimension was nonpredictive (e.g., **A**v-O1, **A**w-O1, **C**x-O2, **C**y-O2).

In a subsequent stage, compound stimuli from a previously predictive and previously nonpredictive dimension were paired with new outcomes (e.g., Ax-O3, Cv-O4). The eye gaze data showed a greater overt attention to the stimulus dimension that had been initially predictive of the outcome, even when compounds were objectively equally predictive. To our knowledge, however, no study has assessed overt attention and learning in an optional-shift task, where outcomes remain constant throughout the task.

A common way of analysing eye-tracking data is by averaging fixation times in each region of interest (ROI) (see Lai et al., 2013). In the optional-shift task, we expected eye-gaze data to reflect the stimuli's objective differences in predictiveness during the initial training stage, with greater fixation times for the relevant stimulus dimensions. Because revaluation and test trials were intermixed during Stage 2, and the new stimuli were objectively equally predictive, we reasoned that participants showing a greater eye-gaze bias for the previously predictive dimension during the first half of revaluation trials should also show a greater accuracy performance during the second half of test trials. The acquired equivalence task required participants to learn about specific combinations of stimulus dimensions, which were, individually, equally predictive. Thus, we did not expect to see any differences in dwell times between stimulus dimensions. Test data from Experiment 3 replicated findings from Experiment 2, and once again confirmed a positive correlation between performance in configural acquired equivalence and optional-shift.

2.3.1 Method

2.3.1.1 Participants, Apparatus & Stimuli, and Procedure

32 students from the University of Nottingham participated (14 men and 18 women, M_{age} = 21.56, *SD* = 2.10, range: 18-25). Students received module credits or an allowance for their participation. Participants were recruited using posters and the School of Psychology online booking system.

For eye-tracking purposes, the audio-visual version of the acquired equivalence and optional-shift tasks described in Experiment 2 was substituted with a second visual version, the one consisting of octopuses with different eyes (A-D) and tentacles (w/x) described in Experiment 1.

The experiment was run on a Tobii TX300 eye-tracker (Tobii Technology, Danderyd, Sweden) with a 51 (width) x 28 (height) cm monitor and a monitor-mounted eye tracker recording gaze at a resolution of 60 Hz that did not require a chin rest. Participants sat approximately 50 cm from the screen. Two regions of interest were established for each compound stimulus. ROIs had different sizes for the snake and octopus cartoons to accommodate for the differences in their shape. The ROI for dimension A-D were 3 cm x 3 cm in size for the octopuses' eyes and 3.5 cm x 3.5 cm for snake tails. The ROIs for dimension w/x were 9 cm (width) x 6 cm (height) for the octopuses' tentacles and 3.5 cm x 3.5 cm for the snake patterns. The eye-tracker recorded participants' pre-response gaze only. That is, it did not record participants' gaze during the feedback part of each trial. 16 participants performed the acquired equivalence task before the optional-shift task; the remainder performed the tasks in the alternative sequence. Each participant completed a different version of each task. For example, the octopus version of the acquired equivalence task followed by the snake version of the optional-shift task. Keyboard

responses were spatially orthogonal across tasks (left/right vs. q/z). All unspecified procedure details were identical to those of Experiment 2.

2.3.2 Results and Discussion

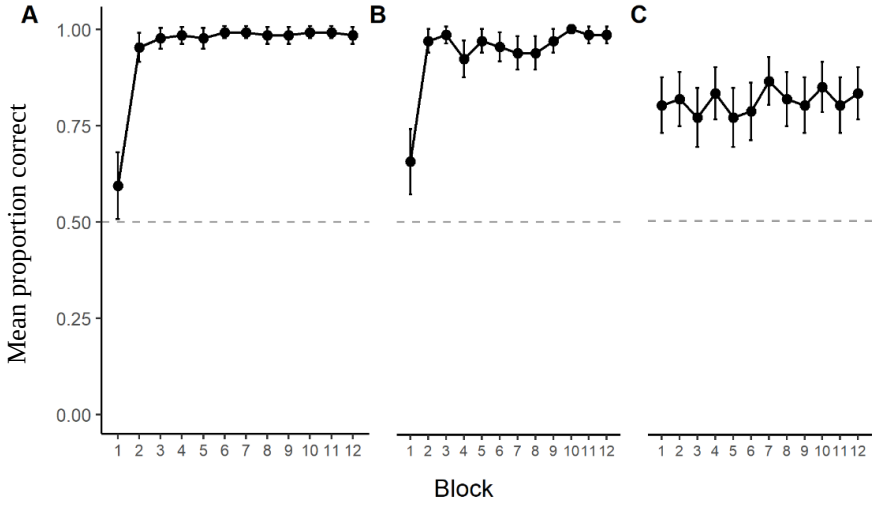
2.3.2.1 Optional-shift

2.3.2.1.1 Behavioural data

Data from Stage 1, shown in **Figure 14(a)** show that participants learned the stimulus-outcomes relationships. Just like in Experiment 2, accuracy reached asymptote after only two blocks and participants were evidently performing above chance during the second half of training, $t(31) = 126.83$, $p < .001$, $d = 22.42$, 90% CI [17.55, 27.02]. An ANOVA once again confirmed participants' acquisition of the discrimination with a main effect of block, $F(4.69, 147.34) = 21.76$, $p < .001$, $\eta_p^2 = .48$, 90% CI [0.41, 0.57].

Figure 14

Collapsed Mean Performance for Stage 1 and Revaluation and Test trials of the Optional-Shift Task



Note. (A) Each block comprised the presentation of compound cues Aw, Ax, Bw and Bx followed by feedback. (B) Each block comprised the presentation of new trials Cy and Dz followed by feedback. (C) Each block comprised test trials Cz and Dy, which were not followed by feedback. Vertical bars represent standard errors of the mean. The horizontal dashed line represents chance performance. Revaluation and test trials were intermixed during the task.

Figure 14(b) shows the discrimination to the new compounds Cy and Dz during Stage 2. Accuracy decreased as compared to the last block of the previous stage, but it recovered immediately. An ANOVA confirmed a main effect of block, $F(3.96, 122.76) = 13.41$, $p < .001$, $\eta_p^2 = .30$, 90% CI [0.18, 0.38].

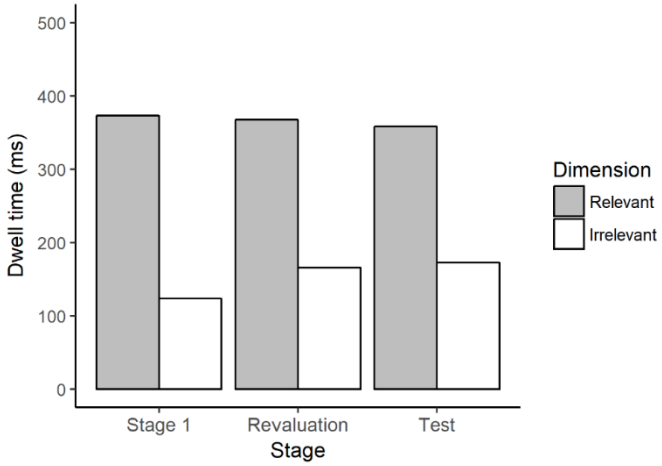
Accuracy to critical test trials Cz and Dy, shown in **Figure 14(c),** confirmed participants' bias toward the dimension that had been predictive during initial training. The results closely matched those from Experiment 2, and evidenced participants' ability to respond to these stimuli despite the lack of any explicit feedback. A one-sample *t*-test confirmed that participants' mean accuracy was reliably above chance, $t(31) = 5.88$, $p < .001$, $d = 1.04$, 90% CI [0.67, 1.40].

2.3.2.1.2 Eye-tracking data

Eye-tracking data, which were collapsed across all trials of each stage, reflected clear changes in gaze direction toward the predictive stimulus dimension. Data are summarised in **Figure 15** and are shown separately for relevant (A-D) and irrelevant (w, x, y, and z) dimensions.

Figure 15

Preresponse Dwell Time on Stimulus Dimensions during the Optional-Shift Task



Note. Relevant dimension refer to snake tails A-D and octopuses' eyes A-D, irrelevant dimension refer to skin patterns and tentacles w, x, y, and z.

From these data it seems clear that that participants biased their dwell times toward the relevant stimuli in all stages. An ANOVA with factors of dimension and stage revealed a significant main effect of dimension, $F(1, 32) = 22.40$, $p < .001$, $\eta_p^2 = .42$, 90% CI [0.19, 0.57], with greater dwell time on relevant than irrelevant stimuli, but no other main effects or interactions (smallest $p = .110$ for the main effect of stage). Preplanned paired *t* tests revealed that dwell time was reliably greater for the relevant than the irrelevant dimension during Stage 1, $t(31) = 6.38$, $p < .001$, d = 1.13, 90% CI [0.75, 1.49], and during Stage 2 for both the revaluation,

$t(31) = 4.23$, $p < .001$, d = 0.75, 90% CI [0.41, 1.07] and test trials, $t(31) = 3.30$, $p = .002$, $d = 0.58$, 90% CI [0.26, 0.89]. These data suggest that the initial predictiveness manipulation during Stage 1 influenced the way participants learned about the different stimuli during Stage 2, where all dimensions were equally predictive of the outcome.

For each participant, we calculated: (I) a single measure of attentional bias, given by the difference in mean dwell time to the relevant dimension minus mean dwell time to the irrelevant dimension across the first half (blocks 1-6) of the revaluation trials during Stage 2. (II) A single learning measure, given by the mean accuracy to test trials during the second half of Stage 2 (blocks 7-12). The correlation between these two measures was reliable $r(30) = .55$, $p < .001$.

The finding that bias in dwell times during initial revaluation trials, which preceded accuracy data later in the task, was positively correlated with performance in later test trials suggests a strong relationship between the two, and adds to previous demonstrations of the relationship between predictiveness and overt attention using eye-tracking (Aristizabal et al., 2016; Hogarth et al., 2010; Le Pelley et al., 2011).

2.3.2.2 Acquired equivalence

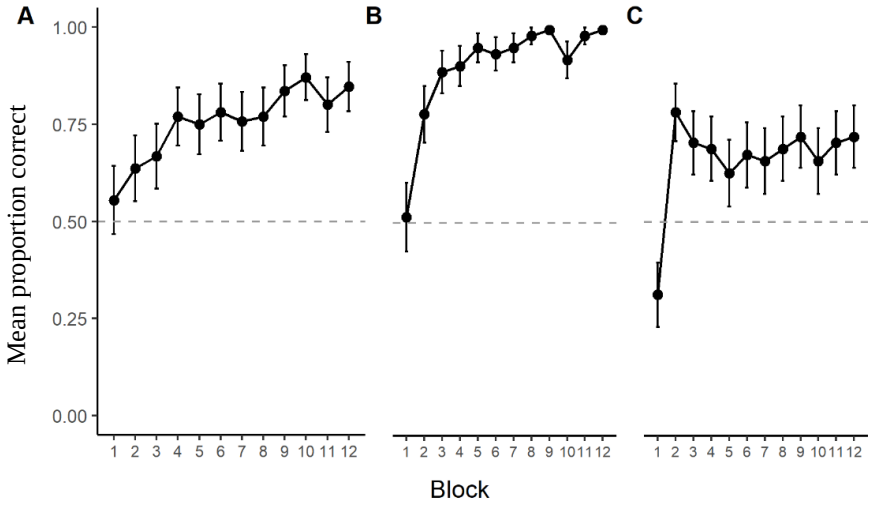
2.3.2.2.1 Behavioural data

Experiment's 2 Stage 1 data are summarised in **Figure 16(a).** An initial look at the data suggests that this cohort of participants performed remarkably well during training. As in Experiment 2, participants learned the contingencies between compound stimuli and outcomes progressively, demonstrating a performance reliably above chance during the second half of Stage 1 training (blocks 7 to 12),

$t(31) = 10.17$, $p < .001$, $d = 1.80$, 90% CI [1.32, 2.26]. An ANOVA revealed a main effect of block, $F(6.38, 191.40) = 12.72$, $p < .001$, $\eta_p^2 = .29$, 90% CI [0.19, 0.36], confirming participants' acquisition of the discrimination.

Figure 16

Collapsed Mean Performance for Stage 1 and Revaluation and Test trials of the Acquired Equivalence Task



Note. (A) Each block comprised the presentation of compound cues Aw, Ax, Bw, Bx, Cw, Cx, Dw and Dx followed by feedback. (B) Each block comprised revaluation trials A and B followed by feedback. (C) Each block comprised test trials C and D, which were not followed by feedback. Vertical bars represent standard errors of the mean. The horizontal dashed line represents chance performance. Revaluation and test trials were intermixed during the task.

Data from Stage 2 were collapsed over version of the task and outcome. Revaluation trials are shown in **Figure 16(b).** Visual examination of the data suggests that these trials proceeded much like in Experiment 2, with participants showing an initial decline in performance upon presentation of single cues A and B, but quickly mastering the new discrimination. An ANOVA confirmed this pattern of

results, yielding an effect of block, $F(4.51, 139.81) = 15.79$, $p < .001$, $\eta_p^2 = .34$, 90% CI [0.22, 0.41].

Data from test trials, shown in **Figure 16(c),** once again replicated the acquired equivalence effect obtained in Experiments 1 and 2**.** Accuracy in these trials started rather low but improved quickly and was consistently good until the end of the task, demonstrating participants' ability to transfer responding to test trials. A one-sample *t*-test confirmed that participants' discrimination was reliably above chance, $t(31) = 2.88$, $p = .003$, $d = 0.50$, 90% CI [0.20, 0.81].

As in Experiment 2, participants' individual performance during test trials in both tasks was correlated. The correlation between these measures was reliable, as showed in **Figure 11(c)**, Pearson's $r(30) = .54$, $p < .001$, replicating the results obtained in the previous experiment. An additional Bayesian correlation provided very strong evidence for a positive relationship between performance in configural acquired equivalence and optional-shift, with a BF = 62.95 in favour of the alternative model. A Fisher's z transformation confirmed that the correlation in performance observed in Experiment 3 differed significantly from the results obtained in Experiment 1, z = -2.93, $p = .003$. However, the correlation between acquired equivalence and attentional set in Experiments 2 and 3 did not differ significantly z = -0.55, $p = .582$, which allows us to confirm Experiment 3 replicated all effects found in Experiment 2.

2.3.2.2.2 Eye-tracking data

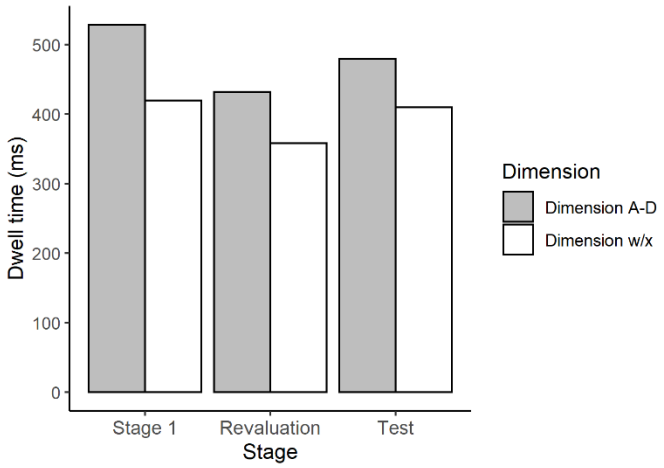
Eye-tracking data were collapsed across all trials of each stage. The eye-tracker was occasionally unable to register gaze location, resulting in missing gaze data. Missing data were infrequent and not systematically occurring on any particular

trial type. A minimum of 80% of eye movement had to be recorded from each participant, which resulted in 6 participants from different counterbalancing subgroups being removed.

Unlike in the optional-shift task, where there was an initial asymmetry in the predictiveness of each stimulus dimension, both dimensions were equally predictive in the acquired equivalence task. We thus anticipated no differences in average fixation time to any particular dimension. A look at the average dwell time data, summarised in **Figure 17**, confirms similar dwell times across dimensions.

Figure 17

Preresponse Dwell Time on Stimulus Dimensions during the Acquired Equivalence Task



Note. Dimension A-D refer to snake tails A-D and octopuses' eyes A-D. Dimension w/x refer to skin patterns and tentacles w, x, y, and z. Both dimensions were relevant for the solution of the discrimination

A repeated measures ANOVA, with factors of dimension and stage, confirmed no differences in dwell times, and yielded no significant main effects or interactions (smallest p = .098 for the main effect of stage). The areas of the screen that each ROI occupied differed in size and location; therefore, no particular

significance can be attributed to the fact that participants still spent some time in those regions even when they were no longer present (i.e., during revaluation and test trials). For example, during trials A+ only one dimension (e.g., the octopuses' eyes) was present in the image. However, it seems reasonable to assume that a participant might have also directed their gaze toward the tentacle area, bigger in size, even if just blank during these trials.

Experiment 3 replicated results from Experiment 2, and added to the demonstration of a positive relationship between performance in a configural acquired equivalence task and an optional-shift task in human participants. In addition to offering a direct replication of Experiment 2, Experiment 3 added to previous demonstrations of the relationship between predictiveness and overt attention, as measured by eye-gaze. Findings of a positive relationship between performance in configural acquired equivalence and attentional set in Experiments 2 and 3 are supportive of Honey et al.'s (2010) claims of a single underlying psychological mechanism. However, the possibility of other intervening variables not specific to acquired equivalence or attentional set influencing the observed correlation still remains. Experiment 4 aimed to address this issue by incorporating an additional control task, as means to confirm the specific correlation between configural acquired equivalence and optional-shift.

2.4 Experiment 4

Experiment 3, which was a direct replication of Experiment 2, demonstrated configural acquired equivalence and attentional set using an optional-shift in human participants and replicated the positive correlation between performance in these two tasks. Experiment 4 intended to replicate these results a third time and incorporated a control *N*-back task to account for any possible non-specific variables (e.g., general interest in the experiment).

In an *N*-back task, participants are required to decide whether the target stimulus in a sequence matches the one that appeared *n* trials ago. The difficulty of the task increases progressively as *n* updates (e.g., 1-back, 2-back, 3-back, etc.). In the *N*-back task reported here, participants were required to observe a sequence of black and white cars and to make a keyboard response whenever any particular car matched the one they saw *n* trials ago (e.g., ABCD**B**AA**B**B**AB** in a hypothetical 3-back sequence).

The *N*-back task has been extensively used as a standard measure of working memory in cognitive and neuroscience research (e.g., Harvey et al., 2005; Jaeggi et al., 2010; Owen et al., 2005; Rac-Lubashevsky & Kessler, 2016). Of most importance to the present experiment, evidence suggests that *N*-back does not correlate with performance in attentional set (Bergvall et al., 2001) or acquired equivalence tasks (Kéri et al., 2005). The results from the *N*-back task were intended to be partialled out from the expected correlation between configural acquired equivalence and optional-shift. A positive correlation between performance in acquired equivalence and optional-shift, even when holding performance in *N*-back constant (Simon, 1954), would have provided robust evidence for a positive correlation between both tasks. However, the task was not analysed and was omitted

following the failure to obtain the anticipated correlation between acquired equivalence and optional-shift (see Results and Discussion section).

2.4.1 Method

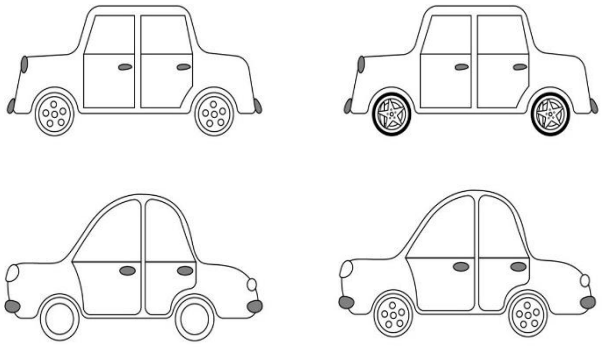
2.4.1.1 Participants, Apparatus & Stimuli

32 students from the University of Nottingham participated (11 men and 18 women, M_{age} = 20.84, *SD* = 2.54, range: 18-28). Students received module credits or an allowance for their participation. Participants were recruited using posters and the School of Psychology online booking system.

Experiment 4 used the same visual stimuli and counterbalancing than Experiment 3 for the configural acquired equivalence and optional-shift tasks. Stimuli for the *N*-back consisted of eight black and white cartoon cars 10 cm (width) x 8 cm (height) presented on a grey background. Cars differed in their roof and wheels, as exemplified in **Figure 18**.

Figure 18

Example Stimuli presented During the N-back Task of Experiment 4



2.4.1.2 Procedure

Participants completed the acquired equivalence, optional-shift and *N*-back tasks. 16 participants performed the acquired equivalence before the optional-shift; the remainder performed the tasks in the alternative sequence to match the effects of any potential temporal variable on each task. All participants completed the *N*-back task last. The experimenter left the room after ensuring participants had understood the tasks and returned only to set up the second and third tasks, respectively. During the *N*-back task, participants were presented with a set of written instructions asking them to *"Pay attention to the sequence of cars that you are going to see. You will have to decide whether the car onscreen is identical to the one you saw X positions before in the sequence. Initially, you will be asked to press the SPACE BAR whenever the car onscreen matches the one you saw 1 positions before. However, the task will get harder! Each car will be presented for 2 seconds, so you will have to make a decision quickly".* Participants saw example 1-back, 2-back ad 3-back trials during the instruction screens.

The *N*-back task was split in three levels that increased the level of difficulty progressively: 1-back, 2-back and 3-back. Each level comprised the presentation of each car twice per block over four blocks (64 trials). Trials were pseudorandomised within each block, so that each block contained four target stimuli. That is, four trials during which participants were expected to press the space bar. Stimuli were presented in the centre of the screen for 2 s. After the participant's response, the feedback *Correct!* Or *Ooops! That was wrong* appeared in the center of the screen for 1 s. All unspecified details are identical to those in Experiment 3.

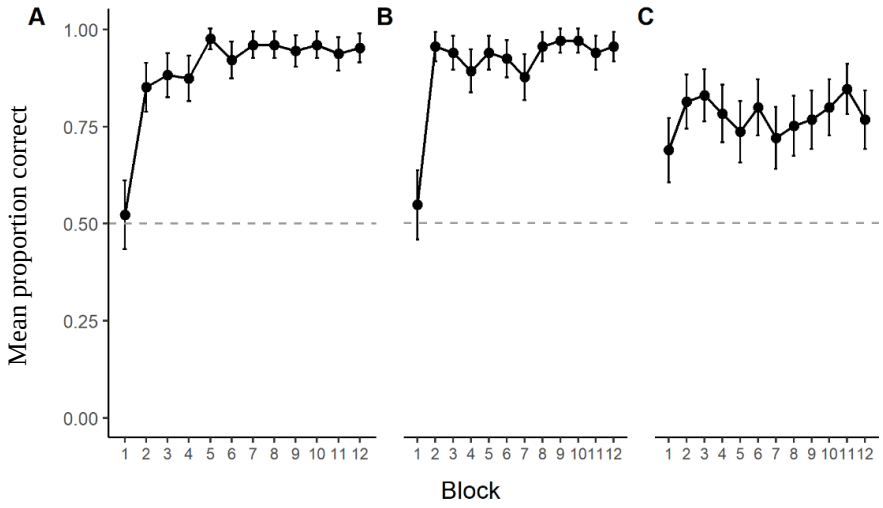
2.4.2 Results and Discussion

2.4.2.1 Optional-shift

Data from Stage 1, summarised in **Figure 19(a)**, show that participants learned the initial discrimination very rapidly, with performance closely matching that of the previous experiments reported in this chapter. A one-sample *t*-test confirmed that participants were confidently performing above chance during the second half of training, $t(31) = 20.41$, $p < .001$, $d = 3.61$, 90% CI [2.79, 4.40]. An ANOVA once again confirmed participants' acquisition of the discrimination with a main effect of block, $F(4.62, 138.06) = 26.41$, $p < .001$, $\eta_p^2 = .47$, 90% CI [0.35, 0.53].

Figure 19

Collapsed Mean Performance for Stage 1 and Revaluation and Test trials of the Optional-Shift Task



Note. (A) Each block comprised the presentation of compound cues Aw, Ax, Bw and Bx followed by feedback. (B) Each block comprised the presentation of new trials Cy and Dz followed by feedback. (C) Each block comprised test trials Cz and Dy, which were not followed by feedback. Vertical bars represent standard errors of the mean. The horizontal dashed line represents chance performance. Revaluation and test trials were intermixed during the task.

Figure 19(b) shows the discrimination to the new compounds Cy and Dz during Stage 2 collapsed over version of the task and outcome. Accuracy dropped to chance levels during the first block but it recovered immediately. An ANOVA confirmed a main effect of block, $F(5.72, 177.32) = 13.26$, $p < .001$, $\eta_p^2 = .30$, 90% CI [0.19, 0.36].

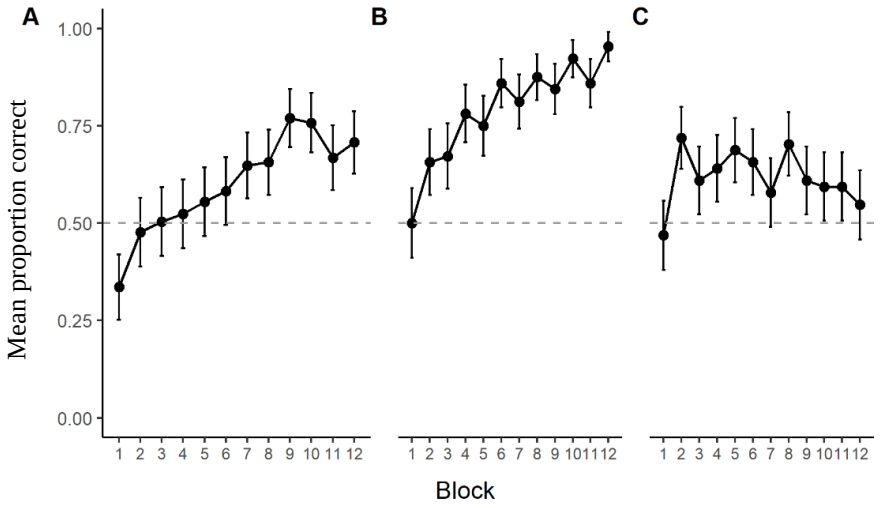
Accuracy to test trials Cz and Dy, shown in **Figure 19(c)**, confirmed participants' bias toward the dimension that had been predictive during initial training. The results are closely matched to those of Experiment 2 and Experiment 3, and unequivocally confirm participants' ability to respond to these stimuli despite the lack of any explicit feedback. A one-sample *t*-test confirmed that participants' mean accuracy was reliably above chance, $t(31) = 5.06$, $p < .001$, $d = 0.89$, 90% CI [0.54, 1.23].

2.4.2.2 Acquired equivalence

Stage 1 data are summarised in **Figure 20(a)**. As in the previous three experiments, participants learned the contingencies between compound stimuli and outcomes progressively, showing a reliable performance during the second half of Stage 1 training, $t(31) = 6.19$, $p < .001$, $d = 1.09$, 90% CI [0.72, 1.46]. This observation was supported by an ANOVA, which showed a main effect of block, $F(6.60, 204.60) = 17.68$, $p < .001$, $\eta_p^2 = .37$, 90% CI [0.27, 0.43].

Figure 20

Collapsed Mean Performance for Stage 1 and Revaluation and Test trials of the Acquired Equivalence Task



Note. (A) Each block comprised the presentation of compound cues Aw, Ax, Bw, Bx, Cw, Cx, Dw and Dx followed by feedback. (B) Each block comprised revaluation trials A and B followed by feedback. (C) Each block comprised test trials C and D, which were not followed by feedback. Vertical bars represent standard errors of the mean. The horizontal dashed line represents chance performance. Revaluation and test trials were intermixed during the task.

Data from Stage 2 revaluation trials were collapsed over version of the task and outcome and are shown in **Figure 20(b).** Data suggest that these trials proceeded somehow more progressively than in the previous experiments, where participants tended to quickly master the discrimination of A and B revaluation trials after an initial decline. Nevertheless, an ANOVA confirmed an effect of block, $F(5.83, 180.73) = 7.64$, $p < .001$, $\eta_p^2 = .20$, 90% CI [0.11, 0.26].

Data from test trials are shown in **Figure 20(c)**. Experiment 4 once again replicated the acquired equivalence effect obtained in the previous experiments**.** Accuracy in these trials started at chance levels but improved quickly and was

consistently maintained until the end of the task, demonstrating participants' ability to transfer responding to test trials. A one-sample *t*-test confirmed that participants' discrimination was reliably above chance, $t(31) = 2.98$, $p = .003$, $d = 0.53$, 90% CI [0.21, 0.83].

As with the experiments previously reported in this chapter, participants' individual performance during test trials in both tasks was correlated. Based on the positive correlations reported in Experiment 2 ($r = .43$) and Experiment 3 ($r = .54$), we expected the correlation between these measures to be reliable also in Experiment 4. However, as illustrated by **Figure 11(d)**, we failed to obtained the anticipated positive correlation, Pearson's $r(30) = -.10$, $p = .709$. The additional Bayesian analysis returned BF = 5.09 in favour of the null model, providing substantial evidence for the lack of a positive relationship between performance in both tasks in Experiment 4, and Fisher's z transformation confirmed that the correlation differed from that of Experiment 2, $z = 2.13$ $p = .033$ and Experiment 3, $z = 2.68$ $p = .007$.

These results are challenging in several ways. The Bayesian analysis of the positive correlation reported in Experiments 2 and 3 provided decisive support in favour of a positive correlation. Whilst all participants were University of Nottingham students of similar ages and education levels who participated under very comparable experimental conditions, these three experiments still tested three different groups of participants and as such, differences between the groups could be expected. Results cannot be attributed to the inclusion of an additional control task, which was delivered only after participants had completed the acquired equivalence and optional-shift tasks. In an attempt to increase the power of the correlation and determine the overall support for a positive relationship between configural acquired equivalence and optional-shift, we pooled participants from Experiments 2, 3 and 4

and (96 participants) and run a Bayesian correlation analysis. The analysis yielded an overall Pearson's r = .31, and a BF = 25.33 in favour of the alternative model, which provides overall strong support for a positive correlation between both tasks.

2.5 General Discussion

In this chapter, four experiments investigated Honey et al.'s (2010) claims that acquired equivalence and attentional set may rely on a common mechanism. Experiment 1 demonstrated acquired equivalence, evidenced by the transfer of responses from revaluation to test trials with no explicit feedback, and showed an anticipated IDS superiority in a CANTAB IDS/EDS. However, it failed to obtain the expected positive correlation between performance in these tasks. Experiment 2 replicated the acquired equivalence effect and demonstrated an attentional set effect with an optional-shift task, which demonstrated participants' preference for the stimulus dimension that had been established as a relevant predictor during initial training without any explicit feedback. Experiment 2 demonstrated a positive relationship between performance in configural acquired equivalence and attentional set when assessed with experimentally matched tasks. Experiment 3 replicated the positive relationship between performance in acquired equivalence and optional-shift found in the previous experiment. It also assessed the relationship between predictiveness and learning with the use of eye-tracking. Dwell times showed how participants' bias toward the cue dimensions that had been initially relevant transferred to a subsequent stage where both dimensions were objectively equally relevant, echoing previous findings. Finally, Experiment 4 sought to provide unequivocal support for a positive correlation between both phenomena with the addition of an *N*-back control task. Although the correlation between acquired equivalence and optional-shift failed to replicate in the last experiment of the series reported here, the pooled Bayesian correlation provided strong overall support for a positive correlation between these two tasks.

Performance in attentional set tasks and configural acquired equivalence has been shown to be selectively impaired in healthy older adults (e.g., Owen et al., 1991; Robinson & Owens, 2013), and affected by selective brain lesions, with rats with lesions to the entorhinal cortex showing good conditional learning but impaired acquired equivalence and attentional set (e.g., Coutureau et al., 2002; Oswald et al., 2001). The novelty of our findings relies on the assessment of these two phenomena in within-subjects, comparable experiments.

Following Honey et al.'s (2010) claims, and based on the aforementioned evidence, we reasoned that performance in both tasks should be expected to correlate positively. Our new findings overall supported this suggestion when both processes were tested using well-matched acquired equivalence and attentional set tasks, but not when testing configural acquired equivalence against IDS/EDS. A possible explanation for these results could involve IDS/EDS and optional-shift, regarded as attentional set tasks, not sharing a common psychological mechanism. For example, performance in IDS/EDS could be governed by psychological processes different from the ones governing performance in both configural acquired equivalence and optional-shift, which could explain the positive correlation found only between these two tasks. Of course, there is an alternative interpretation to these results. The notable differences between IDS/EDS and optional-shift may have resulted in the observed dissociation between configural acquired equivalence and attentional set.

The acquired equivalence and optional-shift tasks used during Experiments 2, 3, and 4 were closely matched to allow direct comparison: both tasks used the same sets of visual and audio-visual stimuli during an identical number of revaluation and test trials. Critically to our test of the network, the optional-shift task yielded a single, direct measure of attentional set, which, taken in conjunction with the closely

matched designs, could be directly correlated against the measure of acquired equivalence. However, in Experiment 1 the differences between the acquired equivalence task and IDS/EDS are clear. The tasks used different stimuli, presented in two different devices, and with different presentation times and number of trials. Whilst the acquired equivalence task measured accuracy to a predetermined fixed number of trials with no feedback, IDS/EDS measured errors to a criterion of six consecutive correct responses. Whilst we derived a single measure of attentional set from the measures of IDS and EDS, our way of assessing attentional set might presumably be affected by participants' varied exposure to stimuli during the IDS and EDS stages. If we consider again how the network is assumed to work, it becomes clear that these differences could have affected the way in which the network forms and trains its connection. For example, the limited exposure to stimuli during each stage of the IDS/EDS could have led to sub-optimal input-to-hidden layer connections, critical to the solution of the discrimination (Honey et al., 2010).

It is worth noting that the increased dwell times to predictive compared to nonpredictive stimuli in Experiment 3 do not preclude a simple connectionist network from accommodating our findings. Highly predictive cues during the initial training stage of the optional-shift task would have been better-able to activate specific hidden units, facilitating the rate of learning of new stimulus-outcome contingencies during revaluation. Even if the bias for predictive stimuli during the first revaluation trials correlated with later accuracy performance, we cannot be sure that the difference in overt attention, as measured by eye-gaze, caused the difference in learning about those stimuli (Le Pelley, 2010).

Our findings do not unambiguously confirm Honey et al.'s claims that there is a relationship between configural learning and attentional set brought about by a

common mechanism. Both phenomena could depend on different psychological mechanisms, both of which could be affected by age and selective brain lesions. The finding of an apparent dissociation between performance in acquired equivalence and the two different measures of attentional set is novel and experimentally interesting in its own right, and could open new avenues of research in the future. This chapter provides a direct comparison between these two procedures in a series of single, within-subjects, comparable experiments for the first time and adds to the generality of demonstrations of the apparent relationship between these two processes demonstrated by ageing and lesions studies.

2.5.1 Conclusion

The experiments reported in this chapter offer partial support to Honey et al.'s (2010) claims. Their importance rely on the fact that, to our knowledge, they are the first ones seeking to directly assess the relationship between configural acquired equivalence and attentional set in within-subjects experiment. Experiment 1 added to the demonstrations of configural acquired equivalence and IDS superiority, but failed to find a positive relationship between the two. Experiments 2 and 3 demonstrated a positive relationship between configural acquired equivalence and optional-shift. However, Experiment 4 failed to replicate this positive correlation. Despite this, a pooled Bayesian correlation offered overall strong support for a positive relationship between the two phenomena. In any case, this correlational approach does not allow us to make any unequivocal conclusions. For example, whilst the BF provided strong evidence for a positive correlation between configural acquired equivalence and optional-shift, we cannot rule out the possibility of non-specific factors underpinning the observed correlation. Chapter 3 will leave aside this

question and focus instead on the experimental assessment of outcome manipulations in configural and non-configural acquired equivalence.

Chapter 3:

Experimental assessment of outcome manipulations in different forms of acquired equivalence

Outcome manipulations in configural acquired equivalence

Previous research has demonstrated that performance in discrimination learning tasks is improved by the use of different outcomes as opposed to a common outcome across reinforced responses. This differential outcome effect (DOE) was first demonstrated by Trapold (1970). Trapold (1970) trained rats on a discrimination in which a left or right lever press was reinforced in the presence of different auditory stimuli (i.e., S1 – R1 and S2 – R2). In an experimental group, rats were trained with differential outcomes (e.g., a food pellet for correct responses to a tone and sucrose for correct responses to a clicker). These rats learned the discrimination faster than did rats in a control group, which received the same reinforcer for all correct responses (i.e., food pellets only or sucrose only). Even when a control and experimental group received differential outcomes in a discrimination task, performance has been found to be enhanced when each stimulus signals a unique outcome (e.g., S1 – R1 – Food and S2 – R2 – Sucrose) compared to when stimuli are rewarded with both differential outcomes equally (e.g., S1 – R1 – Food/Sucrose or S2 – R2 – Food/Sucrose) (e.g., Delamater et al., 2010). Since the first demonstration, the DOE has proven to be a robust effect, consistently replicated under a range of conditions, in both human and non-human animals, and in Pavlovian and instrumental learning tasks (Carlson & Wielkiewicz, 1972; Delamater et al., 2010; Edwards et al., 1982; Estévez et al., 2001; McCormack et al., 2019; Urcuioli, 2005).

Trapold and colleagues argued that in these discriminations, the differential outcomes generated the expectancy of a particular outcome which, in turn, acted as an additional stimulus that evoked a specific lever press response. That is, the expectancy of the outcome was, in itself, part of what was learned during the discrimination. From this interpretation, it follows that the experimental group could

learn to press a lever when expecting one outcome (e.g., S1 – R1 – Food) and a different lever when expecting a different outcome (e.g., S2 – R2 – sucrose). On the other hand, the control group could not use the different expectancies as a means to aid discrimination learning (e.g., S1 – R1 – Food and S2 – R2 – Food) (Trapold, 1970; Trapold & Overmier, 1972; Urcuioli, 2005).

Outcome manipulations are of interest to this thesis because signalling a common outcome is one of the ways in which stimuli come to be treated as equivalent. The DOE has been attributed to an acquired equivalence and distinctiveness of cues (Delamater, 1998, 2012; Delamater et al., 2010; Edwards et al., 1982). When stimuli are followed by unique differential outcomes, their internal representations become more distinct, facilitating the discrimination between the two. Conversely, when stimuli are followed by the same outcome, their internal representations tend to blend and become equivalent, making it harder for the organism to solve the discrimination. It is easy to see, and I have already discussed in previous chapters, how these ideas would extend to a connectionist network such as Honey's (Honey, 2000; Honey et al., 2010). The unique differential outcomes would have an essential role, allowing for the solution of the discrimination by recruiting separate hidden units based on their differential training history. However, without differential outcomes driving the formation of separate internal representations at the hidden layer level, it is difficult to see how the network would solve the initial discrimination.

The configural acquired equivalence tasks mentioned so far in this thesis have all used differential outcomes during training. However, prior research has differed in its manipulation of the outcomes during the revaluation stage of the task. In a typical revaluation experiment, the animal is first trained on a biconditional

configural discrimination in which stimuli A and C signal the delivery of food when paired with stimulus *w* and the absence of food when paired with *x,* and the reversed contingencies are true for stimuli B and D: Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+. In some experiments, stimulus A and B are revalued so that a mild footshock is delivered to the animal in the presence of A but not in the presence of B. That is, animals receive a *different* set of unique differential outcomes during revaluation (e.g., Honey & Watt, 1998; Ward-Robinson & Honey, 2000). In a different version of the revaluation procedure, stimuli are revalued with the *same* set of unique differential outcomes that was used during training (i.e., food) (e.g., Coutureau et al., 2002; Iordanova et al., 2007).

This is experimentally interesting because, to our knowledge, no prior research has directly assessed the effects of presenting the same versus different unique differential outcomes across stages in acquired equivalence. A direct comparison of existing experimental data would be ill-equipped to explore this question, because the use of the same or different outcomes in different tasks is confounded with the intrinsic properties of those outcomes. For example, a stronger acquired equivalence effect in an experiment that used different outcomes across stages (food – footshock) compared to an experiment with the same outcomes (food – food) could simply reflect that a footshock is a more potent reinforcer than food, rather than reflecting the effect of presenting the same or different outcomes per se. The following three experiments address this question by systematically varying a set of non-motivationally significant outcomes within a single experiment, to create counterbalanced experimental subgroups with the same vs. different outcomes across stages that can be compared directly. Experiment 5 found an enhanced acquired equivalence performance when participants experienced different outcomes across

training and revaluation, compared to participants who experienced the same outcomes across stages. Experiment 6 assessed the possibility that the enhanced performance could be attributed to factors non-specific to our outcome manipulation (e.g., arousal) by presenting all outcomes from the onset of the task. Experiment 7 rectified suspected issues with Experiment 6.

3.1 Experiment 5

Experiment 5 was intended to assess the effects of using the same or different outcomes across training and revaluation on the strength of the acquired equivalence effect. To that end, participants were assigned to either group *Same* or group *Different* to complete a configural acquired equivalence task, as summarised in **Table 6**. The acquired equivalence task used here was identical to that described in the previous chapter, but outcomes were systematically varied to create four factorial experimental subgroups. Participants were once again asked to *"Imagine yourself in the role of a marine tour guide. It is your job to keep tourists safe from all dangerous animals. Your boat is about to enter an area densely populated by octopuses that are known to be dangerous to humans".* The instructions indicated that it was participants' task to look at the octopuses and learn which threat each octopus signalled. During Stage 1, outcomes were counterbalanced so half of the participants received "bite/sting" as their outcomes and the other half of the participants received "poison/suffocate". During Stage 2, participants assigned to group Same experienced no change in the outcomes between stages (i.e., bite/sting – bite/sting or poison/suffocate – poison/suffocate). Participants in group Different were presented with the complementary outcomes during the revaluation and test trials of Stage 2 (i.e., bite/sting – poison/suffocate or poison/suffocate – bite/sting). The stimulus-outcomes contingencies were counterbalanced within each experimental subgroup. Stage 2 results demonstrated an enhanced acquired equivalence effect for the group that received *different* outcomes across training and revaluation trials.

Table 6

Experimental Design for Experiment 5

Group Same			Group Different		
Stage 1		Stage 2 Revaluation and Test	Stage 1		Stage 2 Revaluation and Test
a.			c.		
Aw $	Ax *	A $	Aw $	Ax *	A +
Bw *	Bx $	B *	Bw *	Bx $	B –
Cw $	Cx *	**C ?**	Cw $	Cx *	**C ?**
Dw *	Dx $	**D ?**	Dw *	Dx $	**D ?**
b.			d.		
Aw +	Ax –	A +	Aw +	Ax –	A $
Bw –	Bx +	B –	Bw –	Bx +	B *
Cw +	Cx –	**C ?**	Cw +	Cx –	**C ?**
Dw –	Dx +	**D ?**	Dw –	Dx +	**D ?**

Note. Letters A-D represent different eyes and w/x represent different types of tentacles. +, –, * and $ represent outcomes bite, sting, poison and suffocate, respectively. ? indicates the absence of feedback. Revaluation and test trials were intermixed during Stage 2. Group Same received the same outcomes across Stage 1 and Stage 2 trials. Group Different received different outcomes during Stage 1 and Stage 2 trials.

3.1.1 Method

3.1.1.1 Participants, Apparatus & Stimuli

64 students from the University of Nottingham participated (50 women and 14 men, M_{age} = 20.33, *SD* = 3.18, range: 18-32). Participants were informed about the task and debriefed upon completion. All agreed to participate. Participants received module credits or a small allowance for their participation. Participants were recruited using posters and the School of Psychology online booking system. The configural octopus stimuli used in this task were identical to those described in previous experiments, and were presented to participants with the same computer as described in Section 2.1.1.2. All unspecified details are identical to those previously described.

3.1.1.2 Procedure

Prior to the start of the experiment participants were assigned to group Same or Different. Participants were randomly assigned without replacement to guarantee the desired number of participants per group (n = 32 each). Participants were assigned to counterbalancing subgroups in a predetermined fashion (i.e., the first participant in a group was assigned to the first counterbalancing subgroup, the second to the second counterbalancing subgroup, etc.). Participants from both groups underwent, on average, the same treatment during Stage 1. When presented with stimuli A and B during Stage 2, participants from group Different received a new set outcomes, orthogonal to that of Stage 1. For example, for one counterbalancing subgroup in group Different, octopuses went from signalling *Bite* or *Sting* to signalling *Poison* or *Suffocate* during Stage 2. Participants were not explicitly warned about this change, but the keyboard reminder on screen and the feedback changed accordingly, forcing participants to select from the new outcomes. For example, the keypress reminder changed from *left – Bite* and *right – Sting* to *left – Poison* and *right – Suffocate*. All unspecified procedural details are identical to those of the acquired equivalence tasks previously described.

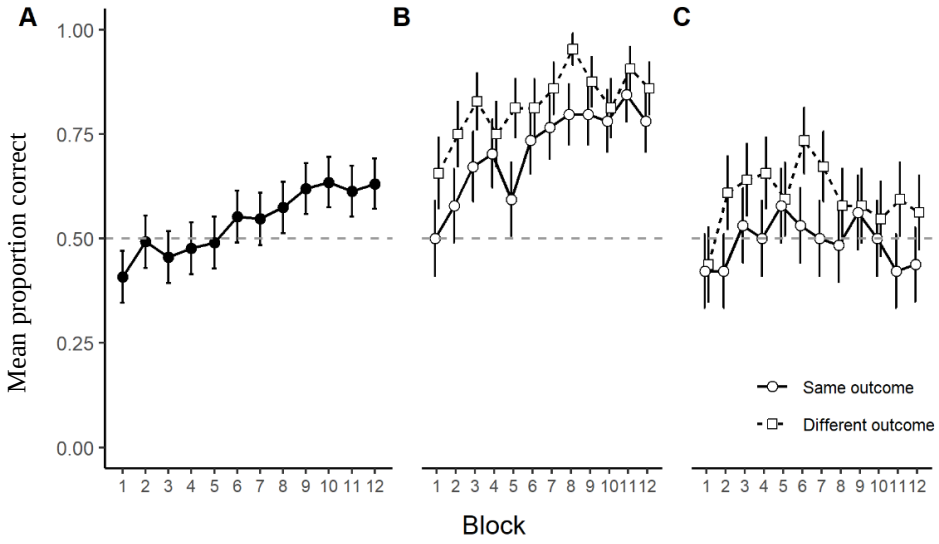
3.1.2 Results and Discussion

During Stage 1, participants progressively learned the contingencies between the octopuses and outcomes. Because groups Same and Different had undifferentiated training, data were collapsed for initial analyses, as summarised in **Figure 21(a)**. Participants performance was rather low during the first few blocks of the task, but a one-sample *t*-test confirmed that it was reliably above chance during

the second half of training (blocks 7 to 12), $t(63) = 4.46$, $p < .001$, $d = 0.56$, 90% CI [0.33, 0.78]. A mixed ANOVA, with between-subjects factors of group and outcome and a within-subjects factor of block was run on these data. The analysis yielded a significant effect of block, $F(9.35, 1122) = 13.06$, $p < .001$, $\eta_\rho^2 = .11$, 90% CI [0.07, 0.12] but no other main effects or interactions were significant (smallest $p = .392$ for the main effect of group). Of most importance for our later analysis, the analysis showed no group or outcome effects at this stage.

Figure 21

Collapsed Mean Performance for Stage 1 and Revaluation and Test trials of the Acquired Equivalence Task



Note. (A) Each block comprised the presentation of compound cues Aw, Ax, Bw, Bx, Cw, Cx, Dw and Dx followed by feedback. Both the Same and Different groups received the same training. (B) Each block comprised revaluation trials A and B, followed by either the same or different outcomes, followed by feedback. (C) Each block comprised test trials C and D, which were not followed by feedback. Vertical bars represent standard errors of the mean. The horizontal dashed line represents chance performance. Revaluation and test trials were intermixed during the task.

Stage 2 revaluation trials A and B were collapsed over group for analysis. Examination of these data showed that participants were quick to learn the new discrimination, as illustrated in **Figure 21(b)**. The apparent advantage in performance for group Different was confirmed by a mixed ANOVA, with the within-subjects factor of block and the between-subjects factors of outcome and group. The analysis yielded a significant effect of group, $F(1, 62) = 5.28$, $p = .025$, $\eta_\rho^2 = .08$, 90% CI [0.01, 0.20] and block, $F(8.03, 497.86) = 6.62$, $p < .001$, $\eta_\rho^2 = .11$, 90% CI [0.05, 0.13]. The analysis revealed no main effect or interactions involving outcome (smallest $p = .244$ for the interaction between group, outcome and block), suggesting an enhanced performance in group Different irrespective of the outcome.

Examination of critical test trials C and D, collapsed over group and summarised in **Figure 21(c)**, reflects an enhanced performance in group Different compared to group Same. A mixed ANOVA, with between-subjects factors of group and outcome and a within-subjects factor of block confirmed this observation. The analysis revealed a main effect of group, confirming an advantage in performance for group Different over group Same, $F(1, 120) = 7.88$, $p = .006$, $\eta_\rho^2 = .10$, 90% CI [0.01, 0.14]. Unexpectedly, the analysis also yielded a main effect of outcome, $F(3, 120) = 3.37$, $p = .021$, $\eta_\rho^2 = .11$, 90% CI [0.01, 0.15]. Holm corrected pairwise comparisons showed that the effect was driven by a lower performance in the 'poison' outcome than any of the other outcomes ($p < .001$). This difference in performance, however, cannot be attributed to an overall preference for any of the outcomes in previous stages, as analyses showed outcomes were all well matched throughout. The analyses showed no other main effects or interactions (smallest $p = .051$ for the main effect of block). To assess whether participants' responses at test reflected the past equivalent training history of stimuli A/C and B/D, we tested

participants' overall mean performance against chance level (50%). Two independent one-sample *t*-test showed that participants' categorisation of stimuli C and D was reliably above chance despite the absence of any explicit feedback in group Different, $t(31) = 2.77$, $p < .001$, $d = 0.49$, 90% CI [0.18, 0.79]; but not in group Same, $t(31) = -.38$, $p = .647$, $d = -0.07$, 90% CI [-0.22, 0.36]. That is, only group Different showed an acquired equivalence effect.

Results from Experiment 5 suggest an advantage for participants in group Different. However, the fact that group Same did not show an acquired equivalence effect challenges a comparison between the two groups. Especially considering that group Same in Experiment 5 performed a task identical to the ones presented in experiments in Chapter 1, where participants did show the anticipated acquired equivalence effect. Additionally, whilst the systematic design of the experiment allows for direct comparison between the groups, differences in performance could be attributed to factors other than the use of different outcomes during revaluation trials. For example, participants in the Different group may have experienced increased arousal from the unexpected change in the outcomes. This could have led to an increased general interest in the task, which could have, in turn, resulted in the observed increased performance. The next experiment aimed to account for any effects non-specific to the change of outcomes by presenting a variation of our acquired equivalence task in which the inclusion of additional non-configural stimuli during training served as a means to present all four possible outcomes from the beginning of the task.

3.2 Experiment 6

The aim of Experiment 6 was to further test the effects of manipulating the outcomes across stages whilst addressing some of the limitations of Experiment 5. Specifically, to reduce group differences in level of arousal when experiencing novel outcomes during the revaluation stage of our acquired equivalence task. To that end, Experiment 6 introduced a variation in our usual configural discrimination by incorporating two additional, non-configural cues (S1 and S2), during training.

Non-configural stimuli consisted of two easily discriminable squid cartoons, which were presented in conjunction with the usual configural stimuli during training in both groups, as illustrated in **Table 7**. These additional cues ensured both groups were exposed to all possible outcomes from the onset of the task, reducing any potential non-specific effects in performance as a result of the differential group treatment. Experiment 6 failed to show any group differences in performance. Indeed, the experiment failed to show the acquired equivalence effect. We discuss and attribute these results to a poor choice of response keys that was rectified in Experiment 7.

Table 7

Experimental Design for Experiment 6 and 7

Group Same		Group Different	
Stage 1	Stage 2 Revaluation and Test	Stage 1	Stage 2 Revaluation and Test
a.		c.	
Aw $ Ax *	A $	Aw $ Ax *	A +
Bw * Bx $	B *	Bw * Bx $	B –
Cw $ Cx *	**C ?**	Cw $ Cx *	**C ?**
Dw * Dx $	**D ?**	Dw * Dx $	**D ?**
S1 +	S1 +	S1 +	S1 $
S2 –	S2 –	S2 –	S2 *
b.		d.	
Aw + Ax –	A +	Aw + Ax –	A $
Bw – Bx +	B –	Bw – Bx +	B *
Cw + Cx –	**C ?**	Cw + Cx –	**C ?**
Dw – Dx +	**D ?**	Dw – Dx +	**D ?**
S1 $	S1 $	S1 $	S1 +
S2 *	S2 *	S2 *	S2 –

Note. Letters A-D represent different eyes and w/x represent different types of tentacles. Non-configural trials S1 and S2 consisted of two very distinct squid drawings. +, –, * and $ represent outcomes bite, sting, poison and suffocate, respectively. ? indicates the absence of feedback. Revaluation and test trials were intermixed during Stage 2. Group Same received the same outcomes across Stage 1 and Stage 2 trials. In group Different, the contingencies between stimuli-outcomes were different across both stages.

3.2.1 Method

3.2.1.1 Participants, Apparatus & Stimuli

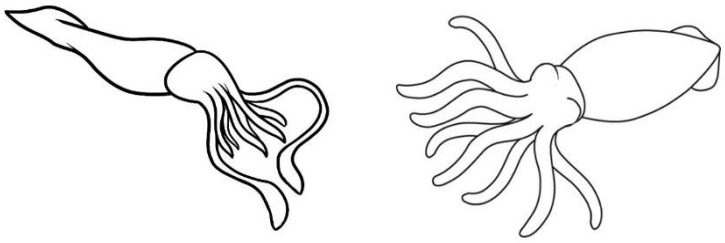
32 students from the University of Nottingham participated (30 women and 2 men, M_{age} = 20.38, *SD* = 2.61, range: 18-28). Participants were informed about the task and debriefed upon completion. All agreed to participate.

Participants received module credits or a small allowance for their participation. Participants were recruited using posters and the School of Psychology online booking system. This task used the configural octopus stimuli described in previous experiments. The two additional non-configural stimuli consisted of distinct

cartoon squids 10 cm (width) x 8 cm (height), as shown in **Figure 22**. All unspecified details were identical to those of Experiment 5.

Figure 22

Example Non-Configural Stimuli presented during Experiment 6 and 7



3.2.1.2 Procedure

Participants were assigned to group Same or Different in a random fashion without replacement to ensure the desired number of participants per group (n = 16 each). During Stage 1, participants received, on average, the same treatment. During training, non-configural stimuli S1 and S2 were each presented four times per block in conjunction with the usual configural discrimination (192 trials). This ensured that participants in both groups were presented with each of the four possible outcomes an equal number of trials. In keeping Experiment 6 as closely matched as possible to Experiment 5, response keys remained unchanged. Instead, each key signalled two complimentary outcomes. That is, *left – Bite OR Poison* and *right – Sting OR Suffocate*.

During revaluation, stimuli continued to signal the same outcomes for group Same. For participants in group Different, stimuli signalled the complimentary outcomes. For example, for one counterbalancing subgroup in group Different,

octopuses went from signalling *Bite* or *Sting* to signalling *Poison* or *Suffocate,* and squids went from signalling *Poison* and *Suffocate* to signalling *Bite* and *Sting*. That is, group Different experienced new stimulus-outcome contingencies, rather than new outcomes per se. This, we reasoned, reduced the differential levels of arousal between the groups.

3.2.2 Results and Discussion

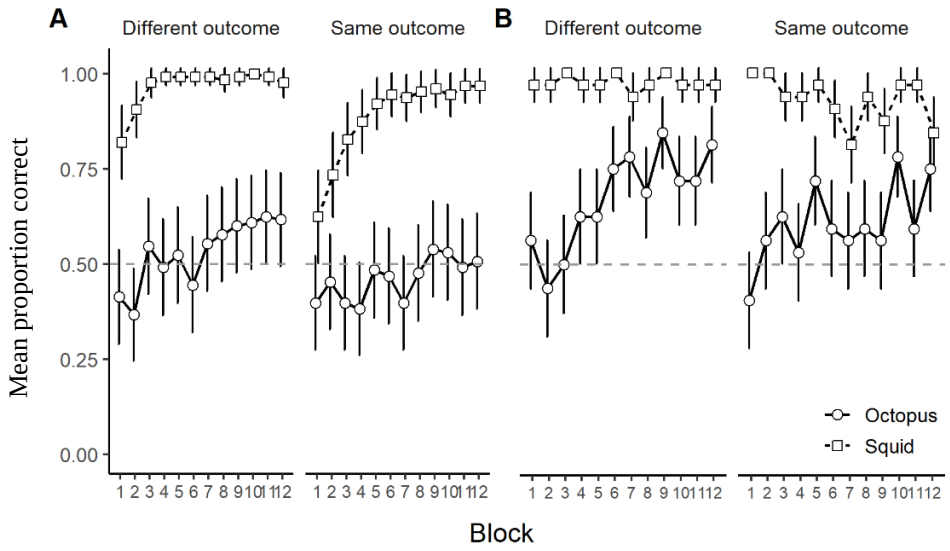
3.2.2.1 Stage 1

The aim of incorporating non-configural cues to our task was to ensure that participants were exposed to all possible outcomes from the onset of the task. However, we expected the discrimination of these stimuli to be trivially easy. Analysing configural and non-configural stimuli together would have resulted in an inflated discrimination performance. Hence, configural and non-configural stimuli were analysed separately. Training data are summarised in in **Figure 23(a)**. In keeping with the simplicity of the non-configural discrimination, the analysis revealed that participants quickly learned how to discriminate between the squids, reaching and maintaining performance levels close to asymptote early in training. However, a quick inspection of the data suggests that octopuses were considerably harder to discriminate. Given the identical treatment of the groups at this stage of the task, no significant differences in performance were expected. However, when looking at the non-configural squid stimuli, a mixed ANOVA with a within-subjects factor of block and between-subjects factors of group (Same vs. Different) and outcome revealed a main effect of group, $F(1, 56) = 16.05$, $p < .001$, $\eta_\rho^2 = .22$, 90% CI [0.08, 0.36] with group Different performing reliably better than group Same. The analysis also showed a main effect of block, $F(4.18, 234.08) = 21.84$, $p < .001$, $\eta_\rho^2 =$

.28, 90% CI [0.19, 0.34] but no other main effects or interactions were significant (smallest p = .085 for the interaction between group and block). Given the undifferentiated treatment of groups Same and Different during these trials we were not anticipating any group differences. However, with an average performance of .97 for group Different and .89 for group Same, it is evident that, overall, both groups mastered the discrimination of these non-configural trials. The difference could be driven by a more progressive acquisition of the discrimination in group Same compared to group Different, illustrated by the lower means per block and the wider error bars. Having assigned the 16 participants randomly to each group and with no differential treatment at this stage, this difference in performance must be coincidental. Furthermore, it is important to remember that we are not so much interested in performance in these non-configural trials per se, which we argued would be trivially easy for participants. Instead, the aim of incorporating these trials was only to reduce potential arousal effects in the subsequent revaluation stage by presenting all possible outcomes from the onset of the task. Given the fact that both groups were exposed to all outcomes and that they both learned this discrimination, we turned our focus to the actual stimuli of interest to investigate the consequences of this initial manipulation in the acquired equivalence effect.

Figure 23

Collapsed Mean Performance for Stage 1 and Revaluation trials of the Acquired Equivalence Task with Configural and Non-Configural Trials



Note. (A) Mean performance for configural (octopus) and non-configural (squids) stimuli during Stage 1 for the Same and Different groups. Each block comprised the presentation of stimuli (AW, AX, BW, BX, CW, CX, DW, DX, S1, S2). (B) Mean correct performance for stimuli during revaluation trials for the Same and Different outcomes groups. Each block comprised the presentation of 4 stimuli (A, B, S1, S2), which were followed by feedback. Error bars represent SEM. The horizontal dashed line indicates chance performance.

A mixed ANOVA with a within-subjects factor of block and between-subjects factors of group and outcome was run on the configural octopus stimuli. Once again, because groups had not received any differential treatment at this point, no group differences were expected. Contrary to our expectations, and in line with the group differences observed in the non-configural discrimination, the analysis yielded a main effect group, $F(1, 56) = 4.20$, $p = .045$, $\eta_p^2 = .07$, 90% CI [0.00, 0.19], suggesting group Different ($M = .53$) performed reliably better than group Same ($M = .46$) at this stage. The analysis also revealed a main effect of block, $F(11, 616) = 4.53$, $p < .001$, $\eta_p^2 = .07$, 90% CI [0.03, 0.09], but no other main effects or

interactions (smallest p = .057 for the interaction between group and outcome). The group differences observed in Stage 1 are problematic. Group differences should not be present at this stage, as they would challenge the interpretation of any possible group differences in subsequent stages of the task. However, it is worth noting that the p-value indicated a marginally reliable result and, importantly, the CI for the effect size included zero in their lower bound, which indicates uncertainty in the effect and a lack of significance (Lee, 2016).

Given the group difference, and to ensure participants had acquired the initial discrimination, we assessed participants' performance over the second half or training against chance for each group individually. Two one-sample t-tests revealed that that participants' performance was only reliably above chance for the Different group, $t(15) = 2.02$, $p = .030$, $d = 0.50$, 90% CI [0.06, 0.94] but not for the Same group, $t(15) = 0.31$, $p = .620$, $d = 0.07$, 90% CI [-0.33, 0.49]. These results are challenging, because acquired equivalence relies upon participants' acquisition of the initial discrimination, and significantly limit any further interpretation of Experiment 6. Nevertheless, data from the revaluation and test trials are still presented. Experimental factors that might have resulted in participants' failure to learn the initial discrimination are discussed and addressed in a follow up experiment.

3.2.2.2 Stage 2 – Revaluation trials

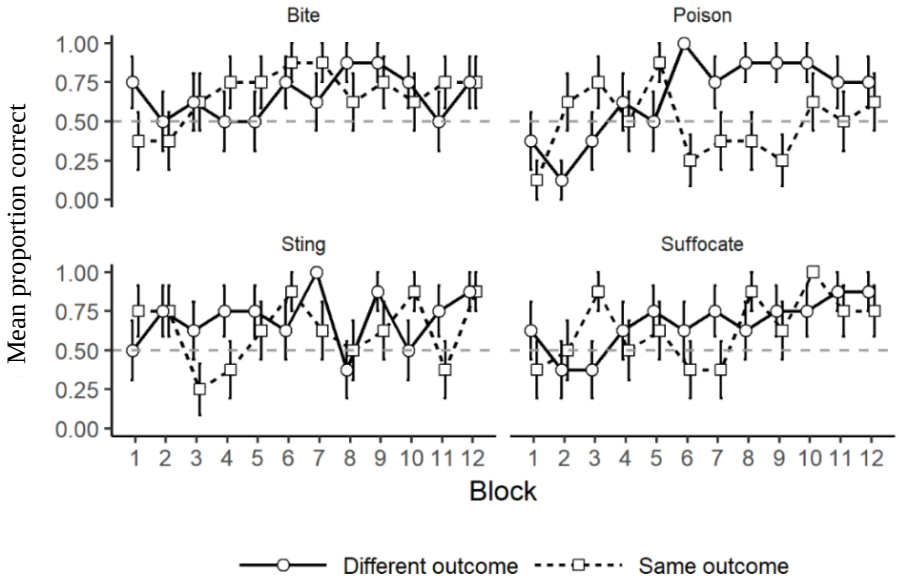
Data for the revaluation trials (A, B, S1 and S2) are summarised in **Figure 23(b)**. The discrimination of the non-configural stimuli was once again trivially easy for both groups. The advantage for group Different transferred from Stage 1 to revaluation trials. A mixed ANOVA confirmed a main effect of group, $F(1, 56) = 4.12$, $p = .047$, $\eta_p^2 = .07$, 90% CI [0.00, 0.19], albeit marginal and with CI indicating

a lack of confidence in the effect. No other main effects or interactions were significant when analysing the non-configural stimuli (smallest $p = .110$ for the interaction between outcome and block).

Data for revaluation trials A and B suggest participants learned this discrimination progressively. A mixed ANOVA revealed a main effect of block, $F(8.25, 462) = 3.18$, $p = .001$, $\eta_\rho^2 = .05$, 90% CI [0.01, 0.07] and an interaction between group, outcome and block, $F(24.75, 462) = 2.16$, $p = .001$, $\eta_\rho^2 = .10$, 90% CI [0.02, 0.10]. The source of the 3-way interaction was examined using four 2 x 12 ANOVAs on the data to see how group and block interacted with each outcome, as shown **Figure 24**. No significant main effects of group or block and no interaction between group and block were found for the 'bite' (smallest $p = .406$) or 'suffocate' outcomes (smallest $p = .094$). No main effects were found for the 'sting' outcome (smallest $p = .623$), but the analysis yielded a significant interaction between group and block, $F(11, 154) = 2.21$, $p = .016$, $\eta_\rho^2 = .01$, 90% CI [0.01, 0.16]. The source of this interaction was examined using simple main-effect analysis using separate error terms for each level of block, which revealed no group differences at any block, smallest $p = .059$ at block 7. The analysis on the 'poison' data revealed no main effects of group or block (smallest $p = .086$) but a group by block interaction, $F(11, 154) = 3.31$, $p < .001$, $\eta_\rho^2 = .08$, 90% CI [0.06, 0.23]. The source of this interaction was examined using simple main-effects analysis with separate error terms at each level of block. This revealed that group Different performed reliably better than group Same in block 2, $F(1,14) = 5.09$, $p = .041$, block 6 $F(1,14) = 21.00$, $p < .001$, block 8, $F(1,14) = 5.09$, $p = .041$ and block 9, $F(1,14) = 9.21$, $p = .009$. These results suggest that the advantage in performance for group Different, already evident during training, was also present during revaluation trials in some subsets of data.

Figure 24

Mean Correct Performance for Revaluation Trials (A and B) collapsed over Group for each Stimulus Outcome



Note. Error bars represent SEM.

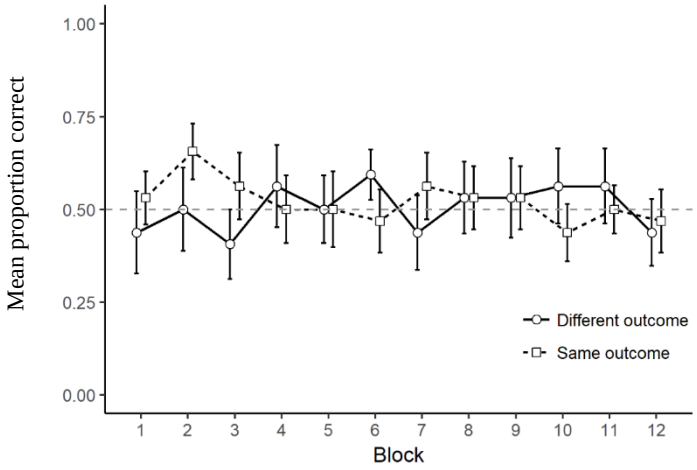
3.2.2.3 Stage 2 – Test trials

Test trials consisted only of octopus stimuli, hence C and D were collapsed over group for analyses and are summarised in **Figure 25**. Here, we anticipated an advantage for group Different, which had experienced different stimulus-outcome contingencies during revaluation trials. However, a mixed ANOVA with the within-subjects factor of block and the between-subjects factor of group and outcome yielded no main effects or interactions (smallest p = .110 for the interaction between group, outcome and allowance). Of most importance, the analysis revealed no main effect of group, $F(1, 56) = 0.06$, p = .811, η_p^2= .00, 90% CI [0.00, 0.05]. Furthermore, two one-sample t-tests, run to assess participants' performance against chance (50%), revealed that participants responses were not reliably different from

chance in both the Same, $t(15) = 0.56$, $p = .290$, $d = .14$, 90% CI [-0.28, 0.55] or Different groups, $t(15) = .09$, $p = .466$, $d = .02$, 90% CI [-0.39, 0.43]. That is, our participants failed to show an acquired equivalence effect regardless of the group they were assigned to.

Figure 25

Mean Correct Performance for Test Trials (C and D) collapsed over Group



Note. Error bars represent SEM.

Results from Experiment 6 are hard to interpret for a number of reasons. For the first time since we run our configural acquired equivalence task, we failed to obtain reliable learning during the initial training stage. Because acquired equivalence relies upon participants' learning the initial contingencies between stimuli and outcomes, it is not surprising that we did not observe performance above chance during test trials. Although this does not explain the absence of an acquired equivalence effect in the Different group, which did learn the initial discrimination reliably, it is worth noting that the results were overall lower than in previous experiments. A comparison between the mean accuracy over the second half of

training for the Different groups in Experiment 5 (M = .62) and Experiment 6 (M = .60) seems to suggest groups performed similarly. However, a closer look at the effect sizes shows that the effect was less robust in Experiment 6 (d = 0.50, 90% CI [0.06, 0.94]) when compared to initial learning of the Different group in Experiment 5 (d = .64, 90% CI [0.32, 0.96]), particularly evident when looking at the lower bound of the confident interval. Additionally, the group differences observed before the groups had any differential treatment invalidate the interpretation of any possible group differences in later stages of the task. We believe the general decrease in performance during Stage 1 and in all subsequent stages of the experiment may have been the result of a poor choice of response keys. Whilst we were motivated by wanting to keep Experiment 6 as similar as possible to previous experiments to allow for meaningful comparisons, it is possible that confounding two outcomes with a single response key might have confused participants, and overall hinder performance. Experiment 7 addressed these limitations by replicating Experiment 6 with a different choice of response keys.

3.3 Experiment 7

Experiment 6 failed to detect any group differences at test in our configural and non-configural acquired equivalence task. However, any interpretation of the absence of group differences was invalidated by the failure of Experiment 6 to obtain reliable learning during training and by the presence of group differences before groups had received differential treatment. We reasoned that a poor choice of response keys (left arrow – *Bite* or *Poison* and right arrow – *Sting* or *Suffocate*) might have resulted in the overall decrease in performance. Experiment 7 was intended as a direct replication of Experiment 6, and differed only in the choice of response keys. Just like in the previous experiment, the addition of the non-configural stimuli was intended to ensure both groups were exposed to all possible outcomes from the onset of the task, reducing any potential non-specific effects in performance as a result of the differential group treatment. To decouple the outcome from the response key, participants in Experiment 7 responded by choosing from four different keys (i.e., *T, D, V* and *H*).

3.3.1 Method

3.3.1.1 Participants

32 students from the University of Nottingham participated (26 women and 6 men, M_{age} = 22.41, *SD* = 3.81, range: 18-35). Participants were recruited using posters and the School of Psychology online booking system. They received module credits or an allowance for their participation.

3.3.1.2 Procedure

In this task, participants were asked to choose between the following response keys: *T – Poison, D – Bite, V – Suffocate* and *H – Sting.* Keys were located in the centre of the keyboard and were spatially orthogonal. Black stickers were used to help participants focus on these four keys. During each trial, the keypress reminder displayed on screen matched the spatial mapping of the keys to aid discrimination. For example, *D – Bite* appeared to the left of the screen and below *T – Poison,* whereas *H – Sting* appeared to the right of the screen and below *T – Poison*). Specifically, the words were displayed as follows in the x and y axes, respectively: Bite (-8, -6), Sting (8, -6), Poison (0, -5), Suffocate (0, -9), where (0, 0) indicates the centre of the screen. All unspecified details were identical to those of Experiment 6.

3.3.2 Results and Discussion

3.3.2.1 Stage 1

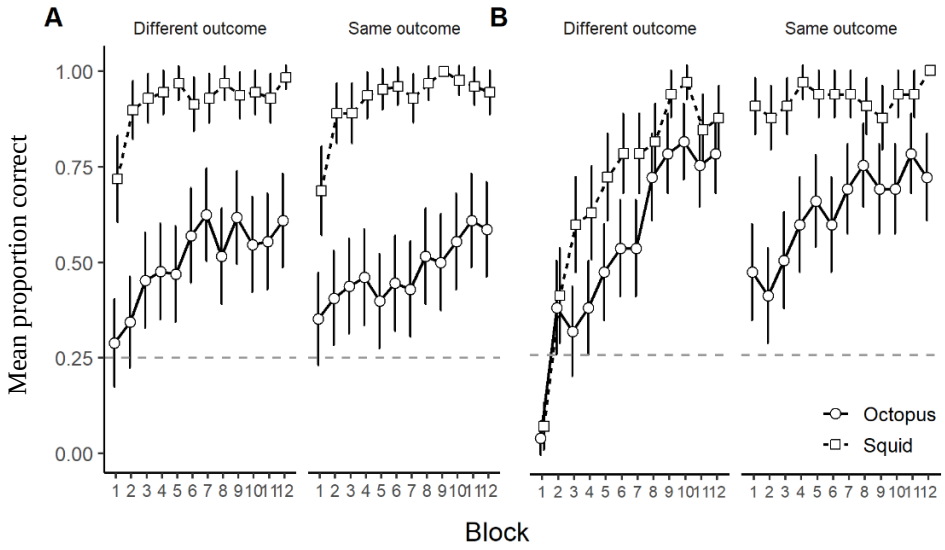
Just like in the previous experiment, the data from Stage 1, summarised in **Figure 26(a)**, show that the discrimination of the non-configural squid was trivially easy when compared to the configural octopuses. Once again we conducted separate analyses on these stimuli. When discriminating non-configural stimuli, participants reached and maintained levels close to asymptote throughout training. A mixed ANOVA with a within-subjects factor of block and between-subjects factors of group (Same vs. Different) and outcome revealed a main effect of block, $F(6.05, 338.80) = 18.05$, $p < .001$, $\eta_\rho^2 = .24$, 90% CI [0.17, 0.29]. The analysis also showed a main effect of outcome, $F(3, 56) = 4.03$, $p = .011$, $\eta_\rho^2 = .18$, 90% CI [0.02, 0.29]. Holm corrected pairwise comparisons showed that participants performance was

reliably better when the outcome was ‘bite’ (M = .94) than when it was ‘poison’ (M = .88) (p = .002). Performance was also reliably better when the outcome was ‘sting’ (M = .96) than when it was ‘poison’ or ‘suffocate’ (M = .91) (p < .001 and p = 001, respectively). Importantly, the analysis showed no main effect of group or interactions involving group (smallest p = .085 for the interaction between group and block).

Although the discrimination of the configural stimuli was harder, the data show a clear improvement in performance compared to Experiment 6, with participants progressively acquiring the discrimination and performance evidently above chance. Of critical importance, learning seemed to progress similarly in both groups. A mixed ANOVA on these data confirmed this observation. The analysis yielded a main effect of block, $F(7.48, 418.88) = 8.05$, $p < .001$, $\eta_\rho^2 = .13$, 90% CI [.07, .16], but no other main effects or interactions were significant (smallest p = .069 for the interaction between group and block). Groups Same and Different did not receive any differential treatment during training, these results are critical because they confirm learning proceeded similarly in both groups and allow for any subsequent differences to be interpreted. Additional one-sample t-tests, conducted to assess performance against chance (25%), showed that participants were performing reliably above chance in the Different, $t(15) = 7.66$, $p < .001$, $d = 1.92$, 90% CI [1.19, 2.61] and Same groups, $t(15) = 6.62$, $p < .001$, $d = 1.66$, 90% CI [1.00, 2.28], confirming that both groups had learned the initial discrimination by the end of Stage 1 training.

Figure 26

Collapsed Mean Performance for Stage 1 and Revaluation trials of the Acquired Equivalence Task with Configural and Non-Configural Trials



Note. (A) Mean correct performance for configural (octopus) and non-configural (squids) stimuli during Stage 1 for group Same and Different. Each block comprised the presentation of stimuli (AW, AX, BW, BX, CW, CX, DW, DX, S1, S2). (B) Mean correct performance for stimuli during revaluation trials for the Same and Different outcomes groups. Each block comprised the presentation of 4 stimuli (A, B, S1, S2), which were followed by feedback. Error bars represent SEM. The horizontal dashed line indicates chance performance.

3.3.2.2 Stage 2 – Revaluation trials

Data for the revaluation trials (A, B, S1 and S2) are summarised in **Figure 26(b)**. When learning about the non-configural stimuli, it is evident that group Same continued to perform at asymptote throughout the revaluation stage. Group Different experienced an initial decline in performance but recovered progressively. A mixed ANOVA with the between-subject variables of group and outcome and the within-subject variable of block on these squid trials revealed a main effect of group, $F(1, 56) = 23.73$, $p < .001$, $\eta_p^2 = .30$, 90% CI [.14, .43], with group Same performing

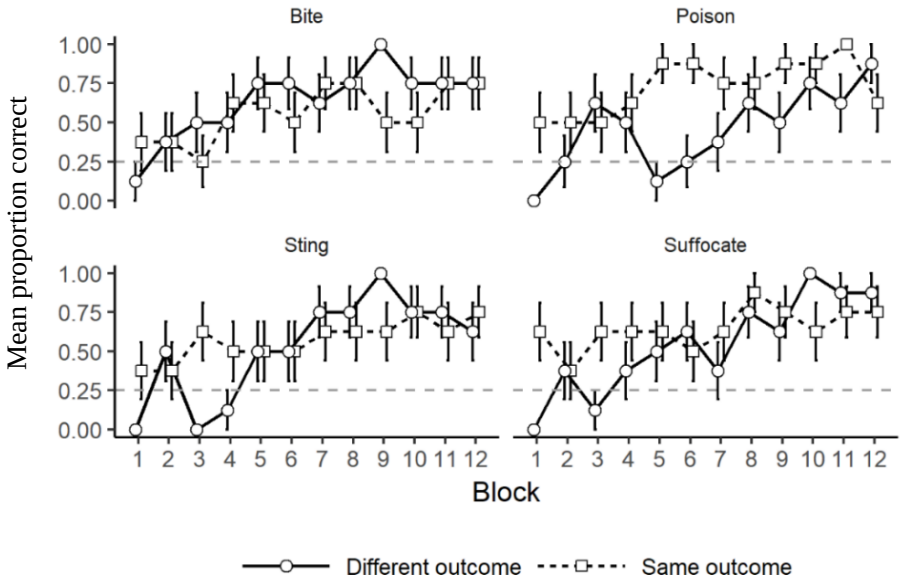
reliably better than group Different ($p < .001$). The analysis also yielded a main effect of block, $F(6.60, 369.60) = 13.58$, $p < .001$, $\eta_\rho^2 = .20$, 90% CI [.13, .24] and an interaction between the two, $F(6.60, 369.60) = 11.40$, $p < .001$, $\eta_\rho^2 = .17$, 90% CI [.10, .21]. No other main effects or interactions were significant when analysing the non-configural squids (smallest $p = .653$ for the interaction between group, outcome and block).

Revaluation trials A and B followed a similar pattern. A mixed ANOVA revealed a main effect of block, $F(8.03, 449.68) = 12.57$, $p < .001$, $\eta_\rho^2 = .18$, 90% CI [.12, .22], reflecting participants progressive learning of the discrimination. Here, the analysis revealed no main effect of group but a significant interaction between group, outcome and block, $F(24.09, 449.68) = 1.64$, $p = .029$, $\eta_\rho^2 = .08$, 90% CI [.00, .08]. Again, the CI suggest a lack of confidence in the effect (Lee, 2016). However, we still examined the 3-way interaction to have a better understanding of how learning proceeded per group. Data were split by outcome and four 2 x 12 ANOVAs were run to examine how group and block interacted with each outcome, as summarised in **Figure 27**. The analysis showed a main effect of block when the outcome was 'bite', $F(11, 154) = 5.55$, $p = .021$ but no main effect of group or interaction between the two. Analysing the outcome 'poison' revealed a main effect of group, $F(1, 14) = 5.71$, $p = .031$ and block, $F(11, 154) = 2.80$, $p = .002$ and an interaction between the two variables, $F(11, 154) = 2.06$, $p = .027$. This interaction was examined using simple main-effects analysis with separate error terms at each level of block. Examination of the data revealed that group Same performed reliably better than group Different in block 1, $F(1, 14) = 7.00$, $p = .021$; block 5, $F(1, 14) = 18.00$, $p = .001$ and block 6, $F(1, 14) = 9.21$, $p = .011$. The analysis of the 'sting' data revealed a main effect of block, $F(11, 154) = 4.31$, $p < .001$ and an interaction

between group and block, $F(11, 154) = 2.07$, $p = .025$. Inspection of this interaction, using simple main-effects analysis with separate error terms at each level of block, showed that group Same performed reliably better than group Different only in block 3, $F(1, 14) = 11.67$, $p = .004$. There was a main effect of block when the outcome was 'suffocate', $F(11, 154) = 3.81$, $p < .001$ and an interaction between group and block, $F(11, 154) = 2.10$, $p = .023$. Analysis of simple main-effects with separate error terms at each level of block revealed that group Same performed reliably better than group Different in block 1, $F(1, 14) = 11.67$, $p = .004$ and in block 3, $F(1, 14) = 5.09$, $p = .041$. That is, the advantage for group Different evident in Experiment 5 already during revaluation trials was not observed in Experiment 7.

Figure 27

Mean Correct Performance for Revaluation Trials (A and B) collapsed over Group for each Stimulus Outcome



Note. Error bars represent SEM.

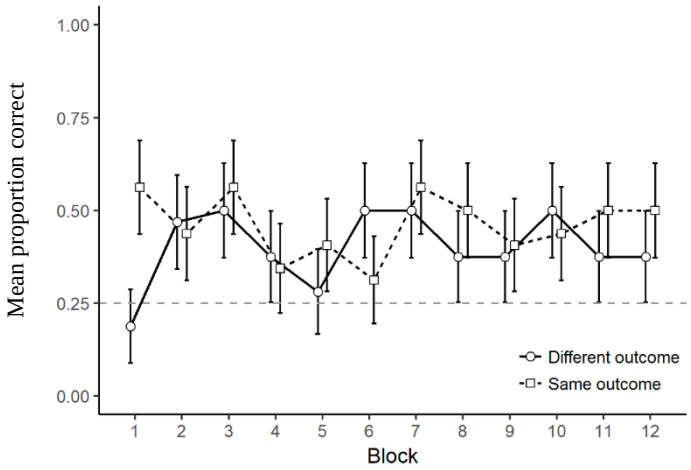
3.3.2.3 Stage 2 – Test trials

Test data consisted only of octopuses, hence trials were collapsed over group for analyses. Inspection of the data, summarised in **Figure 28**, suggested participants reliably generalised responding to stimuli C and D without explicit training. This observation was confirmed by two one-sample *t*-tests, run to assess participants' performance against chance (25%). Participants in both the Same, $t(15) = 3.22$, $p = .003$, $d = 0.81$, 90% CI [0.32, 1.27] and Different group, $t(15) = 2.69$, $p = .008$, $d = 0.67$, 90% CI [0.21, 1.12] performed significantly above chance on test trials. That is, both groups showed the acquired equivalence effect.

Here, and after incorporating non-configural cues to account for potential non-specific effects, an advantage in performance for group Different would have been a robust replication of the results observed in Experiment 5. However, a mixed ANOVA with the within-subjects factor of block and the between-subjects factor of group and outcome showed no main effects or interactions (smallest $p = .057$ for the interaction between group and block). Of critical importance to our experiment, the analysis revealed no main effect of group, $F(1, 56) = 0.70$, $p = .406$, $\eta_\rho^2 = .01$, 90% CI [0.00, 0.11].

Figure 28

Mean Correct Performance for Test Trials (C and D) collapsed over Group



Note. Error bars represent SEM.

These results are not what we anticipated following results from Experiment 5, where participants who experienced different outcomes across stages performed reliably better during revaluation and test trials. On the one hand, these results could indicate that the differences in performance observed in Experiment 5 were indeed explainable in terms of simple arousal or increased interest in the task, instead of reflecting the changes in the critical connections across the network, and that the presentation of all outcomes from the start reduced or eliminated theses effects. If this were the case, this could be seen as a challenge to the network which we reasoned should anticipate differences in performance between groups Same and Different (see General Discussion). On the other hand, the results could suggest that intermixing configural and non-configural trials could have consequences for the development of connections across a network with the characteristics of Honey's. In Experiment 7, the non-configural squid stimuli each had full predictive value (i.e., each signalled an outcome unambiguously). On the other hand, configural stimuli

were comprised of ambiguous elements, which had full predictive value only when presented in specific combinations. Evidence shows that factors such as the previous experience with non-configural stimuli can influence subsequent configural learning in humans and other organisms (Melchers et al., 2008). For example, initial training with the non-configural squid stimuli could have encouraged an incorrect non-configural approach to the configural octopus stimuli, leading to a different internal organisation of the corresponding input-to-hidden and hidden-to-output connections and influencing subsequent performance in the task. In any case, the results from this set of experiments are inconclusive in either supporting or challenging the description of Honey's network because we are basing our interpretations in an informal description of the model. The following chapter will compliment this analysis with the use of one formal implementation of the network described by Honey and colleagues.

So far in this thesis, I have presented configural acquired equivalence tasks as a means to illustrate acquired equivalence that cannot be explained in terms of mediated conditioning, but is amenable to a connectionist approach. However, some non-configural discriminations can serve as examples as well. The following three experiments, based on experiments in Delamater (1998), will investigate the effects of outcome manipulations in these type of non-configural discriminations.

Outcome manipulations in non-configural acquired equivalence

Delamater (1998) demonstrated that reversal discrimination learning proceeds more readily when stimuli from the same modality are trained with different outcomes across stages than when they are trained with the same outcomes across stages. In Experiment 3 of this paper, rats were trained on a discrimination between two auditory and two visual stimuli in which only one stimulus from each modality was paired with a distinct outcome, A1+, A2-, V1- and V2*, where '+' and '*' represent food pellets and sucrose, respectively. Once rats had learned the discrimination, one group received a reversal with the same outcomes within modality (i.e., A1-, A2+, V1* and V2-). A second group of rats received a reversal stage in which stimuli signalled different outcomes within each modality (i.e., A1-, A2*, V1+ and V2-). The group that experienced different outcomes within modality learned the discrimination more rapidly than the group with the same outcomes within modality.

These results are significant because just like in the previous set of experiments, they cannot be explicable in terms of DOE. That is, assuming the initial outcome expectancies are as follows: A1 – O1, A2 – O2, V1 – O3 and V2 – O4, it is not evident how the associations formed during training would facilitate learning for subjects in the group with different outcomes within stimulus modality (A1 – O3, A2 – O4, V1 – O1 and V2 – O2) any more than subjects in the other group (A1 – O2, A2 – O1, V1 – O4 and V2 – O3). Additionally, these results are not anticipated by a mediated conditioning account. In the same vein as in a configural acquired equivalence task, outcomes alone are non-informative to the discrimination. Consider as an example a discrimination that goes from A1+, A2- in training to A1-, A2+ during reversal. By the end of training, A1 will evoke the representation of the

outcome (e.g., food). During the reversal, the presentation of A1 will initially continue to evoke the representation of food. However, these trials will not be reinforced during the reversal. Conversely, A2 will not evoke the representation of food at the start of the reversal stage. However, reinforcement will now occur in these trials. That is, the outcome representation evoked by each stimulus will be non-informative during the reversal, because sometimes it will signal reinforcement and sometimes it will not.

Instead, these ideas are naturally captured by a connectionist network approach, based on the acquired equivalence and distinctiveness of the internal representations. When stimuli from the same modality are paired with the same outcomes, the internal representations of those cues become more equivalent. Conversely, when stimuli from the same modality are paired with different outcomes, their representations become more distinct and, therefore, easier to discriminate, resulting in the observed enhanced performance. Indeed, Delamater offered this alternative explanation and presented a connectionist network that simulated performance in these tasks successfully, by dint of allowing the hidden unit layer to have a set of hidden units dedicated to process each stimulus modality (auditory and visual), in addition to a multimodal set of hidden units (Delamater, 2012). In this chapter's discussion however, I will argue that Honey's network could theoretically account for these differences in performance without the need to invoke separate sets of modality-specific and multimodal hidden units.

The next set of three experiments sought to replicate these findings using tasks inspired by Experiment 3 in Delamater (1998) and to extend the generality of these findings by investigating the effect of within and between modality outcome manipulations in a non-configural acquired equivalence task with human

participants. Experiment 8 failed to find group differences in performance during a reversal stage that used either same outcomes within modality or different outcomes within modality. Experiment 9 rectified the counterbalancing of Experiment 8 and included additional exemplars to increase the difficulty of the task. However, it still failed to find group differences. Experiment 10 attempted a more direct replication of Delamater's (1998) experiment by using stimuli from the auditory and visual modalities. Once again, we failed to obtain any group differences in performance.

3.4 Experiment 8

Experiment 8 was intended to investigate the effects of reversals with the same or different outcomes within modality in acquired equivalence. To that end, participants were assigned to either group *Same* or group *Different,* as exemplified in **Table 8**, to complete a non-configural acquired equivalence task inspired by Experiment 3 in Delamater (1998). In this task, participants were asked to imagine themselves walking around a forest where they could encounter different wild animals, and to determine in which direction each animal would run away upon encountering them. B1 and B2 represent two distinct brown bears and S1 and S2 two snakes, together they made our two distinct stimulus dimensions. Our stimuli depart from those used in Delamater's experiments. Instead of using stimuli from two distinct modalities (auditory and visual), we chose distinct visual stimuli. We simply reasoned that the bear and snake stimuli were sufficiently different to be considered as belonging to two discrete stimulus dimensions, and that participants might find these stimuli more engaging as part of a discrimination learning task.

L, R, U and D represent 'left', 'right', 'up' and 'down', the four directions in which any given animal could escape. For example, trial B1 – L indicates that bear B1 would be followed by outcome 'left'. During reversals, stimuli continued to signal the same outcomes within dimension for group Same. For participants in group Different, stimuli now signalled different outcomes within dimension.

Here, it is important to note that each stimulus dimension signalled outcomes from one spatial plane only. For example, in a counterbalancing subgroup, bear stimuli might move left and right, and snake stimuli up and down, but the spatial planes would never be interchanged between stimulus modalities. The choice of outcomes in the task differs from the original Delamater (1998) experiment in a

number of ways, which will be discussed in more detail in the General Discussion. However, although all stimuli signalled distinct outcomes during training, we reasoned that the internal representations of stimuli from each dimension should still become more equivalent, by dint of signalling outcomes limited to a spatial axis: *B1 – Left – Horizontal, B2 – Right – Horizontal, S1 – Up – Vertical* and *S2 – Down – Vertical*. If we assume this is true, it follows that a discrimination with different outcomes within dimension should still be easier to learn than a reversal with the same outcomes within dimension.

Table 8

Experimental Design for Experiment 8

Training	Reversal	Reacquisition
a. Same outcomes within dimension		
B1 - **L**	B1 - **R**	B1 - **L**
B2 - **R**	B2 - **L**	B2 - **R**
S1 - **U**	S1 - **D**	S1 - **U**
S2 - **D**	S2 - **U**	S2 - **D**
b. Different outcomes within dimension		
B1 - **L**	B1 - **U**	B1 - **L**
B2 - **R**	B2 - **D**	B2 - **R**
S1 - **U**	S1 - **L**	S1 - **U**
S2 - **D**	S2 - **R**	S2 - **D**

Note. B1 and B2 represent the bear stimuli and S1 and S2 the snake stimuli. The Left response (L) required participants to press D on the keyboard. The Right response (R) required participants to press H. The Up response (U) required participants to press T, and the Down response (D) to press V. The response keys were spatially mapped on the keyboard to indicate left, right, up and down, respectively.

3.4.1 Method

3.4.1.1 Participants

32 students from the University of Nottingham participated (25 women and 7 men, M_{age} = 24.78, *SD* = 4.83, range: 18-37). Participants were informed about the task and debriefed upon completion. All agreed to participate. Participants received module credits or a small allowance for their participation. Participants were recruited using posters and the School of Psychology online booking system.

3.4.1.2 Apparatus & Materials

The stimuli were two front-facing images of cartoon brown bears and two front-facing images of cartoon snakes in full colour, 6 (width) x 6 (height) cm in size, as shown in **Figure 29**. Images were presented on a grey background. During training, the contingencies between stimuli and outcomes were counterbalanced so that for half of the participants, bears moved in the horizontal plane and snakes moved in the vertical plane (B1-L, B2-R, S1-U, S2-D or B1-R, B2-L, S1-D, S2-U). For the other half of the participants, bears and snakes moved in the complementary plane (B1-U, B2-D, S1-L, S2-R or B1-D, B2-U, S1-R, S2-L). During the reversal stage, stimuli in the Same outcomes within dimension signalled the outcome opposite to the one they had signalled during training within the same spatial dimension. For example, for a participant for whom stimuli had signalled B1-L, B2-R, S1-U, S2-D during training, stimuli now signalled B1-R, B2-L, S1-D, S2-U, resulting in four possible counterbalancing subgroups. Stimuli in the Different outcomes within dimension signalled an outcome from the complementary spatial dimension compared to training. For example, given the training B1-L, B2-R, S1-U, S2-D, stimuli in the reversal stage signalled B1-U, B2-D, S1-L, S2-R. Contingencies

in the reversal stage for this group were partially counterbalanced, resulting in four counterbalancing subgroups.

Figure 29

Bear and Snake Stimuli used in Experiment 8

Note. Only one stimulus was present in any given trial

3.4.1.3 Procedure

Participants were assigned randomly without replacement to groups *Same* outcomes within dimension and *Different* outcomes within dimension to ensure the desired number of participants per group (n = 16 each). All participants read an instruction sheet that emphasize participants' rights to terminate the task at any time. After ensuring participants had understood the task, the experimenter left the room and returned once the task was finished to debrief participants.

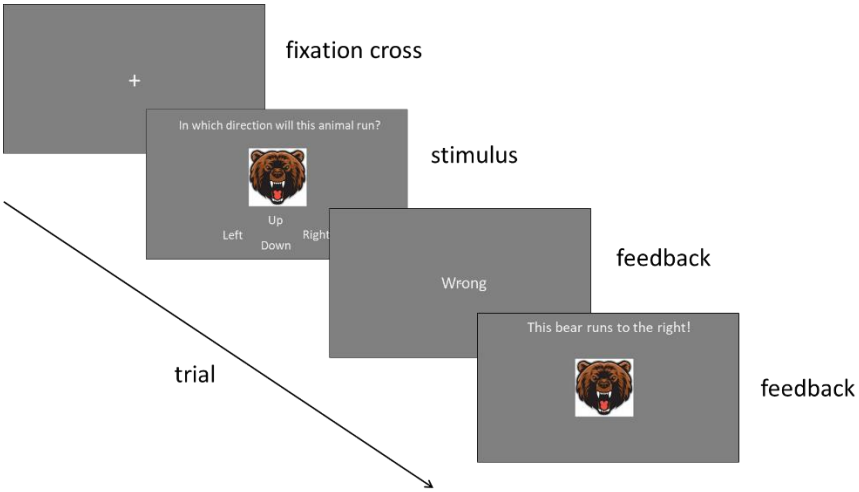
Prior to the start of the experiments, participants read a set of written instructions asking them to *"imagine yourself walking in a forest. You will encounter some wild animals as you walk around. Luckily, all wild animals will run away from you when they see you. It is your task to look at the different animals and learn to predict in which direction they will run away. You will have 5 seconds to guess the direction. Initially you will have to guess, but you will receive feedback on your responses so you can learn to respond correctly."* Participants were instructed to press the following keys: left – press D, right – press H, up – press T and down –

press V. Black stickers were used to highlight the four response keys on the keyboard.

The training stage comprised the presentation of trials B1, B2, S1, S2 twice per block over four blocks (32 trials). The presentation of stimuli was block randomised. Each trial begun with a fixation cross located in the centre of the screen for 0.5 s. A stimulus was then presented in the centre of the screen for 5 s. The sentence *"In which direction will this animal runaway?"* remained on top of the stimulus during the duration of the trial. The words "left", "right", "up", and "down" were displayed below the stimulus throughout the duration of the trial. The location of the responses was mapped in accordance with the location they indicated. For example, the word "left" appeared to the left and below the word "up". Specifically, the words were displayed as follows in the x and y axes, respectively: left (-8, -6), right (8, -6), up (0, -5), down (0, -9), where (0, 0) indicates the centre of the screen. After the participant's response, the feedback *"CORRECT!"* or *"wrong"* appeared in the centre of the screen for 1 s, followed by the picture of the stimulus and the text *"This bear runs to the left!"* (for participants for whom a specific bear signalled a left response key) for 2 s. During feedback, stimuli were made to move across the appropriate horizontal or vertical axis to give the illusion of movement towards the corresponding direction. An example trial is shown in **Figure 30**.

Figure 30

Example Layout of a Trial during Experiments 8 and 9



Note. The stimulus displayed during feedback moved across the screen in the corresponding direction.

A reversal stage followed immediately after training. The reversal comprised the presentation of stimulus B1, B2, S1, and S2 twice per block over two blocks (16 trials). Stimuli were block randomised. Trials developed as previously described, but outcomes were reversed with regards to the training stage. For half of the participants (group Same outcomes within dimension), outcomes for each stimulus dimension (bears or snakes) were reversed within each (horizontal and vertical) axis. For example, the bear requiring a left response during training required a right response during the first reversal stage. For the other half of the participants (group Different outcomes within dimension), outcomes for each stimulus dimension were reversed across spatial axes. For example, a snake requiring a left response during training required an up response during the first reversal. Participants completed a total of four reversal stages (i.e., Reversal – Reacquisition – Reversal – Reacquisition, 64 trials) and received no indication that a reversal stage was about to start.

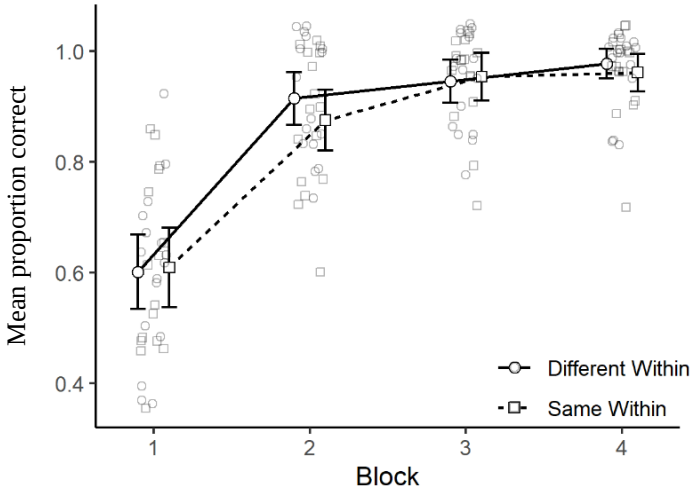
3.4.2 Results and Discussion

3.4.2.1 Training data

Data from the training stage of Experiment 8 are presented in **Figure 31**. The data indicates participants' progressive acquisition of the initial contingencies, reaching and maintaining levels of performance close to asymptote from block two. The training for both groups was identical and no group differences were expected at this stage. A mixed ANOVA with the between-subjects factors of group (Same vs Different outcomes within dimension) and the within-subjects factor of stimulus (B1, B2, S1, and S2), outcome (left, right, up, and down) and block (blocks one to four) was conducted on training data. Of critical importance, the analysis yielded no main effect of group, $F(1, 30) = 0.28$, $p = .598$, $\eta_\rho^2 < .001$, 90% CI [0.00, 0.04] and no interactions involving group (smallest $p = .203$ for the interaction Group x Outcome x Block), confirming no group differences at this stage.

Figure 31

Collapsed Mean Performance for the Training Stage of the Acquired Equivalence Task.



Note. Each block comprised the presentation of stimuli B1, B2, S1, S2 twice. Error bars represent SEM. Semi-transparent grey circles and squares represent participants' average individual performance in each group.

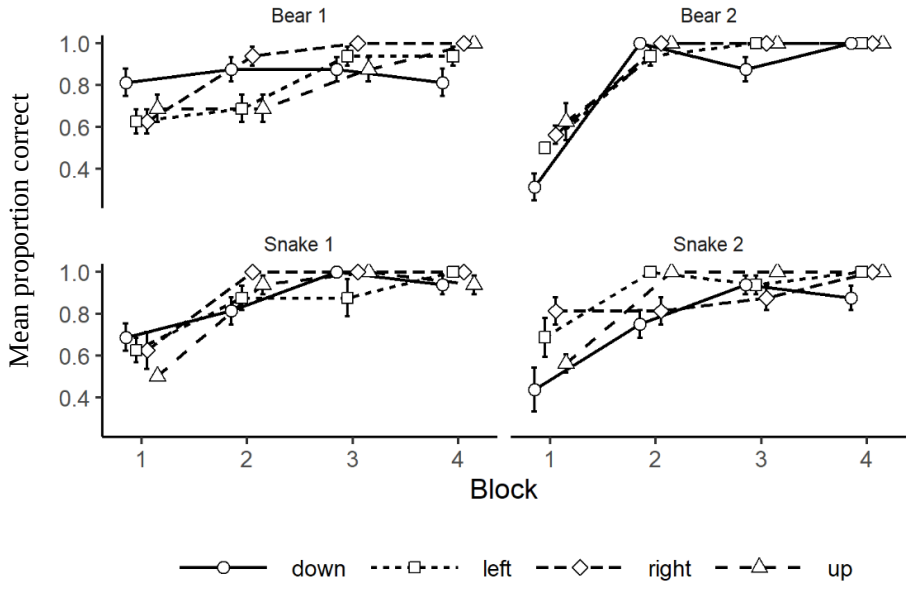
The analysis confirmed a main effect of block, $F(3, 288) = 105.40$, $p < .001$, $\eta_\rho^2 = .52$, 90% CI [0.46, 0.57], reflecting participants' acquisition of the discrimination. Additionally, the analysis yielded no main effect of stimulus, $F(3, 96) = 0.49$, $p = .692$, $\eta_\rho^2 = .01$, 90% CI [0.00, 0.05] but a main effect of outcome, $F(3, 96) = 3.07$, $p = .031$, $\eta_\rho^2 = .09$, 90% CI [0.00, 0.16]. However, the 90% CI of the effect size suggested the effect was not reliable and, upon closer examination, only the difference between performance for the right ($M = .89$) and the down outcomes ($M = .81$) approached significance at $p = .07$.

The analysis also revealed a 3-way interaction between Stimulus, Outcome and Block, $F(22.41, 239.04) = 2.11$, $p = .003$, $\eta_\rho^2 = .16$, 90% CI [0.03, 0.16]. The source of the interaction was examined using four 4 x 4 ANOVAs on the data split

per stimulus (B1, B2, S1, S2), as shown in **Figure 32**. The analysis revealed a main effect of block when the stimulus was B1 ($p < .001$) but no main effect of outcome or interaction between the two ($p = .576$ and $p = .280$, respectively). The same pattern of results was found for B2, with the analysis showing only a main effect of block ($p < .001$), but no main effect of outcome ($p = .067$) or interaction ($p = .107$). The analysis also showed a main effect of block for stimulus S1 ($p < .001$) but no main effect of outcome ($p = .463$) or interaction between the two ($p = .441$). The analysis revealed a main effect of block ($p < .001$) and a main effect of outcome ($p = .001$) for stimulus S2 but no interaction between both factors ($p = .120$). Pairwise comparisons following the main effect of outcome revealed that participants were reliably better responding to the left and up outcomes compared to the down outcome when presented with S2 ($p = .016$ and $p = .029$, respectively). No other interactions were significant.

Figure 32

Mean Correct Performance during Training collapsed over Outcome for each Stimulus



Note. Error bars represent SEM.

3.4.2.2 Reversals

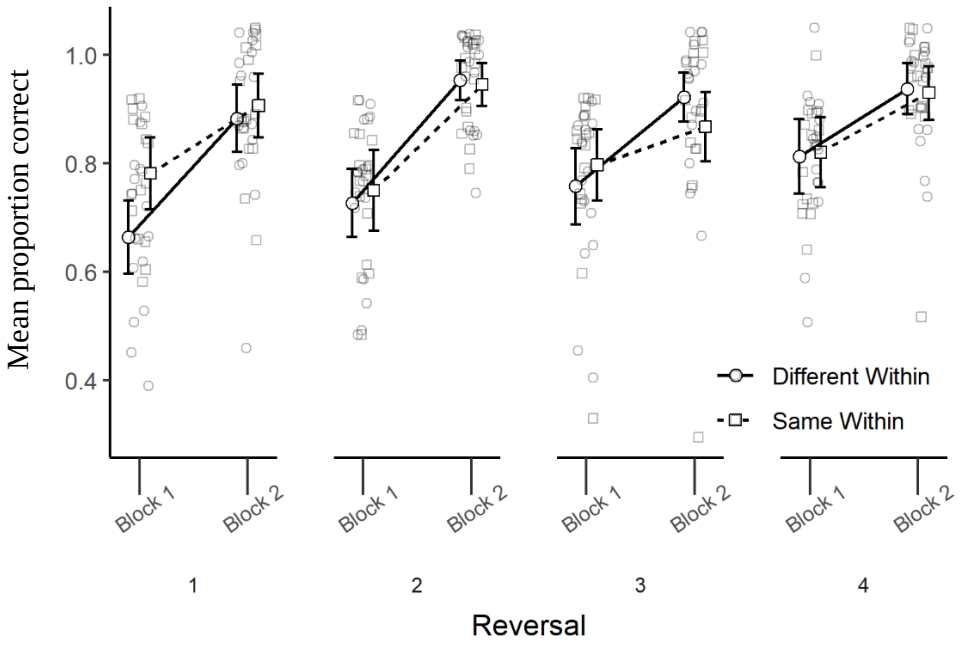
Data for the four reversal stages, further split by block within each reversal, are presented in **Figure 33**. In here, we expected performance to reflect the differential group treatments, with group Different performing better than group Same. Data were collapsed over stimulus and a mixed ANOVA with the between-subjects factors of group (Same vs Different outcomes within dimension) and the within-subjects factor of reversal (reversal one to four), outcome (left, right, up, and down) and block (block one and two) was conducted on reversal data.

Of critical importance to Experiment 8, the analysis did not show the anticipated main effect of group, $F(1, 30) = 1.05$, $p = .313$, $\eta_\rho^2 = .03$, 90% CI [0.00, 0.18], failing to reveal any differences based on the group's differential treatment during reversals. The analysis revealed a main effect of reversal, $F(3, 90) = 3.04$, $p =$

.003, η_p^2 = .11, 90% CI [0.00, 0.17]. Although the lower bound of the confident interval points at the effect lacking reliability, pairwise comparisons indicated that performance in reversal four was significantly higher than performance in the first reversal (p = .018). The ANOVA also yielded a main effect of block, F(1, 30) = 91.51, p < .001, η_p^2 = .75, 90% CI [0.60, 0.82], with participants overall performance in the second block of the reversals (M = .92) reliably better than overall performance in the first block of the reversals (M = .76) (p < .001). No reliable main effect of outcome was found, F(3, 90) = 1.85, p = .144, η_p^2 = .06, 90% CI [0.00, 0.13].

Figure 33

Mean Proportion of Correct Responses during Reversal stages



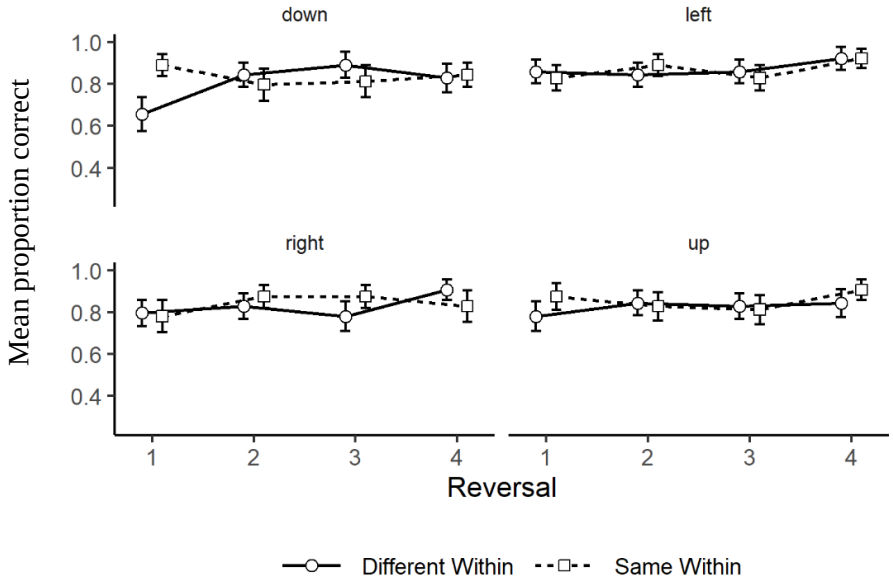
Note. Each block comprised the presentation of stimuli B1, B2, S1, S2 twice (16 trials per reversal). Error bars represent SEM. Semi-transparent grey circles and squares represent participants' average individual performance in each group.

The interaction involving Group, Reversal and Outcome was reliable, $F(1.89, 56.70) = 2.36$, $p = .025$, $\eta_\rho^2 = .07$, 90% CI [0.00, 0.18]. Despite the lower bound of the confident interval, we examined the 3-way interaction to have a better understanding of how these three variables interacted. Data were split by outcome and four 2 x 4 ANOVAs were run to examine the source of the 3-way interaction, as shown in **Figure 34**.

No main effects or interaction between group and reversal were found when the outcome was 'left' (smallest $p = .221$ for the main effect of stage). No main effects or interaction between group and reversal were found when the outcome was 'right' (smallest $p = .253$ for the interaction Group x Reversal). No main effects or interaction between group and reversal were found when the outcome was 'up' (smallest $p = .322$ for the main effect of group). The analysis revealed no main effects when the outcomes was 'down' (smallest $p = .318$ for the main effect of stage), but a significant interaction between group and reversal, $F(2.64, 79.20) = 4.12$, $p = .002$. This interaction was examined using simple main-effects analysis with separate error terms at each level of reversal. Examination of the data revealed that group Same performed reliably better than group Different during the first reversal, $F(1, 62) = 11.88$, $p < .001$.

Figure 34

Mean Correct Performance during Reversals collapsed over Group for each Outcome



Note. Error bars represent SEM.

Results failed to show the anticipated advantage in performance for group Different outcomes within dimension over group Same. These results do not accord with Delamater (1998). One possible explanation as to why we failed to observe any group differences could be that the task seemed trivially easy for participants. During training, participants reached levels of performance close to asymptote from block 2. That is, after having experienced each stimulus-outcome contingency only twice. It seems plausible to think that genuine group differences might have been masked by this. Additionally, the task was only partially counterbalanced for group Different. For example, outcome 'right' always became 'down', but never 'up' during the reversal. Experiment 9 attempted to rectify these two issues by increasing the number of exemplars presented and applying a full counterbalancing to the task.

3.5 Experiment 9

Results from Experiment 8 failed to show the anticipated increased performance in the group that experienced reversals with different outcomes within dimension. Experiment 9 sought to increase the difficulty and sensitivity of the task to find potential group differences in reversal performance. To that end, additional exemplars were added to the initial discrimination training, as summarised in **Table 9**. It also extended the counterbalancing of group Different with regards to the previous experiment.

Table 9

Experimental Design for Experiment 9

Training	Reversal	Reacquisition
a. Same outcomes within dimension		
B1 - **L** B2 - **R**	B1 - **R** B2 - **L**	B1 - **L** B2 - **R**
B3 - **R** B4 - **L**	B3 - **L** B4 - **R**	B3 - **R** B4 - **L**
B5 - **L** B6 - **R**	B5 - **R** B6 - **L**	B5 - **L** B6 - **R**
S1 - **U** S2 - **D**	S1 - **D** S2 - **U**	S1 - **U** S2 - **D**
S3 - **D** S4 - **U**	S3 - **U** S4 - **D**	S3 - **D** S4 - **U**
S5 - **U** S6 - **D**	S5 - **D** S6 - **U**	S5 - **U** S6 - **D**
b. Different outcomes within dimension		
B1 - **L** B2 - **R**	B1 - **U** B2 - **D**	B1 - **L** B2 - **R**
B3 - **R** B4 - **L**	B3 - **D** B4 - **U**	B3 - **R** B4 - **L**
B5 - **L** B6 - **R**	B5 - **U** B6 - **D**	B5 - **L** B6 - **R**
S1 - **U** S2 - **D**	S1 - **L** S2 - **R**	S1 - **U** S2 - **D**
S3 - **D** S4 - **U**	S3 - **R** S4 - **L**	S3 - **D** S4 - **U**
S5 - **U** S6 - **D**	S5 - **L** S6 - **R**	S5 - **U** S6 - **D**

Note. B1-B6 represent six bear stimuli and S1-S6 six snake stimuli. The Left response (L) required participants to press D on the keyboard. The Right response (R) required participants to press H. The Up response (U) required participants to press T, and the Down response (D) to press V. The response keys were spatially mapped on the keyboard to indicate left, right, up and down, respectively.

3.5.1 Method

3.5.1.1 Participants

32 students from the University of Nottingham participated (26 women and 6 men, M_{age} = 22.06, *SD* = 4.88, range: 18-38). Participants were informed about the task and debriefed upon completion. All agreed to participate. Participants received module credits or a small allowance for their participation. Participants were recruited using posters and the School of Psychology online booking system.

3.5.1.2 Apparatus & Materials

The stimuli were six front-facing images of cartoon bears and six front-facing images of cartoon snakes in full colour, 6 (width) x 6 (height) cm in size, as shown in **Figure 35**. Images were presented on a grey background. During training, the contingencies between stimuli and outcomes were counterbalanced so that for half of the participants, bears moved in the horizontal plane and snakes moved in the vertical plane, and for the other half of the participants, bears and snakes moved in the complementary plane. The reversal stage for the group with the Same outcomes within dimension proceeded just as in Experiment 8. Stimuli signalled an outcome from the complementary spatial plane compared to training in group Different. However, unlike in Experiment 8, stimuli-outcome contingencies were fully counterbalanced, resulting in eight counterbalancing subgroups. Like that, we ensured any one response (e.g., left) was reversed to any one response in the opposite spatial plane (i.e., up and down) equally often.

Figure 35

Bear and Snake Stimuli used in Experiment 9

Note. Only one stimulus was present in any given trial

3.5.1.3 Procedure

Participants were randomly assigned without replacement to groups *Same* outcomes within dimension and *Different* outcomes within dimensions (n = 16 each). All participants read an instruction sheet that emphasize participants' rights to terminate the task at any time. After ensuring participants had understood the task, the experimenter left the room and returned once the task was finished to debrief participants.

The training stage comprised the presentation of twelve trials: B1, B2, B3, B4, B5, B6, S1, S2, S3, S4, S5 and S6 presented once per block over eight blocks (96 trials). The presentation of the stimuli was block randomised. Trials were identical to those described in Experiment 8. A reversal stage followed the training stage. Each reversal comprised the presentation of each stimulus twice per block over two blocks

(48 trials). Stimuli were block randomised. During reversal, outcomes were reversed with regards to the training stage. For participants in group Same, outcomes for each stimulus dimension (bears or snakes) were reversed within each (horizontal and vertical) axis. For group Different, outcomes for each stimulus dimension were reversed across spatial axes. Participants completed two reversal stages (i.e., Reversal and Reacquisition) and received no indication that a reversal stage was about to start.

3.5.2 Results and Discussion

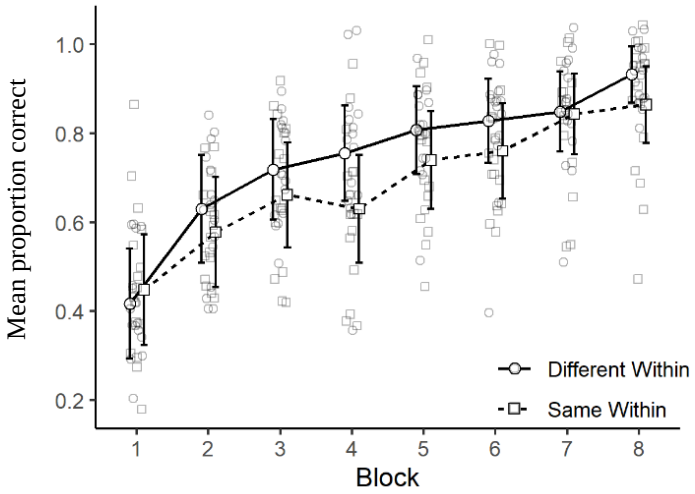
3.5.2.1 Training data

Data from the training stage are presented in **Figure 36**. In contrast to the training data from Experiment 8, inspection of the data indicates that participants learned the initial discrimination more progressively, reaching levels of performance close to asymptote later in training. This suggests that the inclusion of extra exemplars had the intended effect of increasing the task difficulty, albeit not drastically. Just like in the previous experiment, the training for both groups was identical and no group differences were expected at this stage. A mixed ANOVA with the between-subjects factors of group (Same vs. Different) and the within-subjects factor of stimulus, outcome and block was conducted on training data. Of critical importance, the analysis yielded no main effect of group, $F(1, 30) = 3.46$, $p = .073$, $\eta_p^2 = .10$, 90% CI [0.00, 0.28] and no interactions involving group (smallest $p = .298$ for the interaction Group x Outcome x Block), confirming no group differences at this stage.

The analysis also confirmed a main effect of block, $F(1.96, 58.8) = 47.82$, $p < .001$, $\eta_p^2 = .61$, 90% CI [0.47, 0.69], reflecting participants' progressive acquisition of the discrimination. No other main effects or interactions were significant (smallest $p = .314$ for the interaction between group and block).

Figure 36

Collapsed Mean Performance for the Training Stage of the Acquired Equivalence Task.



Note. Each block comprised the presentation of stimuli B1, B2, B3, B4, B5, B6 and S1, S2, S3, S4, S5, S6 once per block. Error bars represent SEM. Semi-transparent grey circles and squares represent participants' average individual performance in each group.

3.5.2.2 Reversals

Data for the two reversal stages, further split by block within each reversal, are summarised in **Figure 37**. We anticipated the inclusion of additional exemplars and the more progressive acquisition of the initial discrimination to facilitate the observation of possible group differences in performance. Just like in the previous

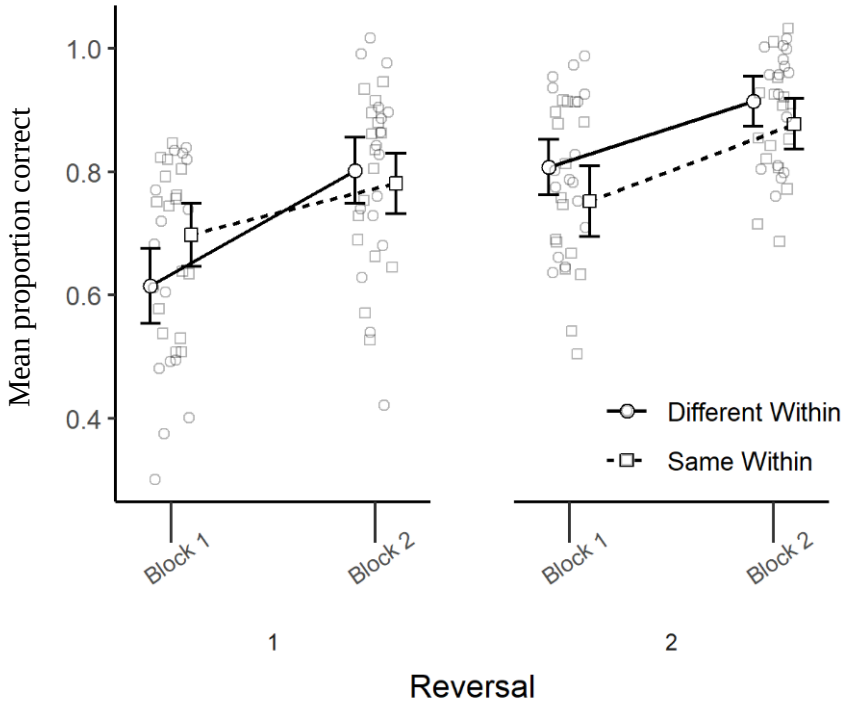
experiment, we expected group Different to perform better than group Same. Data were collapsed over stimulus and a mixed ANOVA with the between-subjects factors of group (Same vs. Different outcomes within dimension) and the within-subjects factor of reversal (first and second reversal), outcome (left, right, up, and down) and block (block one and two) was conducted on reversal data.

Although examination of the reversal data suggested a numerical advantage for group Different during the second reversal, the analysis did not yield the anticipated main effect of group, $F(1, 30) = 0.04$, $p = .849$, $\eta_\rho^2 = .002$, 90% CI [0.00, 0.07]. Of interest to our task, the analysis revealed an interaction between group and reversal, $F(1, 30) = 4.78$, $p = .037$, $\eta_\rho^2 = .14$, 90% CI [0.00, 0.32]. However, when data were split per reversal to examine the effect of group, no reliable main effects of group were found ($p = .512$ and $p = .193$ for the first and second reversal, respectively).

The analysis revealed a main effect of reversal, $F(1, 30) = 42.02$, $p < .001$, $\eta_\rho^2 = .58$, 90% CI [0.37, 0.70], with performance in reversal two reliably higher than performance in the first reversal ($p < .001$). The analysis also yielded a main effect of block, $F(1, 30) = 78.84$, $p < .001$, $\eta_\rho^2 = .72$, 90% CI [0.57, 0.80], with participants overall performance in the second block of the reversals ($M = .84$) reliably better than overall performance in the first block of the reversals ($M = .72$) ($p < .001$). No other reliable main effects or interactions were found.

Figure 37

Mean Proportion of Correct Responses during Reversal stages



Note. Each block comprised the presentation of stimuli B1-B6 and S1-S6 twice (48 trials per reversal). Error bars represent SEM. Semi-transparent grey circles and squares represent participants' average individual performance in each group.

Whilst the training data suggested that the additional stimuli had increased the difficulty of the initial discrimination, and reversal data seemed to suggest a numerical advantage for group Different, at least during the second reversal, the results from Experiment 9 once again failed to show any reliable group differences in performance. Another reason why we might not be observing any differences in performance could be attributed to one of the most evident differences between our task and Delamater's; the choice of stimulus dimensions. Experiment 8 and 9 used two distinct dimensions within the visual modality instead of stimuli from different modalities. We chose the stimuli and created the "walk in the forest" narrative in an attempt to make the task more appealing for participants, whilst still presenting two

sets of stimuli that we assumed to be clearly distinct. However, the possibility that stimuli were not from dimensions distinct enough remains. Furthermore, Delamater (2012) proposed in his own connectionist model a solution to this discrimination that involved specific sets of auditory and visual modality-specific hidden units combined with shared multimodal hidden units. For these reasons, the last experiment of this series attempted to replicate Delamater's findings by presenting stimuli from an auditory and visual modality.

3.6 Experiment 10

Results from the previous two experiments failed to show differences in performance between a group of participants who received a reversal with the same outcomes within dimension and a group who received different outcomes within each stimulus dimension. The last experiment in this series attempted a last replication of findings from Experiment 3 in Delamater (1998). Experiment 10 was a direct replication of Experiment 9. However, in line with the original study, it presented participants with stimuli from distinct auditory and visual modalities.

3.6.1 Method

3.6.1.1 Participants

Due to the COVID-19 outbreak in the spring of 2020, only 20 students of the 32 that were intended were recruited in this study (17 women and 3 men, M_{age} = 25.15, *SD* = 5.56, range: 18-38). Of those 20 participants, 10 were tested in our usual small room in the Psychology building at the University of Nottingham. These participants received module credits or a small allowance for their participation. Participants were recruited using posters and the School of Psychology online booking system. The remainder 10 participants were tested in a quiet room in their homes to keep the experimental setting as consistent as possible. All participants were informed about the task and debriefed upon completion. All agreed to participate.

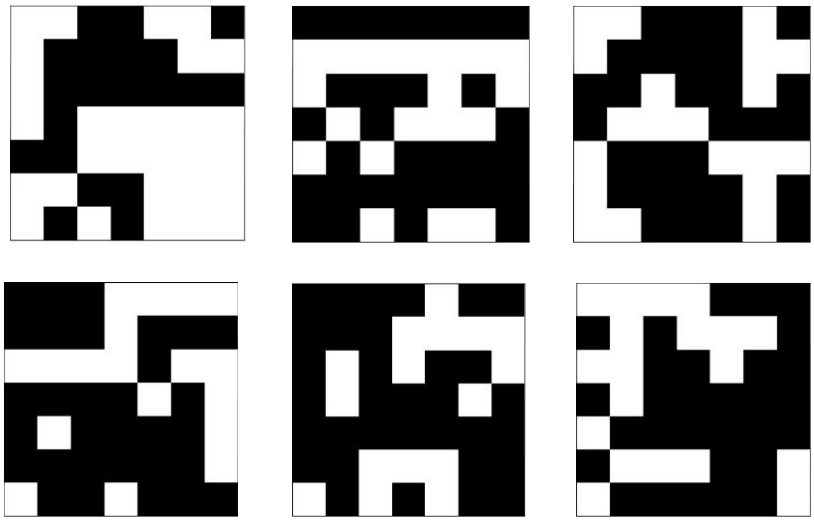
3.6.1.2 Apparatus & Materials

Visual stimuli were six black and white checkboards that contained 49 cells (7 x 7), as shown in **Figure 38**. The checkboard stimuli were made by generating a random distribution of real numbers ranging from equal or greater than zero to one on a 7 x 7 cell table. Cells containing numbers equal or lower than 0.5 were coloured in black and the remainder cells in white. Checkboard one contained 20 black and 29 white squares. Checkboard two contained 30 black and 19 white squares. Checkboard three contained 29 black and 20 white squares. Checkboard four contained 31 black and 18 white squares. Checkboard five contained 33 black and 16 white squares and checkboard six contained 30 black and 19 white squares. Overall, checkboards were made of approximately 58% black and 41% white squares arranged in a random fashion. Stimuli were 6 by 6 cm in size. Auditory stimuli consisted of six different melodies played by string, wind and percussion orchestra instruments. Specifically, a cello, a clarinet, a mandolin, a drum, a saxophone and a violin. After the presentation of each stimulus, a red dot approximately 1.5 cm in diameter appeared in the centre of the screen and moved upwards, downwards, to the left or to the right of the screen in the corresponding direction. After the participant's response, the feedback *"CORRECT!"* or *"wrong"* appeared in the centre of the screen for 1 s, followed by the red dot and the text *"The dot moves to the left!"* (for participants for whom a specific auditory or visual stimulus signalled a left response) for 2 s. 10 participants were tested using the computer described in all previous experiments. The 10 participants tested in their homes completed the task in a laptop with a 30 (width) x 21.24 (height) cm screen. All participants wore a pair of headphones (Panasonic RP-HT225) and completed the experiment in a quiet room in

an attempt to keep the experimental settings consistent. All unspecified details are identical to those of Experiment 9.

Figure 38

Visual Stimuli used in Experiment 10



Note. Only one stimulus was present in any given trial.

3.6.1.3 Procedure

Participants were randomly assigned without replacement to groups *Same* outcomes within dimension and *Different* outcomes within dimensions to ensure an equal number of participants per group (n = 10 each). All participants read an instruction sheet that emphasize participants' rights to terminate the task at any time. After ensuring participants had understood the task, the experimenter left the room and returned once the task was finished to debrief participants.

Prior to the start of the experiments, participants read a set of written instructions that read *"In this experiment, you will see different black and white checkboards and hear different musical instruments. Each checkboard and*

instrument will be followed by a red dot that will move to the left, right, up or down on the screen. It is your task to pay close attention to each checkboard and musical instrument and learn to predict in which direction the dot will move. You will have 5 seconds to guess the direction. Initially you will have to guess, but you will receive feedback on your responses so you can learn to respond correctly". Trials were identical to those described in Experiment 9 with the difference that the presentation of each auditory or visual stimulus was followed by a red dot that moved in one of four possible directions on the screen.

3.6.2 Results and Discussion

3.6.2.1 Training data

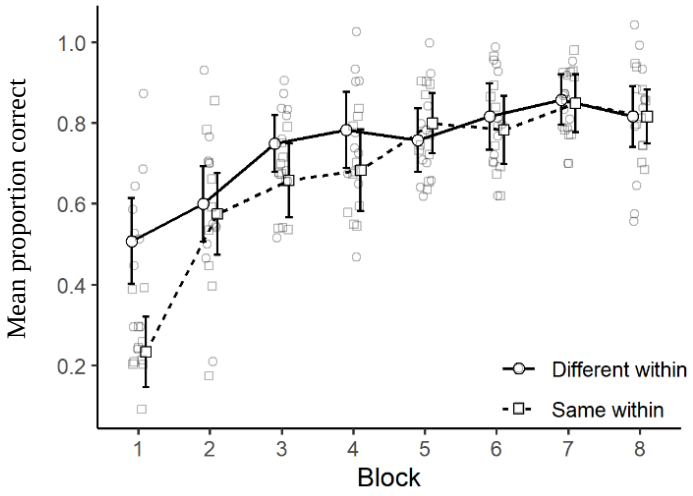
Experiment 10 used a new set of visual and auditory stimuli, in keeping with Delamater's (1998) original stimuli dimensions. To assess these stimuli were matched in difficulty and ensure that participants were learning about both modalities comparably, we halted data collection after eight participants (four from each group) and conducted a one-way ANOVA to assess a possible effect of stimulus modality on participants' performance. Just like in the previous experiment, the training for both groups was identical and the data were collapsed. The initial inspection of the data suggested an advantage for the auditory (M = .78) over the visual modality (M = .71). However, the main effect of stimulus dimension was not reliable, $F(1, 7) = 3.07$, $p = .123$, $\eta_\rho^2 = .33$, 90% CI [0.00, 0.58]. After ensuring that performance was matched, participant recruitment was resumed.

Data from the training stage are summarised in **Figure 39**. Just like in Experiment 9, the inspection of the data indicates that participants learned the initial discrimination more progressively, reaching levels of performance close to

asymptote later in training. The data shows a clear advantage for group Different in the first block of training. However, this spurious difference disappeared quickly, reflecting the identical training. A mixed ANOVA with the between-subjects factors of group (Same vs. Different) and the within-subjects factor of dimension (auditory vs. visual), and block was conducted on training data. Of most importance, the analysis yielded no main effect of group, $F(1, 18) = 1.85$, $p = .191$, $\eta_\rho^2 = .09$, 90% CI [0.00, 0.31], confirming no overall group differences at this stage. The interaction between group and block was reliable, $F(7, 126) = 4.41$, $p = .003$, $\eta_\rho^2 = .19$, 90% CI [0.06, 0.19], confirming the observed advantage for group Different over group Same only in block 1, $F(1, 18) = 13.52$, $p = .001$. The analysis also yielded a main effect of block, $F(7, 126) = 45.92$, $p < .001$, $\eta_\rho^2 = .72$, 90% CI [0.63, 0.75], and a main effect of dimension, $F(1, 18) = 12.76$, $p = .002$, $\eta_\rho^2 = .41$, 90% CI [0.11, 0.60], reflecting that, although we tried to ensure performance for each stimulus modality was well matched, there was an overall advantage for the auditory ($M = .76$) over the visual dimension ($M = .65$).

Figure 39

Collapsed Mean Performance for the Training Stage of the Acquired Equivalence Task.



Note. Each block comprised the presentation of stimuli V1, V2, V3, V4, V5, V6 and A1, A2, A3, A4, A5, A6 once per block. Error bars represent SEM. Semi-transparent grey circles and squares represent participants' average individual performance in each group.

3.6.2.2 Reversals

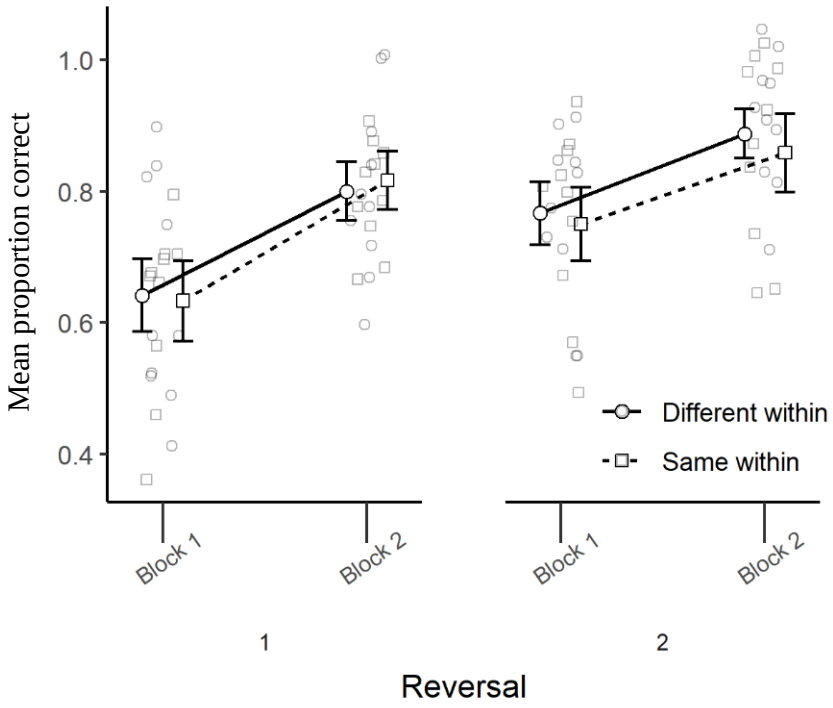
Data for the two reversal stages, split by block within each reversal, are summarised in **Figure 40**. Experiment 8 and Experiment 9 failed to show any group differences. However, the choice of stimuli in these two experiments differed considerably from stimuli in the original study. We reasoned that the use of auditory and visual stimuli made this task the closest to replicating Experiment 3 by Delamater (1998). Thus, we still expected group Different to perform better than group Same. A mixed ANOVA with the between-subjects factors of group (Same vs. Different) and the within-subjects factor of reversal (first and second reversal), dimension (auditory vs. visual) and block (block one and two) was conducted on reversal data.

Upon initial inspection, the reversal data looked similar to that of the previous experiment, suggesting a small numerical advantage for group Different during the second reversal. However, once again the analysis failed to show a reliable effect of group, $F(1, 18) = 0.04$, $p = .852$, $\eta_\rho^2 = .002$, 90% CI [0.00, 0.10] or any interaction involving group ($p = .533$ for the interaction between group and reversal).

The analysis replicated the main effect of reversal found in our previous experiments, $F(1, 18) = 18.93$, $p < .001$, $\eta_\rho^2 = .51$, 90% CI [0.21, 0.68], with performance in the second reversal reliably higher than performance in the first reversal ($p < .001$). The analysis also yielded a main effect of block, $F(1, 18) = 172.19$, $p < .001$, $\eta_\rho^2 = .90$, 90% CI [0.80, 0.93], with participants overall performance in the second block of the reversals ($M = .84$) reliably better than overall performance in the first block of the reversals ($M = .70$) ($p < .001$). The analysis showed a main effect of dimension, $F(1, 18) = 10.76$, $p < .001$, $\eta_\rho^2 = .37$, 90% CI [0.08, 0.56], in keeping with the increased performance for the auditory ($M = .81$) over the visual ($M = .73$) stimulus modality ($p = .003$).

Figure 40

Mean Proportion of Correct Responses during Reversal stages



Note. Each block comprised the presentation of stimuli A1-A6 and V1-V6 twice (48 trials per reversal). Error bars represent SEM. Semi-transparent grey circles and squares represent participants' average individual performance in each group.

3.7 General Discussion

Experiments in Chapter 3 sought to investigate the effects of outcome manipulations in different forms of acquired equivalence. The first set of experiments within Chapter 3 served as exploratory investigations into the effect of presenting the same or different outcomes across training and revaluation in a configural acquired equivalence task. Experiment 5 compared performance in a group of participants that completed our usual acquired equivalence task, with the *same* set of differential outcomes across training and revaluation, and a group that received a *different* set of differential outcomes across stages. The findings revealed an enhanced acquired equivalence in the group that received different outcomes. This finding is important because previous research into configural acquired equivalence has not systematically manipulated the reinforcers to assess the strength of the acquired equivalence effect. However, participants who completed our usual task failed to show acquired equivalence for the first time in this thesis. Experiment 7, which rectified Experiment 6, presented additional non-configural stimulus-outcome contingencies to both groups during training in an attempt to minimise potential non-specific effects that might have resulted from the differential group treatments in Experiment 5. The results revealed that, although both groups showed the expected acquired equivalence, the group differences disappeared.

These three experiments addressed a series of theoretically and experimentally motivated questions. One of the motivations in conducting this series of experiments was the DOE, which has demonstrated that subjects learn a discrimination more rapidly when trained with differential, over non-differential outcomes, in a variety of circumstances and organisms (see Urcuioli, 2005). All the experiments presented in the first half of the chapter used differential outcomes.

However, these studies have asked whether a change to different differential outcomes during a second stage would have an effect in performance compared to the use of the same differential outcomes. Leaving aside the results from Experiment 7 for now, it is unclear how traditional interpretations of the DOE would account for the differences in performance observed during revaluation and test in Experiment 5. For example, Trapold and Overmier (1972) proposed that Pavlovian S – O and instrumental O – R associations contribute to the discrimination learning in a differential outcome task. That is, through training, when a stimulus (S1) is presented, it is assumed to evoke the representation of an outcome (O1). When the correct response (R1) is reinforced in the presence of this O1 representation, then this specific O1 – R1 association will be learned. In these terms, the biconditional discrimination in our task could be represented as *Aw – O1 – R1, Ax – O2 – R2, Bw – O2 – R2, Bx – O1 – R1, Cw – O1 – R1, Cx – O2 – R2, Dw – O2 – R2,* and *Dx – O1 – R1*, where O1, O2, R1 and R2 represent the differential outcomes and corresponding responses, respectively. This being the case, it is not obvious how a DOE mechanism could anticipate our experimental data. Participants who received the same differential outcomes should have, if anything, benefited from the outcome expectancies established during the initial training: *A – O1 – R1 and B – O2 – R2.* On the other hand, participants who experienced different outcomes during revaluation would have had to update their now inaccurate outcome expectancies: *A – O3 – R1 and B – O4 – R2.*

However, the connectionist architecture described in previous chapters could be extended to explain findings from Experiment 5. By the end of training in a discrimination of the form Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw- and Dx+, we would expect the connections between each input and their corresponding hidden

and output units (*acw+*, *acx-*, *bdw-*, *bdx+*) to be at asymptote. For group Same, this asymptotic activation would result in the presentation of A+ eliciting great activation to its corresponding hidden and output unit (*acw+*) during revaluation. This strong activation of the hidden unit will drive Hebbian and anti-Hebbian changes in the strength of the connections across the network (Honey et al., 2010). On the one hand, it will ensure that the connection between input A and hidden unit *acw* and the reciprocal connections between the hidden and output unit *acw+* remain strong. However, this asymptotic activation of *acw+* driven solely by A in the absence of inputs C and w will cause a proportionally strong reduction in the weight of the connections between these inputs and hidden unit *acw+*, critical to the acquired equivalence effect. That is, with no explicit feedback, the link between input C and hidden unit *acw+* will quickly be extinguished. The same will apply upon presentation of B- and hidden unit *bdw-*. Group Different will start the revaluation with the exact same asymptotic connections between inputs, hidden and output units. The presence of A in an A* revaluation trial will tend to strongly activate hidden units *acw+* or *acx-*, but for the sake of argument I will assume that hidden unit *acw+* wins this competition. Because the outcome anticipated by hidden unit *acw+* does not occur in an A* trial, there will be no strengthening of the reciprocal connections between hidden unit *acw* and output unit +. A sub-asymptotic activation of *acw+* should, in turn, leave less scope for the reduction in the connections between w and, critically C, and their corresponding hidden unit. Instead, hidden unit *acw* should be expected to quickly develop new connections to the new output unit *. Of course, we should expect the connection between critical inputs C and D and their corresponding hidden unit to extinguish eventually, but if group Different

experiences a slower reduction in the weights of those connections, they should, at least for some time, have an advantage over group Same.

Experiment 7, which presented all outcomes during training by incorporating non-configural stimuli, eliminated group differences. One possibility for these findings is that presenting all outcomes from the onset of the task controlled for the differences in participants' level of arousal or interest in the task that may have translated in the observed group differences in performance, thus eliminating them. By presenting all outcomes we should have certainly reduced group differences in levels of arousal compared to the previous experiment. Instead of suddenly receiving novel outcomes during revaluation, participants in group Different received different permutations of the stimulus-outcome contingencies presented during training. However, although less dramatic, this difference should have still resulted in increased arousal or renewed interest in the task in group Different. At least more so than in group Same. Theoretically, the same connectionist analysis that applied to Experiment 5 could be extended to Experiment 7, which suggests we should have still anticipated group differences. However, this will be discussed in depth in Chapter 4, which includes simulated data from a formal mathematical instantiation of Honey's network in order to explore the effects of incorporating additional non-configural stimuli to the network.

The second set of experiments within Chapter 3 served as replication investigations into the effect of reversing differential outcomes between and within stimulus modalities in a non-configural acquired equivalence task, based on Experiment 3 in Delamater (1998). Overall, however, this series of experiments was unable to provide any evidence for group differences in reversal acquisition. Experiment 8 compared performance in a group of participants that received a series

of reversals with the *same* differential outcomes within stimulus dimension with a group that received reversals with *different* outcomes within stimulus dimension. Contrary to our expectations, the findings revealed no group differences in reversal acquisition, with both groups performing at levels close to asymptote across the task. Experiment 9 used a full counterbalancing and intended to increase the difficulty of the task, in an attempt to unmask possible group differences, by tripling the number of exemplars presented during the discrimination. However, results once again failed to show any group differences in reversal acquisition. Experiment 10 attempted a more direct replication of Delamater's (1998) by presenting stimuli from the original visual and auditory modalities. It should be noted this experiment was interrupted and has a smaller sample size compared to the other experiments in this series. However, no reliable group differences were found.

There are some evident differences, starting with the fact that we used an instrumental procedure, between our experiments and Delamater's that could have led to our failure to replicate the findings. For example, in Delamater's experiment, rats received trials with two differential appetitive reinforcers (food pellets and sucrose) and trials with no reinforcer. In our tasks, participants received four neutral differential trials outcomes (left, right, up, and down). However, the difference of most interest to these experiments is the choice of stimuli. In Experiments 8 and 9 our two stimulus modalities were snake and bear cartoons. We argued that each snake cartoon should have elements in common with the other snake cartoons, but few common elements with the bear stimuli. That is, they should still be considered as separate modalities. However, it is possible that these two sets of stimuli had more commonalities than we initially anticipated. For example, both sets of stimuli are animals, both are wild animals specifically, both have the potential to be dangerous

to people, etc. It remains possible that participants might not have treated these stimuli as belonging to different modalities, essential to the motivations behind these experiments. Of course, even if we assume this is true, Experiment 10 should have rectified this issue, with visual checkboards and melodies played by different instruments unequivocally belonging in two separate stimulus dimensions. However, although underpowered, data in Experiment 10 did not seem to suggest reliable group differences.

Leaving aside our failure to replicate the results, just like the first series of experiments within this chapter, the differences in reversal acquisition cannot simply be attributed to a DOE. Instead, these findings could naturally be captured by a connectionist network. Delamater (2012) proposed a three-layered connectionist network that could accommodate his findings. The network differs from a Honey-like network in that it assumes that modality-specific and multimodal units coexist at the hidden layer level, in a feature meant to add biological plausibility by recognising that the nervous system allows multimodal and unimodal processing pathways (Poremba et al., 2003). The network successfully simulated differences in reversal acquisition by allowing internal representations across hidden units to converge when stimuli from the same modality were reinforced with the same outcomes across training and reversals, and to diverge when different outcomes were presented across stages. However, the need to invoke different hidden unit modalities to accommodate Delamater's findings is not evident. A network with the characteristics of Honey's network should, in theory, be able to anticipate different performances between the groups in we make a number of reasonable assumptions. By the end of training, each individual stimulus within a modality should be expected to activate its hidden and output unit at an asymptotic level. This would

result in hidden units *a1x+*, *a2x-*, *a1y**, *a2y$*, where *x* and *y* represent the elements common to the visual and auditory dimension, respectively, and +, - , * , and $ represent four differential outcomes. Here, we need to assume that stimulus generalisation will be high between members of a stimulus modality, by dint of their similarity and common elements, but that no generalisation will occur across members of a different modality. This being true, it is clear that a reversing outcomes within a stimulus modality should be harder for the network than reversing them across stimulus modalities.

Although they might not provide unequivocal overall conclusions, taken together, the findings in this chapter are experimentally interesting for several reasons. On the one hand, they explore the effects of outcomes manipulations in different forms of acquired equivalence. In our configural acquired equivalence task, findings showed that changing outcomes across training and revaluation enhanced the acquired equivalence effect. These results are novel and could merit consideration when designing acquired equivalence tasks because, to the best of our knowledge, no single experiment had compared the effect of changing outcomes in the strength of acquired equivalence. Our failure to replicate Delamater's findings, albeit with some notable differences, is interesting in its own right, and might indicate that, at least in an instrumental task, participants are equally proficient at making reversals with the same or different outcomes across dimensions.

3.7.1 Conclusion

Experiment 5 found an enhanced acquired equivalence effect when different outcomes were used across training and revaluation, but not when additional non-configural cues were added to the initial discrimination training. The attempts to

replicate Delamater (1998) failed to obtain any group differences in reversal performance. However, taken together, experiments in this chapter add to the experimental literature on outcome manipulations. In any case, the experiments presented in this chapter offer a series of experimental data that, I argued, could be theoretically accommodated by a network with the characteristics of Honey's network. Chapter 4 sought to qualify, or refute, this connectionist approach by testing these data against a formal computational instantiation of the Honey's model (Robinson et al., 2019).

Chapter 4:

A Hebbian learning network: simulating empirical evidence of outcome manipulations

A number of disciplines (e.g., machine learning or neuroscience) use connectionist or neural networks as a way of modelling specific aspects of a given process. In psychology, a generic connectionist network could consist of a series of dedicated input and output units as a way of representing a standard view of associative learning, through connections forming between conditioned stimuli (e.g., a bell) and unconditioned stimuli (e.g., food), respectively. However, current theories go beyond the scope of this standard view of associative learning and maintain that a higher level of stimulus processing happens at a *hidden layer* that exists between the input (CS) and output layers (US) (e.g., Delamater, 2012; Honey et al., 2010; Honey & Ward-Robinson, 2002; Pearce, 1994). These theories have in common the notion that the features of CSs are represented and initially processed at an input layer. However, they assume further processing occurs at a hidden layer level, which is initially independent of any specific input. The hidden layer is assumed to reflect a deeper level of processing because it reflects changes in the connections between inputs and elements of the hidden layer, but also changes in the connections between outputs and elements of the hidden units, which become tuned to specific patterns of sensory inputs and outcomes.

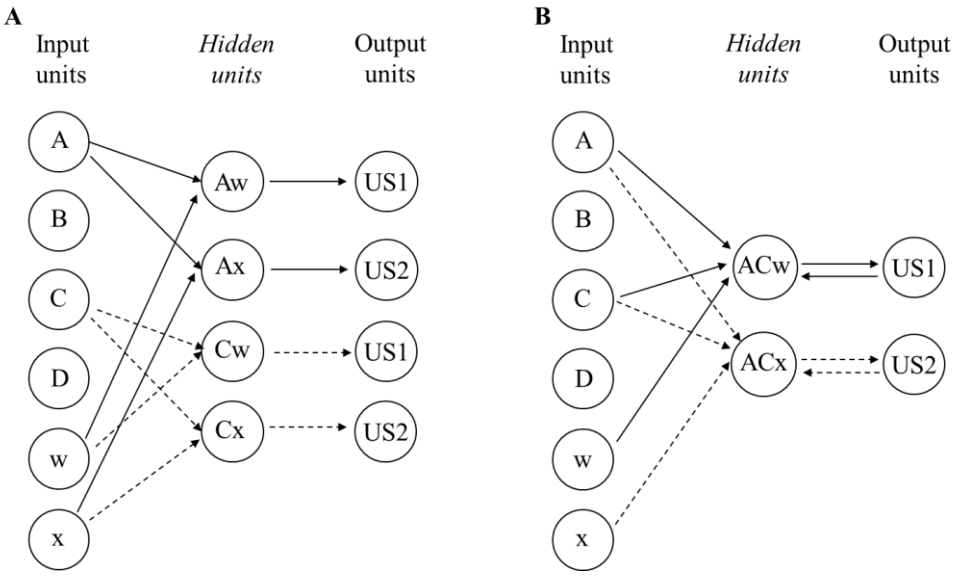
As I have been discussing throughout this thesis, a phenomenon that clearly illustrates the suitability of these neural network architectures is configural acquired equivalence. In brief, an animal receives a conditional discrimination in which only specific combinations of stimuli and contexts are reinforced (e.g., Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+). After the revaluation of A and B, a transfer of response is observed from stimulus A to C and from B to D with no explicit training, reflecting their common training history (e.g., Coutureau et al., 2002; Honey & Watt, 1998; Iordanova et al., 2007; Ward-Robinson & Honey, 2000). An analogous design

has been used consistently throughout Chapter 2 and Chapter 3 of this thesis. Because this design equates all binary combinations of stimuli and outcomes, it cannot be explained in terms of simple mediated conditioning (see General Introduction). Consider as an example the shock revaluation of A+ and B- after the initial biconditional appetitive training (e.g., Ward-Robinson & Honey, 2000). Without *w* or *x*, the presentation of A followed by a footshock might evoke the representation of Aw+ or Ax-. This would mean that the representations of food or no food are equally likely to enter into an association with the footshock. Additionally, because Aw and Ax are as similar to Cw and Cx as they are to Dw and Dx, there would be no grounds for generalisation from A+ to stimulus C any more than to D. Configural acquired equivalence is also not amenable to all connectionist accounts. For example, Pearce's (1987, 1994) connectionist model, illustrated in **Figure 41(a)**, posits that each specific input pattern (e.g., Aw) will develop connections to a single hidden unit, which will store the internal representation of that specific configural input pattern. Each hidden unit, representing each configural stimulus, will, in turn, develop links to the corresponding output (e.g., Food), thus solving the initial configural discrimination. However, under a symmetrical development of connections between inputs and hidden units, there is no reason to anticipate an increased transfer of responding from A to C and from B to D of the kind observed in experimental studies. Because hidden units only code for conjunctions of stimuli that have been presented together, the fact that A and C (or B and D) have never been combined with one another leaves this model ill-equipped to account for the observed acquired equivalence effect and others (e.g., congruent vs. incongruent acquisition - Honey & Ward-Robinson, 2001).

Honey's connectionist model (Honey et al., 2010; Honey & Ward-Robinson, 2002), exemplified in **Figure 41(b)**, circumvents this issue by allowing for similar input patterns to activate the same hidden unit when they are trained to predict the same outcome. The model does so by incorporating an output-to-hidden layer set of connections that has already been discussed in this thesis. In brief, any given input pattern (e.g., Aw) will come to activate a hidden unit and corresponding output (e.g., Food). However, when a similar input pattern is presented during training (e.g., Cw), the connections of inputs C and w to the hidden units will be influenced by the now active common output unit (e.g, Food), and the same hidden unit will likely be selected. These reciprocal connections between hidden and output units allow for inputs that have never been presented together to share hidden units, based on their shared features and common outcomes. These shared hidden units are critical to explain the generalisation between A and C (and B and D) after revaluation and to accommodate extant experimental data, like the faster acquisition of congruent vs. incongruent discriminations (e.g., Hodder et al., 2003; Honey & Ward-Robinson, 2002; Robinson & Owens, 2013) or whole vs. part reversal acquisition (e.g., Delamater & Joseph, 2000; Honey & Ward-Robinson, 2001; Robinson & Owens, 2013).

Figure 41

Pearce's Connectionist Network Architecture (A) vs. Honey's Connectionist Network Architecture (B)



Note. Configural trials Aw+, Ax-, Cw+ and Cx- are used to illustrate Pearce's (A) and Honey's (B) solution to a biconditional discrimination. In Pearce's model, each configural hidden unit is connected to a specific outcome (US1 = Food, US2 = No food). This allows for the solution of the initial discrimination, but fails to account for the observed post-revaluation acquired equivalence. In Honey's model, the reciprocal connections between hidden and output units allow for the sharing of hidden units between similar inputs that are trained to signal the same outcomes. This allows for the solution of the initial discrimination and also accommodates post-revaluation acquired equivalence.

Until recently, Honey's model had only been described informally. However, Robinson et al., (2019) reported a series of successful simulations of configural discrimination learning problems, albeit with some parameter dependency, using a formal computational instantiation of Honey's learning network. This chapter qualifies the computer instantiation of the model, by testing the implementation's behaviour against the experimental data reported in the previous chapters of this thesis. Note that a direct quantitative comparison between simulated and

experimental behavioural data would be misleading because it is not clear how simulated learning would translate into actual behaviour (e.g., how output unit activation would translate into participants' accuracy). However, it is possible to test the model by assessing the relationship between inputs and outcomes more qualitatively. For example, comparing the direction of the findings and ensuring that the most meaningful relationships between the stimuli and outcomes in the simulated data mirror experimental data (e.g., transfer from A to C and to B to D without explicit feedback). To that end, I first report the simulation of acquired equivalence tasks analogous to our usual acquired equivalence task, which presents revaluation and test trials intermixed in a single stage, instead of the usual 3-Stage procedure in which revaluation and test trials happen at different points in time. Then, a direct comparison between the simulation of an acquired equivalence task with the same and different outcomes across stages will show that the instantiation of the model does a good job capturing the group differences from Experiment 5. Simulations of Experiment 7 qualify our understanding of what happens to the network when configural and non-configural inputs are presented at once. Finally, I will assess the sensitivity of the model against non-configural discriminations of the kind reported in Delamater (1998), by reporting simulations analogous to Experiment 8, Experiment 9 and Experiment 10 in Chapter 3. Simulations were run on the Hebbian learning network modelled by George (2018)[1] and used in Robinson et al. (2019). I adapted the source code and modified the different scripts to meet the requirements of my simulations and make them analogous to the experimental tasks that I have reported in Chapters 2 and 3 of this thesis. Unless otherwise stated, all simulations

[1] The source code for the Hebbian learning network can be accessed and downloaded from the following GitHub repository: https://github.com/DavidNGeorge/HebbianNN.git

were run on 1000 networks with learning rate parameters of: .05, .25, and .25 for the input-to-hidden, hidden-to-output, and output-to-hidden projections, respectively. The networks received 50 epochs of training and 2 epochs of revaluation.

4.1 Simulating a 2-Stages Configural Acquired Equivalence Task

The ability for the instantiation of Honey's learning network to demonstrate configural acquired equivalence in a revaluation task has previously been established (Robinson et al., 2019). In their simulations, Robinson et al. (2019) demonstrated how the model was capable of solving an initial biconditional discrimination and to transfer responding to stimuli that had been initially trained as equivalent without explicit training. In their test of the network, Robinson and colleagues simulated the usual revaluation procedure consisting of three stages: training, revaluation, and test. In this thesis, I have used our adapted 2-Stage version of the procedure, summarised in **Table 10**, where training is followed by intermixed revaluation and test trials.

The task has done a good job of demonstrating the acquired equivalence effect, with Experiments 1-7 consistently capturing the phenomenon, with minimal exceptions. Although our 2-Stage revaluation acquired equivalence task is similar to the commonly used 3-Stage one, it would be unsafe to simply assume that intermixing revaluation and test trials would not have any consequences for the model, nor that it would not have any effect on acquired equivalence itself. Therefore, we reasoned that a failure for the instantiation of the Honey model to demonstrate acquired equivalence in a simulation with only two stages would undermine it, in its departure from the empirical findings presented in this thesis.

Table 10

2-Stages Configural Acquired Equivalence Experimental Design

Acquired equivalence	
Stage 1	Stage 2 Revaluation and test
Aw + Ax –	A +
Bw – Bx +	B –
Cw + Cx –	**C ?**
Dw – Dx +	**D ?**

Note. Although with different sets of stimuli, this experimental design was used in Experiment 1, Experiment 2, Experiment 3, Experiment 4 and Group Same in Experiment 5, Experiment 6 and Experiment 7. Letters represent different combinations of stimulus dimensions (e.g., an octopus with sleepy eyes and tentacles covered in suckers). +/- represent different trial outcomes (e.g., bite and sting). In all experiments stimuli C and D received no feedback.

4.1.1 Simulation description

Stimuli were modelled in the following way. In the Matlab script, each input unit coded for a stimulus (i.e., A, B, C, D, w, x) and each output unit coded for a specific outcome (US1 or US2). The training and intermixed revaluation and test trials were captured by two separate matrices of inputs and outputs. Inputs and outputs could be turned on (1) or off (0) in their respective matrices at any given stage. For example, in a simulated Aw+ training trial, all inputs would be off except for A and w and only one output (e.g.,US1) would be on, as shown in **Figure 42**. In keeping with our task, I will refer to the output units in the following simulations as *Bite* and *Sting*, respectively. That is, the output representing US1 being on (1) would be the equivalent of participants receiving, for example, outcome 'bite'. On that same trial, the output representing US2 (or 'sting') would be off (0). Note, however,

that these are simple verbal labels that do not provide the network with any intrinsic information about the outcome.

In this simulation, the network was initially trained on a biconditional discrimination of the form *Aw – Bite, Ax – Sting, Bw – Sting, Bx – Bite, Cw – Bite, Cx – Sting, Dw – Sting, Dx – Bite*. After training, the network was given two epochs of revaluation, where A and B continued to signal 'bite' and 'sting', respectively. Inputs C and D were also presented during the revaluation stage, but with the outcomes off (0) as a proxy for no feedback. The network was subsequently tested with the four inputs A-D, in the absence of w or x.

Figure 42

Example Script to Simulate our 2-Stage Revaluation Configural Acquired Equivalence Task

```
          %A B C D w x
tInput = [1 0 0 0 1 0;... %matrix of input patterns
          0 1 0 0 1 0;... %each row is a different pattern
          0 0 0 0 0 0;... %each column corresponds to an input unit
          0 0 0 0 0 0;...
          0 0 0 0 0 0;...
          0 0 0 0 0 0];

        %+ -
tOut = [1 0;... %matrix of output patterns
        0 1;... %each row is a different pattern
        0 0;... %each column corresponds to an output unit
        0 0;...
        0 0;...
        0 0];
```

Note. The example script shows simulated training trials Aw – US1 and Bw – US2. All simulations followed the same matrix structure. In the training Input matrix each column represents a distinct input. Different input pattern combinations (e.g., Aw) are represented on each row. In the training Output matrix each column represents an output (e.g., US1). Inputs and outputs can be switch off (0) or on (1) on any given simulation. Simulations consisted of training input/output, revaluation input/output and test input/output matrices.

4.1.2 Simulation results

The activation to each output unit (i.e., *Bite* or *Sting*) in the presence of inputs A-D prior and after revaluation is shown in **Table 11**. Prior to the revaluation stage, the activation to output units *Bite* and *Sting* was substantial (i.e., > .63) but undifferentiated when presenting stimuli A-D without w and x. This reflects the network's initial configural training. Without crucial inputs w and x, the activation to both output units was similar. A 4 x 2 ANOVA with the within-network factors of stimulus (A-D) and output (Bite vs. Sting) confirmed no main effect of stimulus, $F(3, 2997) = 0.17$, $p = .918$, $\eta_\rho^2 < .001$ or output, $F(1, 999) = 0.98$, $p = .323$, $\eta_\rho^2 < .001$ and no significant interaction between the two, $F(1.89, 1888.11) = 0.40$, $p = .660$, $\eta_\rho^2 < .001$, indicating no differences in output unit activation prior to the revaluation of inputs A/B.

Table 11

Mean Activation Levels to the Ouput Units when presenting Stimuli A-D Before and After Revaluation in a Simulation of a 2-Stage Configural Acquired Equivalence Task with the Same Outcomes across Stages

Input-to-Hidden	Hidden-to-Output	Output-to-Hidden	Test Stimulus	Output Unit: Before Revaluation Bite	Before Revaluation Sting	Post-Revaluation Test Bite	Post-Revaluation Test Sting
0.05	0.25	0.25	A	0.636	0.670	0.996	0.080
			B	0.644	0.658	0.075	0.996
			C	0.650	0.651	0.842	0.414
			D	0.643	0.655	0.410	0.838

Note. The networks were trained on our adapted 2-Stage acquired equivalence task, with revaluation and test trials intermixed. The networks were trained with the same outcomes across training and revaluation and test trials. Two pairs of stimuli were equivalent (AC and BD). Acquired equivalence is evidenced by the activation levels in the Post-Revaluation Test column. Note the transfer of output activation from stimulus A to C and from B to D after revaluation.

Looking at the Post-Revaluation Test data, it is evident that after the revaluation of *A – Bite and B – Sting*, stimulus A generated great activity in the *Bite* hidden unit and only marginal activity in the *Sting* output unit and that stimulus B generated the opposite pattern of results. As anticipated, A 2 x 2 ANOVA showed no main effect of stimulus (A vs. B), $F(1, 999) = 0.01$, $p = .921$, $\eta_\rho^2 < .001$ or output (Bite vs. Sting), $F(1, 999) = 0.24$, $p = .623$, $\eta_\rho^2 < .001$, but it confirmed the interaction between stimulus and outcome previously noted, $F(1, 999) = 37510.79$, $p < .001$, $\eta_\rho^2 = .974$.

Of most interest is how the network performed during the simulated test trials C and D, which had been paired with dummy outputs during the revaluation. From the Post-Revaluation Test data column, it is clear that the pattern of output activation transferred from the revaluation to the test trials despite the absence of explicit feedback. That is, stimulus C strongly activated the *Bite* output unit and stimulus D the *Sting* output unit. A pattern of activation qualitatively identical to participants' performance in our acquired equivalence task. An ANOVA confirmed this observation, once again revealing no main effect of stimulus (C vs. D), $F(1, 999) = 0.01$, $p = .921$, $\eta_\rho^2 < .001$ or output (Bite vs. Sting), $F(1, 999) = 0.24$, $p = .623$, $\eta_\rho^2 < .001$ during test trials, but a reliable interaction between these two, $F(1, 999) = 837.43$, $p < .001$, $\eta_\rho^2 = .456$. Note how, inputs C and D had never been presented with US1 or US2 during revaluation. Thus, the observed pattern of activation necessitates of the formation of shared hidden units (i.e., ACw – US1, BDx – US2) during the initial training to explain the resulting generalisation of output activity between inputs AC – US1 and BD – US2. The simulation of the Honey network was

able to successfully demonstrate acquired equivalence in our adapted 2-Stage procedure, accommodating the empirical findings discussed throughout this thesis.

4.2 Simulating Outcome Manipulations in a Configural Acquired Equivalence Task

Experiment 5 in Chapter 3, summarised in **Table 12**, demonstrated the acquired equivalence effect in a task that used the same outcomes across training and revaluation, but also different outcomes across stages. In their paper, Robinson et al. (2019) briefly noted how simulations of configural acquired equivalence using the *same outcomes* during training and revaluation yielded smaller differences in activation to output units than simulations using *different outcomes* during training and revaluation, in which could be taken as computational support for our experimental data. This observation led them to pose the question: *Would the use of different outcomes across stages lead to an enhanced acquired equivalence effect?* Our acquired equivalence task differs from the one simulated in Robinson et al. (2019) in that we present revaluation and test trials intermixed in a single stage. Additionally, no quantitative attempt to actually compare levels of output activation between the two simulations was made in Robinson et al.'s (2019) paper. The simulation in the previous section showed that the current model can accommodate experimental data from a 2-Stage task with the same outcomes across stages. However, the question of whether the model will simulate the task with different outcomes and whether it will anticipate these differences in activation remains to be answered.

Table 12

Experimental Design for Experiment 5

Group Same			Group Different		
Stage 1		Stage 2 Revaluation and Test	Stage 1		Stage 2 Revaluation and Test
a.			c.		
Aw $	Ax *	A $	Aw $	Ax *	A +
Bw *	Bx $	B *	Bw *	Bx $	B –
Cw $	Cx *	**C ?**	Cw $	Cx *	**C ?**
Dw *	Dx $	**D ?**	Dw *	Dx $	**D ?**
b.			d.		
Aw +	Ax –	A +	Aw +	Ax –	A $
Bw –	Bx +	B –	Bw –	Bx +	B *
Cw +	Cx –	**C ?**	Cw +	Cx –	**C ?**
Dw –	Dx +	**D ?**	Dw –	Dx +	**D ?**

Note. Letters represent different combinations of octopuses eyes (A-D) and different types of tentacles (w and x). +/- represent outcomes 'bite' and 'sting' and */$ represent 'poison' and 'suffocate', respectively. Stimuli C and D received no feedback when presented after training. Subgroups were counterbalanced to cancel out effects specific to any one outcome.

In the previous chapter, I noted how a direct comparison between configural acquired equivalence tasks that have used the same (e.g., Honey & Watt, 1999; Ward-Robinson & Honey, 2000) vs different outcomes across stages (e.g., Coutureau et al., 2002; Iordanova et al., 2007) would be inadequate to answer the question at hand. However, Experiment 5 circumvented the problems associated with attempting to compare prior research by counterbalancing the contingencies between stimuli and outcomes, and allowed for a direct comparison between performance in the two tasks, with results suggesting an advantage in performance for the group that experienced different outcomes across stages. When interpreting these results, I noted how an asymptotic connection between inputs A and B and their corresponding hidden units after training could lead to a faster extinction of the connections between C and D and their corresponding hidden units after revaluation

in the group with the same outcomes across stages, which would explain the diminished performance compared to the group with different outcomes. The simulations reported next are analogous to Experiment 5 and will allow for the assessment of the effects of outcome manipulations during revaluation and for a formal interpretation of any potential differences in performance. Thus, simulations reported next were intended to: (i) demonstrate that the implementation of the model also anticipates configural acquired equivalence in a 2-Stage procedure with *different* outcomes across training and revaluation, and (ii) asses the effect of presenting the same or different outcomes across training and revaluation in the strength of acquired equivalence. Specifically, test for potential group differences by directly comparing the level of output activation in a simulation with the same outcomes across training and revaluation with the level of output activation in a simulation with different outcomes across stages.

4.2.1 Simulation description

Just like in the previous simulation, each input unit coded for a stimulus (i.e., A, B, C, D, w, x) and each output unit coded for a particular outcome (US1, US2, US3 and US4) that I will label as *Bite, Sting, Poison* and *Suffocate*, respectively. Just like in the previous simulation, outputs could be on (1) or off (0) on any given trial and test trials C and D were followed by a dummy output (0), as a proxy for no feedback. The preceding section reported the simulation of our usual 2-Stage task with the same outcomes across training and revaluation. Therefore, this section will focus on the simulation of a 2-Stage procedure with *different* outcomes presented across training and revaluation.

In this simulation, the network was initially trained on a biconditional discrimination of the form *Aw – Bite, Ax – Sting, Bw – Sting, Bx – Bite, Cw – Bite, Cx – Sting, Dw – Sting, Dx – Bite*. However, because the Matlab script requires matrices for input and outputs to be the same size across all stages, all four outcomes had to be present in both the training and the revaluation matrix. This seems to depart from our experimental design in that in Experiment 5 participants were presented with only two possible outcomes at any given stage. However, although *Poison* and *Suffocate* were available during training, and *Bite* and *Sting* during revaluation and test, they were not active upon presentation of the inputs. Thus, activation to these output units was expected to be negligible.

After being trained on the initial biconditional discrimination, the network received two revaluation epochs with different outcomes: *A – Poison and B – Suffocate*. Note that referring to outputs as *Bite, Sting, Poison* or *Suffocate* simply helps us make meaningful comparisons between simulated data and our experimental

tasks in previous chapters. For the network, however, these outputs are simply *different output units* from the ones active during training.

4.2.2 Simulation results

The activation to each of the four output units in the presence of inputs A-D prior and after revaluation is shown in **Table 13**. Just like in the previous simulation, the activation to output units *Bite* and *Sting* in the presence of A-D prior to revaluation was substantial (> .63) but undifferentiated. A 4 x 2 ANOVA with the within-network factors of stimulus (A-D) and output (Bite vs. Sting) confirmed no main effect of stimulus, $F(3, 2997) = 2.46$, $p = .062$, $\eta_\rho^2 = .003$ or output, $F(1, 999) = 0.36$, $p > .548$, $\eta_\rho^2 < .001$ and no significant interaction between the two, $F(1.29, 1288.71) = 0.74$, $p = .526$, $\eta_\rho^2 < .001$. As anticipated, the activation to the *Poison* and *Suffocate* output units prior to revaluation was negligible, given the initial biconditional discrimination did not involve training these output units. The marginal activation (< .05) reflected the initial random activity of the network, which would have quickly adjusted as training proceeded.

Table 13

Mean Activation Levels to the Ouput Units when presenting Stimuli A-D Before and After Revaluation in a Simulation of a 2-Stage Configural Acquired Equivalence Task with Different Outcomes across Stages

Input-to-Hidden	Hidden-to-Output	Output-to-Hidden	Test Stimulus	Output Unit							
				Before Revaluation				Post-Revaluation Test			
				Bite	Sting	Poison	Suffocate	Bite	Sting	Poison	Suffocate
0.05	0.25	0.25	A	0.636	0.647	0.042	0.043	0.282	0.286	0.956	0.002
			B	0.633	0.641	0.044	0.043	0.294	0.277	0.002	0.952
			C	0.642	0.636	0.043	0.044	0.362	0.358	0.869	0.002
			D	0.661	0.636	0.043	0.042	0.376	0.354	0.003	0.861

Note. The networks were trained on our adapted 2-Stage acquired equivalence task, with revaluation and test trials intermixed. The networks were trained with different outcomes (i.e., US1 and US2 to US3 and US4) across training and revaluation and test trials. Two pairs of stimuli were equivalent (AC and BD). Acquired equivalence is evidenced by the activation levels in the Post-Revaluation Test column. The patterns of activation clearly transferred from A to C and from B to D.

Two epochs of revaluation of *A – Poison* and *B – Suffocate* were sufficient for input A to generate great activity in the *Poison* output unit and only marginal activation in the *Suffocate* hidden unit and for B to generate the opposite pattern of activation. A 2 (A vs. B) x 2 (Poison vs. Suffocate) ANOVA confirmed no main effects of stimulus or output, $F(1, 999) = 3.13$, $p = .077$, $\eta_\rho^2 = .003$ and $F(1, 999) = 3.19$, $p = .074$, $\eta_\rho^2 = .003$, respectively, but a reliable interaction between the two, $F(1, 999) = 87581.34$, $p < .001$, $\eta_\rho^2 = .999$. Critically, stimulus C and D, which had been revalued with dummy outputs, generated the equivalent pattern of activity. That is, stimulus C strongly activated the *Poison* output unit and stimulus D the *Suffocate* output unit. This observation was confirmed by a 2 (C vs. D) x 2 (Poison vs. Suffocate) ANOVA, which yielded no main effect of stimulus, $F(1, 999) = 1.93$, $p = .164$, $\eta_\rho^2 = .001$ or outcome, $F(1, 999) = 1.93$, $p = .165$, $\eta_\rho^2 = .001$, but a reliable interaction, $F(1, 999) = 17849.58$, $p < .001$, $\eta_\rho^2 > .957$, confirming the network's ability to successfully demonstrate acquired equivalence in our adapted 2-Stage procedure also with different outcomes across stages.

There was some residual activation (< .38) remaining from the training stage to the *Bite* and *Sting* output units. The activation was greater when stimulus C or D were presented (> .30) as compared to A or B (< .30). The greater reduction in the activation to output units *Bite* and *Sting* by A and B would have been caused by the presentation of the different output units *Posion* and *Suffocate* only in the presence of these trials, which would have facilitated the network's readjustment of weights towards the new outcomes. However, C and D were never presented in conjunction with the new outcomes, which would have resulted in these stimuli eliciting marginally more residual activity in the former outputs. A 4 (A-D) x 2 (Bite vs.

Sting) ANOVA confirmed this observation, yielding a main effect of stimulus, $F(1.56, 1558.44) = 80.52$, $p < .001$, $\eta_\rho^2 = .074$, but no main effect of outcome, $F(1, 999) = 0.12$, $p = .723$, $\eta_\rho^2 < .001$ or reliable interaction between the two factors, $F(1.56, 1558.44) = 0.39$, $p = .763$, $\eta_\rho^2 < .001$. Holm corrected pairwise comparisons confirmed that A and B generated significantly less activation in *Bite* and *Sting* than C and D post-revaluation ($p < .001$) but that neither A and B ($p > .350$) not C and D ($p > .350$) differed from each other in the level of activity they generated.

4.2.3 Same vs. Different outcomes across stages: comparing simulated absolute levels of activation.

After demonstrating the network's ability to simulate configural acquired equivalence in a 2-stage procedure using the same and different outcomes across stages, our focus turned to testing whether the network anticipated significant differences in performance between both simulations, in accordance with experimental data. To test whether the network can accommodate for these results, we report a between-networks comparison of the absolute level of activation to the correct output units in a simulation using the same outcomes across training and revaluation (group Same), vs. the absolute level of activation to the correct output units in a simulation using different outcomes across stages (group Different). From Experiment 5, it follows that this potential advantage might already be present during revaluation trials. Thus, we report the results of these between-networks comparisons for revaluation trials (A and B) and test trials (C and D) separately. The reported absolute level of activation to the correct output unit was calculated as the average activation to the correct output unit minus the average activation to the incorrect output unit for each input. Comparisons were run on the data from the simulations described above (sections 4.1.2 and 4.2.2).

4.2.4 Simulating group Same vs. group Different: Revaluation trials (A/B)

Unsurprisingly, the data for revaluation trials in both the Same and Different outcomes simulations showed that stimuli A and B generated great activity to the corresponding correct output unit post-revaluation. Although an initial look at the revaluation data in **Table 11** (Same; > .99) and **Table 13** (Different; > .95) could suggest a better performance in group Same, the analysis of the *absolute* levels of activation offered different conclusions. The averaged absolute levels of activation pointed at a small advantage for the simulation in group Different (M = .95) over the simulation in group Same (M = .92). A more granular look at the absolute activation of each of the 1000 individual networks per group revealed that, whilst no single network generated the incorrect pattern of output activation in the Different outcomes simulations in the presence of A or B, 33 networks generated more activation in the **wrong** output unit than in the correct one in the Same outcomes simulations. A 2 x 2 mixed ANOVA, with the within-networks factor of stimulus (A vs. B) and the between-networks factor of group (Same vs. Different), yielded a main effect of group, $F(1, 1998) = 38.43$, $p < .001$, $\eta_\rho^2 = .019$ but no main effect of stimulus $F(1, 1998) = 1.97$, $p = .160$, $\eta_\rho^2 < .001$ or interaction between the two, $F(1, 1998) = 1.97$, $p = .160$, $\eta_\rho^2 < .001$. These results accord with the revaluation data from Experiment 5, which revealed an enhanced performance in the Different group already present during revaluation trials.

4.2.5 Simulating group Same vs. group Different: Test trials (C/D). Evidence for an enhanced acquired equivalence effect.

The instantiation of the Honey model accommodated the advantage in performance for group Different observed in revaluation trials in Experiment 5. However, the real test to the network involved replicating the advantage found for group Different over group Same also during test trials. A quick look back at **Table 11**, shows that whilst stimulus C generated more activity in the *Bite* than the *Sting* output unit in the Same outcomes simulation, the incorrect output unit (i.e., *Sting*) still generated a lower, yet noticeable level of activation (> .40). Stimulus D produced a very closely matched opposite pattern of activation: strong activation to the *Sting* output unit and reduced, yet noticeable, activation to the incorrect output unit *Bite* (> .40). In the simulation where different outputs were used across stages, summarised in **Table 13**, the differences in activation to the correct and incorrect output units are more drastic. Stimulus C strongly activated the correct *Poison* output unit and it only generated a marginal activation in the incorrect output unit *Suffocate* (< .01). The equivalent opposite pattern of activation was generated by stimulus D, with strong activation to output unit *Suffocate* and only negligible activation to the incorrect output unit *Poison* (< .01). A more granular analysis of these data revealed that, in the lines of revaluation trials, individual networks behaved considerably different in each group. In the simulation of group Same, 541 individual networks generated the wrong pattern of activation. That is, more than half of the networks generated more activity in the incorrect than the correct output unit during test trials. In the simulation with Different outcomes, no single network generated the wrong pattern of activation. A Stimulus (C vs. D) x Group (Same vs. Different) ANOVA confirmed our initial observations. The analysis yielded a main effect of group, $F(1,$

1998) = 724.49, $p < .001$, $\eta_\rho^2 = .266$, but no main effect of stimulus, $F(1, 1998) = 0.95$, $p = .330$, $\eta_\rho^2 < .001$ or interaction between the two, $F(1, 1998) = 0.63$, $p = .428$, $\eta_\rho^2 < .001$. That is, the simulation with Different outcomes across stages produced a significantly greater absolute level of correct output activation ($M = .86$) compared to the simulation with the Same outcomes across stages ($M = .43$), confirming an enhanced acquired equivalence effect and replicating the results reported in Experiment 5. Results from the revaluation and test simulated trials demonstrate that the current Hebbian learning network is able to accommodate participants' actual performance, including group differences. They provide formal computational support to our discussion in the preceding chapter, and confirm that an asymptotic connection between A → *acw*+ and B → *bdx*- must have led to a faster extinction of the connections between C → *acw*+ and D → *bdx*- in the absence of explicit feedback in group Same compared to group Different.

4.3 Simulating Outcome Manipulations: Configural and Non-Configural Acquired Equivalence

Experiment 7 (Chapter 3), intended to replicate the group difference observed in Experiment 5 after incorporating non-configural trials to our usual configural acquired equivalence task, as summarised in **Table 14**. These filler non-configural trials ensured that participants were exposed to all outcomes from the onset of the task. Participants demonstrated acquired equivalence in the task with both the same and different outcomes across stages. This experimental design sought to account for effects non-specific to the change of outcome (e.g., arousal), and did so by minimising group differences in their exposure to the four possible outcomes. However, no group differences in performance were found. It is plausible that the inclusion of the four outcomes during training accounted for the failure to observe group differences. For example, if the enhanced performance in group Different were to be attributed to a simple increase in the interest in the task caused by the new outcomes during revaluation, presenting all outcomes from the start of the task might have been enough to level the interest in the task for participants in both groups. There was some evidence of the addition of non-configural trials having interceded with learning about the critical configural trials in Experiment 6, where performance during training was conspicuously low. However, Experiment 7 rectified this issue, and showed that participants learned the initial configural and non-configural discrimination successfully. It is still plausible for the inclusion of these non-configural stimuli to have somehow affected the formation of a Honey-like set of connections. Indeed, all simulations presented in Robinson et al. (2019) were configural in nature, and there is only a mention in passing to non-configural acquired equivalence. Thus, the simulations presented next were intended to qualify

the instantiation of the model further by testing whether it could demonstrate the acquired equivalence effect in a task with intermixed configural and non-configural trials, and assess to what extent simulated data matched experimental data. I present simulated data with *the same* outcomes and *different* outcomes across stages, analogous to the Same and Different groups in Experiment 7.

Table 14

Experimental Design for Experiment 7

Group Same		Group Different	
Stage 1	Stage 2 Revaluation and Test	Stage 1	Stage 2 Revaluation and Test
a.		c.	
Aw $ Ax *	A $	Aw $ Ax *	A +
Bw * Bx $	B *	Bw * Bx $	B –
Cw $ Cx *	**C ?**	Cw $ Cx *	**C ?**
Dw * Dx $	**D ?**	Dw * Dx $	**D ?**
S1 +	S1 +	S1 +	S1 $
S2 –	S2 –	S2 –	S2 *
b.		d.	
Aw + Ax –	A +	Aw + Ax –	A $
Bw – Bx +	B –	Bw – Bx +	B *
Cw + Cx –	**C ?**	Cw + Cx –	**C ?**
Dw – Dx +	**D ?**	Dw – Dx +	**D ?**
S1 $	S1 $	S1 $	S1 +
S2 *	S2 *	S2 *	S2 –

Note. Letters represent different combinations of octopuses eyes (A-D) and different types of tentacles (w and x). +/- represent outcomes 'bite' and 'sting' and */$ represent 'poison' and 'suffocate', respectively. Stimuli C and D received no feedback when presented after training. S1 and S2 represent two distinct squid cartoon drawings. Each was presented four times during training to ensure each outcome was presented an equal number of times.

4.3.1 Simulation description

New input units (S1 and S2) were added to the training and revaluation matrices in the script. When simulating data from the group Same, the contingencies between inputs and outputs remained the same across stages. In simulated data from group Different, all inputs activated the alternative output during the revaluation stage. For example, trial *Aw – Bite* became *A – Poison* during revaluation and *S1 – Poison* became *S1 – Bite*. All unspecified details are identical to those in the previous simulations.

4.3.2 Simulating configural and non-configural acquired equivalence - *Same* outcomes across training and revaluation.

The activation to each output unit in the presence of inputs A-D and S1-S2 prior and after revaluation is shown in **Table 15**. Prior to the revaluation stage, the activation to output units *Bite* and *Sting* was substantial (i.e., > .67) but undifferentiated when presenting stimuli A-D without w and x, reflecting the network's initial configural training. The non-configural nature of stimuli S1 and S2 was reflected in the great level of activation to *Poison* and *Suffocate* from these two inputs (> .99). An ANOVA with the within-network factors of stimulus (A, B, C, D, S1 and S2) and output (Bite, Sting, Poison and Suffocate) confirmed this observation. The ANOVA showed an interaction between stimulus and output, $F(2.85, 2847.15) = 3715.79$, $p < .001$, $\eta_\rho^2 = .788$. Both main effects were also reliable. The main effect of stimulus, $F(2.55, 2547.45) = 486.41$, $p < .001$, $\eta_\rho^2 = .327$ reflected, overall, greater activation to stimuli A-D than S1 or S2 ($p < .001$, Holm adjusted corrections). However, levels of activation between stimuli A-D and between S1-S2 were undifferentiated. The analysis also yielded a main effect of

output, $F(1.23, 1228.77) = 1816.25$, $p < .001$, $\eta_\rho^2 = .645$, reflecting that, overall, the outputs *Poison* and *Suffocate* received less activation than *Bite* and *Sting* ($p < .001$, Holm adjusted corrections). These results demonstrate that adding non-configural trials to the initial configural discrimination resulted in the anticipated learning and in the correct pattern of output level activation. These results mirror the empirical results obtained in Experiment 7, with participants in group Same showing good levels of performance in the configural trials and asymptotic performance in the non-configural trials before revaluation.

Table 15

Mean Activation Levels to the Ouput Units when presenting Stimuli A-D Before and After Revaluation in a Simulation with Configural and Non-Configural Stimuli (Same Outcomes)

Input-to-Hidden	Hidden-to-Output	Output-to-Hidden	Test Stimulus	Output Unit: Before Revaluation				Post-Revaluation Test			
				Bite	Sting	Poison	Suffocate	Bite	Sting	Poison	Suffocate
0.05	0.25	0.25	A	0.710	0.687	0.000	0.000	0.997	0.051	0.000	0.000
			B	0.679	0.699	0.000	0.000	0.055	0.997	0.000	0.000
			C	0.714	0.691	0.000	0.000	0.899	0.377	0.000	0.000
			D	0.678	0.698	0.000	0.000	0.393	0.873	0.000	0.000
			S1	0.000	0.000	0.999	0.055	0.000	0.000	0.999	0.054
			S2	0.000	0.000	0.055	0.999	0.000	0.000	0.054	0.999

Note. The networks were trained on our acquired equivalence task with additional non-configural trials S1 and S2. The networks were trained with the same outcomes across training and revaluation and test trials. Configural trials signalled the pair of outcomes (e.g., *bite* and *sting)* complimentary to the pair of outcomes signalled by non-configural cues (e.g., *poison* and *suffocate).* Two pairs of stimuli were equivalent (AC and BD). Acquired equivalence is evidenced by the activation levels in the Post-Revaluation Test columns.

After the revaluation of *A – Bite* and *B – Sting*, the same outcomes used to simulate training, it is evident that stimulus A generated great activity in the *Bite* hidden unit and only marginal activity in the *Sting* output unit and that stimulus B generated the opposite pattern of results. Unsurprisingly, S1 and S2, which

continued to unequivocally signal *Poison* and *Suffocate* during revaluation trials, generated great levels of activity in their respective output units, and only negligible activity in the *Bite* and *Sting* output units. Here, An ANOVA showed no main effect of stimulus, $F(1.65, 1648.35) = 1.72$, $p = .160$, $\eta_\rho^2 = .017$ or output, $F(1.95, 1948.05) = 1.39$, $p = .244$, $\eta_\rho^2 = .014$, but it confirmed the interaction between stimulus and output previously noted, $F(2.16, 2157.84) = 33621.08$, $p < .001$, $\eta_\rho^2 = .971$. The activity levels of the revaluation trials were once again in accordance with the empirical data presented in Experiment 7. Participants in group Same showed very good levels of discrimination for revaluation trials (A and B) and non-configural trials (S1 and S2).

Critically, test trials C and D, which had not received any explicit revaluation, generated a pattern of activation equivalent to that of stimuli A and B. That is, C strongly activated the *Bite* output unit and D the *Sting* output unit. An ANOVA confirmed this observation, revealing no main effect of stimulus (C vs. D), $F(1, 999) = 1.19$, $p = .276$, $\eta_\rho^2 < .001$ or output (Bite vs. Sting), $F(1, 999) = 0.99$, $p = .320$, $\eta_\rho^2 < .001$ during test trials, but a reliable interaction between these two, $F(1, 999) = 1243.03$, $p < .001$, $\eta_\rho^2 = .554$. That is, with the same outcomes across the training and revaluation stages, the simulation of our configural acquired equivalence task with the addition of non-configural stimuli produced the anticipated pattern of output unit activation, with generalisation occurring from stimulus A to C and B to D, mirroring experimental data from group Same in Experiment 7. This shows that the ability for the current model to simulate acquired equivalence can also be extended to experimental designs where configural and non-configural stimuli are intermixed, at least when the same outcomes are used across training and revaluation.

4.3.3 Simulating configural and non-configural acquired equivalence - *Different* outcomes across training and revaluation.

The activation to each of the four output units in the presence of inputs A-D and S1-S2 prior and after revaluation is shown in **Table 16**. Because the simulated training stage in this simulation and in the previous simulation are identical (i.e., groups received identical training but differential revaluation stages) the levels of output activation prior to revaluation are the same and are therefore omitted. Activation to output units after two epochs of revaluation with different outcomes: *A – Poison, B – Suffocate and S1 – Bite, S2 – Sting* is summarised in the post-revaluation columns of **Table 16.** Here, it is important to note that after revaluation, the residual activation to the output units used during training (i.e., *Bite* and *Sting* for A/B and *Poison* and *Suffocate* for S1/S2) was noticeably substantial (i.e., > .65).

Table 16

Mean Activation Levels to the Ouput Units when presenting Stimuli A-D Before and After Revaluation

Input-to-Hidden	Hidden-to-Output	Output-to-Hidden	Test Stimulus	Output Unit							
				Before Revaluation				Post-Revaluation Test			
				Bite	Sting	Poison	Suffocate	Bite	Sting	Poison	Suffocate
0.05	0.25	0.25	A	0.688	0.705	0.000	0.000	0.653	0.657	0.064	0.000
			B	0.694	0.707	0.000	0.000	0.658	0.669	0.000	0.060
			C	0.692	0.687	0.000	0.000	0.671	0.669	0.053	0.000
			D	0.694	0.711	0.000	0.000	0.676	0.697	0.000	0.043
			S1	0.000	0.000	0.999	0.053	0.037	0.000	0.985	0.035
			S2	0.000	0.000	0.052	0.999	0.000	0.038	0.038	0.982

Note. The networks were trained on our acquired equivalence task with additional non-configural trials S1 and S2. The networks were trained with different outcomes across training and revaluation and test trials. During training, configural trials signalled the pair of outcomes (e.g., *bite* and *sting)* complimentary to the pair of outcomes signalled by non-configural cues (e.g., *poison* and *suffocate).* These contingencies were reversed during revaluation. Two pairs of stimuli were equivalent (AC and BD). Acquired equivalence is evidenced by the activation levels to the *Poison* and *Suffocate* outputs in the Post-Revaluation Test columns. Note, however, the level of residual activation present in post-revaluation trials.

This differs from previous simulated data, where two epochs of revaluation were sufficient for the network to show a great level of activation in the new output units, and only marginal residual activation in the output units used during training. The addition of non-configural cues during training in this simulation resulted in two epochs of revaluation not being enough for the network to readjust its connection weights to the same extent. These results are qualitatively similar to the empirical data from group Different in Experiment 7, which showed a more progressive acquisition of the revaluation trials than that of group Same. For the purpose of our simulation, however, we focus on the activation generated in the new output units, leaving aside all other activation.

Although low in absolute terms, the presentation of input unit A generated the expected greater activity in the *Poison* as compared to the *Suffocate* output unit after revaluation. The opposite was true upon presentation of input unit B. Conversely, S1 and S2 generated more activity in the *Bite* and *Sting* output units, respectively. That is, although arguably very low, the simulation did generate the correct pattern of output activation in that all input units activated the correct, and not the incorrect, output units after revaluation. It is clear that inputs S1 and S2 showed a great level of residual activation to outputs *Poison* and *Suffocate,* which were reinforced during training, after revaluation. A direct quantitative comparison between these simulated data and the experimental data from Experiment 7 is not possible. However, these results can be taken as qualitatively similar to participants' performance in group Different, which showed that participants were also worse at learning the non-configural discrimination when compared to group Same. An ANOVA with the factors of stimulus (A, B, S1, S2) and output (Bite, Sting, Poison and Suffocate) confirmed the expected interaction between stimulus and output, $F(2, 7992) = 304.24$, $p < .001$, $\eta_\rho^2 = .071$.

Of most interest to our test of the network are test trials. Test trials C and D generated a pattern of activation equivalent to that of stimuli A and B. That is, C activated the *Poison* output unit and D the *Suffocate* output unit with no explicit feedback. The ANOVA confirmed the expected interaction between stimulus (C vs. D) and output (Poison vs. Suffocate), $F(1, 999) = 101.93$, $p < .001$, $\eta_\rho^2 = .093$ but no main effect of stimulus or output. These results show that the network was able to generate the correct pattern of output activation, albeit low in absolute terms, in a simulation with intermixed configural and non-configural trials even when different outcomes were used across stages. Results were qualitatively comparable to those of

Experiment 7, which showed that although participants in group Different demonstrated acquired equivalence, their learning was more progressive and their performance numerically lower compared to participants in group Same.

4.3.4 Same vs. Different outcomes across stages: comparing simulated absolute levels of activation on test trials (C/D) in a simulation with configural and non-configural inputs.

The instantiation of the Honey model was able to generate the correct pattern of output activation in simulations with intermixed configural and non-configural inputs and different output manipulations. These results extend the generality of the model, which had not been tested against this possibility. A quick look back at **Table 15** and **Table 16** however shows very different levels of correct output activation to test trials depending on whether the same (> .87) or different (> .04) outcomes were used across stages. It is worth remembering that we cannot directly equate levels of output activation with participants' accuracy performance. However, it is still instructive to look at test simulated performance in detail and to explicitly compare performance in both simulations to confirm statistical differences and identify qualitative similarities between simulated and experimental acquired equivalence data in a task with intermixed configural and non-configural trials.

In the simulation with the same outcomes across stages it is clear that inputs C and D generated more activity in their corresponding correct output units. Yet the incorrect output units showed lower but noticeable levels of activity (> .37). A look at the individual 1000 networks involved in this simulation revealed that 403 networks generated the wrong pattern of activation. That is, almost half of the networks generated more activity in the incorrect than the correct output unit. These

results are in line with those of the simulation describe in section 4.2.5, and show that adding non-configural inputs in a simulation with the same outcomes across stages had little effect on acquired equivalence. Although the overall levels of output activation were low in the simulation with different outcomes, the incorrect output units generated negligible levels of activation (< .00). The individual network data revealed that no single network generated the incorrect pattern of output activity, once again in line with the simulations described in section 4.2.5. These data evidence that, whilst adding non-configural inputs *and* changing outcomes across stages resulted in clearly lower overall patterns of activation, the network was still capable of showing acquired equivalence. These differences in absolute levels of correct output activation were confirmed by an ANOVA with the within-networks factor of Stimulus (C vs. D) and the between-networks factor of Group (Same vs. Different), which revealed a main effect of group, $F(1, 1998) = 974.12$, $p < .001$, $\eta_\rho^2 = .328$, but no main effect of stimulus, $F(1, 1998) = 0.99$, $p = .320$, $\eta_\rho^2 < .001$ or interaction between the two, $F(1, 1998) = 0.94$, $p = .332$, $\eta_\rho^2 < .001$. That is, the simulation with the Same outcomes across stages produced a significantly greater absolute level of correct output activation ($M = .47$) compared to the simulation with different outcomes across stages ($M = .04$). These results are qualitatively comparable to test results in Experiment 7, in which both the Same and Different groups showed acquired equivalence and group Same showed a numerical, albeit unreliable, advantage ($M = .46$) over group Different ($M = .40$).

The simulation of a configural acquired equivalence task with the addition of non-configural trials and different outcomes across stages was able to produce results in the anticipated direction, with generalisation occurring from stimulus A to C and B to D, respectively. However, it differed from the simulations previously

reported in the substantial levels of residual activation to the output units used during training after the revaluation took place. The low levels of activation to output units could be seen as undermining this simulation. However, this does not need to be the case. The simulation does not depart from empirical findings of acquired equivalence. Here, we need to focus on the qualitative relationships between the inputs and outputs that are most meaningful to our task, and generalisation did occur between the correct input and output units, just as anticipated. We simply cannot make a meaningful comparison between the output levels in the computer simulation and participants' actual performance. Furthermore, whilst the minimal two epochs of revaluation were sufficient to show good levels of post-revaluation output activation when using the same outcomes across stages, the low activation in the simulation with different outcomes could reflect an insufficient number of epochs of revaluation. Looking back at **Table 16**, it is evident that non-configural inputs S1 and S2 generated a great level of activity in the outputs active during training even after the revaluation took place. This suggests that the addition of the non-configural cues, coupled with the change of outcomes across stages, resulted in two epochs of revaluation not being enough for the network to readjust its connections and generate levels of activation comparable to those of the previous simulations. This will be discussed in more detail in the General Discussion of this chapter.

Although the current network has done a good job accommodating the experimental findings presented in this thesis, adding non-configural inputs resulted in a significant decrease in overall levels of output activation that had not been anticipated. This could be seen as a challenge to the model, which was built with configural discriminations in mind. Thus, an alternative way of testing the model would involve testing its ability to simulate non-configural acquired equivalence.

The next section focuses on simulating Delamater's (1998) non-configural test of acquired equivalence as a means to further qualify the present instantiation of the model.

4.4 Simulating Non-Configural Acquired Equivalence

Configural acquired equivalence is particularly relevant to illustrate the need of further processing at a hidden layer level, and demonstrates that mediated conditioning alone cannot account for the phenomenon. However, some non-configural experimental designs are also inexplicable in terms of mediated conditioning. For example, Delamater (1998) trained rats to discriminate between two auditory and two visual stimuli where one stimulus from each modality was paired with either a food pellet or sucrose (A1-, A2+, V1*, V2-, where A indicates auditory and V indicates visual stimuli). Once rats had mastered the discrimination, a group of rats received a reversal stage in which the stimuli signalled the same outcomes within each stimulus modality (A1+, A2-, V1-, V2*). A second group of rats received a reversal in which the stimuli signalled the outcomes previously associated with the opposite stimulus modality. That is, different outcomes within modality (A1*, A2-, V1-, V2+). Rats in the *different* outcomes within modality group acquired the new discrimination faster than rats in the *same* outcomes group. In this thesis, I have tried to replicate Delamater's findings and extend the generality of the effect in Experiments 8, 9 and 10, as exemplified in **Table 17**. Experiments differed in the number of exemplars (four in Experiment 8 and 12 in Experiment 9 and Experiment 10) and the stimulus modalities (visual in Experiment 8 and Experiment 9 and audio-visual in Experiment 10). However, all experiments, including the closest replication in terms of stimulus modality, failed to obtain any group differences in reversal acquisition.

Table 17

Experimental Design for Experiment 10

Training	Reversal	Reacquisition
a. Same outcomes within dimension		
A1 - **L** A2 - **R**	A1 - **R** A2 - **L**	A1 - **L** A2 - **R**
A3 - **R** A4 - **L**	A3 - **L** A4 - **R**	A3 - **R** A4 - **L**
A5 - **L** A6 - **R**	A5 - **R** A6 - **L**	A5 - **L** A6 - **R**
V1 - **U** V2 - **D**	V1 - **D** V2 - **U**	V1 - **U** V2 - **D**
V3 - **D** V4 - **U**	V3 - **U** V4 - **D**	V3 - **D** V4 - **U**
V5 - **U** V6 - **D**	V5 - **D** V6 - **U**	V5 - **U** V6 - **D**
b. Different outcomes within dimension		
A1 - **L** A2 - **R**	A1 - **U** A2 - **D**	A1 - **L** A2 - **R**
A3 - **R** A4 - **L**	A3 - **D** A4 - **U**	A3 - **R** A4 - **L**
A5 - **L** A6 - **R**	A5 - **U** A6 - **D**	A5 - **L** A6 - **R**
V1 - **U** V2 - **D**	V1 - **L** V2 - **R**	V1 - **U** V2 - **D**
V3 - **D** V4 - **U**	V3 - **R** V4 - **L**	V3 - **D** V4 - **U**
V5 - **U** V6 - **D**	V5 - **L** V6 - **R**	V5 - **U** V6 - **D**

Note. A1-A6 represent six distinct auditory stimuli (musical instruments) and V1-V6 six visual stimuli (black and white checkboards). L, R, U and D indicate a left, right, up or down keyboard response, respectively.

Delamater (2012) proposed a multimodal neural network to accommodate the reversal data. The network most significantly differs from the instantiation of Honey's model in that it assumes that there exist different modality pathways connecting hidden units with their respective outputs. In keeping with the nervous system's processing of sensory information, it assumes that some hidden units capture the physical features of each stimulus modality (auditory or visual) and a different set of hidden units allows for multimodal processing. The network captures the group differences reported in Delamater (1998) by allowing the internal representations across the hidden units to converge when the stimuli from the same

modality are reinforced with the same set of outcomes, and to become more distinct when the stimuli from the same modality are reinforced with different outcomes.

The instantiation of Honey's model has done a reasonable job of accounting for the experimental data presented in this thesis and other forms of acquired equivalence (Robinson et al., 2019) without resorting to different pathways for dedicated hidden units. In the previous chapter, I discussed reasons why we might have failed to replicate and extend the generality of Delamater's findings. Here, simulated data might offer computational support, or refute, our extant experimental data and help qualify whether different modality pathways are needed to accommodate the data. Thus, the simulations reported next intended to test the instantiation of Honey's model by simulating non-configural experimental data from Experiment 8, Experiment 9 and Experiment 10, analogous to Experiment 3 in Delamater (1998).

4.4.1 Simulation description

Stimuli were modelled in the following way. Each stimulus was assumed to activate a distinct input unit and a second input unit, common to all exemplars of that particular dimension. That is, inputs V1 to V6 coded for each visual stimulus. These six stimuli also activated input unit *x*, which represented the common visual features shared by all visual stimulus. Inputs A1 to A6 coded for each auditory stimulus. Similarly, all activated input unit *y*, which represented the common features shared by all auditory stimuli.

The network was initially trained on 50 epochs of *V1x – US1, V2x – US2, V3x – US2, V4x – US1, V5x – US1, V6x – US2 and A1y – US3, A2y – US4, A3y – US4, A4y – US3, A5y – US3* and *A6y – US4* pairings. In keeping with our task, US1, US2, US3 and US4 represent *left*, *right*, *up* and *down*, respectively. Just like in the previous simulations, these are simply verbal labels. At the end of training, the network was cloned and received two simultaneous 50 epoch reversals, which allowed for a within-network comparison of the effects of the different conditions of the reversals. In one reversal, analogous to our experimental *Same* group, the network received the opposite input-output pairings *within* each dimension (i.e., *V1x – US2, V2x – US1, V3x – US1, V4x – US2, V5x – US2, V6x – US1 and A1y – US4, A2y – US3, A3y – US3, A4y – US4, A5y – US4 and A6y – US3*). In the other reversal, analogous to our experimental *Different* group, the network received the opposite input-output pairings *across* dimensions (i.e., *V1x – US3, V2x – US4, V3x – US4, V4x – US3, V5x – US3, V6x – US4 and A1y – US1, A2y – US2, A3y – US2, A4y – US1, A5y – US1 and A6y – US2*).

Auditory and visual stimuli are assumed to have common features within each modality, but no common features shared across modality. However, it could be

argued that stimuli in a purely visual discrimination of the kind used in Experiment 8 and Experiment 9 will have additional common shared elements. That is, on top of a feature common to all bear stimuli (e.g., B1x) and a different feature common to all snake stimuli (e.g., S1y) these simulated data will be assumed to have one additional input to code for a feature common to all stimuli: all of them are animals (i.e., B1xa, S1ya). In keeping with previous simulations, we report simulations with learning rate parameters (ε) of: .05, .25, and .25 for the input-to-hidden, hidden-to-output, and output-to-hidden projections, respectively.

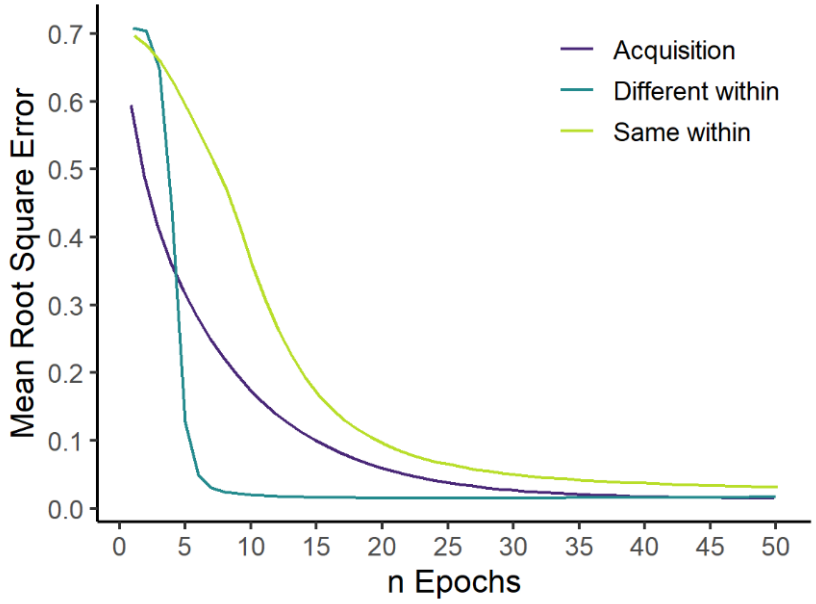
4.4.2 Simulating Delamater (1998): two visual dimensions

Simulated acquisition and reversal data for Experiment 8 is shown in **Figure 43.** The acquisition data shows a clear progressive reduction in the average mean root square error as the number of epochs progressed, illustrating the network's acquisition of the discrimination. The mean error rate for acquisition data was .102, 90% CI [.069, .133], with a standard deviation = .13. Of more interest is how the network performed during simulated reversals. Further inspection of **Figure 43** shows that the simulated data for group Different experienced a more abrupt decline in the mean root squared error compared to the simulation of group Same. That is, the simulated data for group Different shows a faster rate of learning compared to group Same. Overall, the average root square error for the simulation of the reversal with Different outcomes within dimension was .068 (*SD* = .17), 90% CI [.028, .108] and the average root square error for the simulation of the reversal with the Same outcomes within dimension was .174 (*SD* = .21), 90% CI [.125, .223]. That is, a difference of .104, 90% CI [.102, .105]. A paired *t*-test confirmed the advantage for the simulation with Different outcomes within modality, $t(999) = 144.22$, $p < .001$.

These results accord with Delamater's experimental data but depart from participants' data in Experiment 8 who did not show reliable differences in reversal acquisition.

Figure 43

Simulations of the Acquisition Data and a Reversal with the Same and Different Outcomes – analogous to Experiment 8



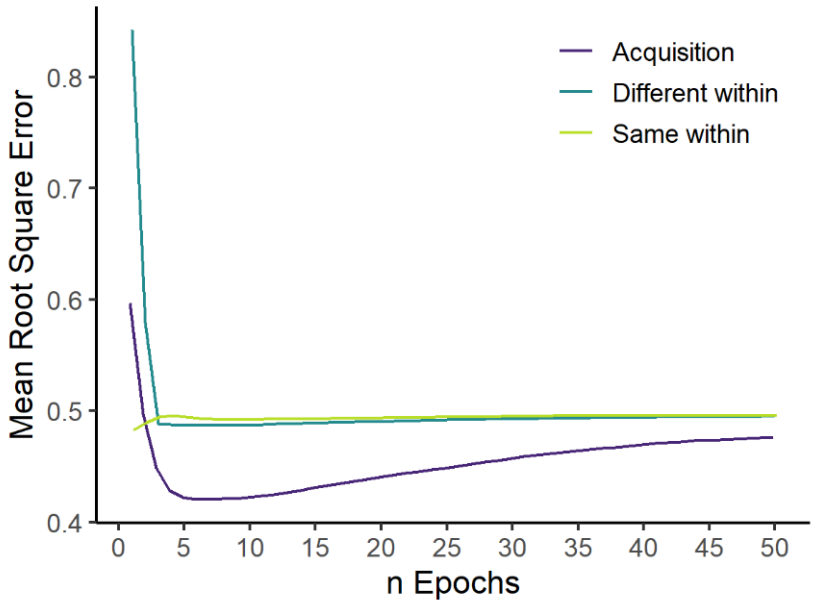
Note. Simulation of the acquisition data (identical for both sets of networks) and the reversal data with the same outcomes within input modality (e.g., L, R → R, L) and different outcomes within input modality (e.g., L, R → U, D) in a simulation with four inputs and a common element across all inputs. The error rate declines as the network learns. This decline is reliably faster in the simulation with different outcomes within stimulus modality.

The Simulation of Experiment 8 demonstrate that the network is able to accommodate group differences in reversal acquisition when the coding reflects two stimulus modalities with more than one feature in common. However, it remains a

possibility that additional inputs and outputs could affect the network's performance. Simulated data for Experiment 9, which increased the number of exemplars from four to 12, is summarised in **Figure 44.** Here, increasing the number of inputs resulted in a rather different pattern of activity. The initial acquisition data shows an abrupt early decline in mean root square error that *increases* progressively as training proceeds. Over the 50 epochs of initial training, the mean root error rate for acquisition data was .454 (*SD* = .03), 90% CI [.447, .460], notably higher than in a simulation with fewer inputs. Simulated reversal data, also summarised in **Figure 44**, clearly indicates a different behaviour from that of a simulation with fewer inputs. Most notably, the advantage for the simulation with Different outcomes within dimension previously observed was lost. The simulation for group Different started with a higher mean root square error that declined quickly as time went. However, the simulation for group Same started and maintained a lower mean root square error throughout the reversal stage. Overall, the average root square error for the simulation of the reversal for group Different was .500 (*SD* = .05), 90% CI [.488, .512] and the average root square error for the simulation for group Same was .494 (*SD* = .002), 90% CI [.493, .494]. That is, a difference of activation of .006, 90% CI [.002, .006]. Although the overall average root square errors for both simulations were higher than in the previous simulation, and the absolute difference marginal, a paired *t*-test confirmed an advantage for the simulation with the *Same* outcomes within dimension, $t(999) = 48.86$, $p < .001$. These data qualify the previous simulation and suggest that the network's ability to accommodate Delamater's (1998) data is dependent upon the number of common elements shared by the inputs, the number of input and output units, or both.

Figure 44

Simulations of the Acquisition Data and a Reversal with the Same Outcomes Within Dimensions and Different Outcomes Within Dimensions with an Increased Number of Inputs – analogous to Experiment 9



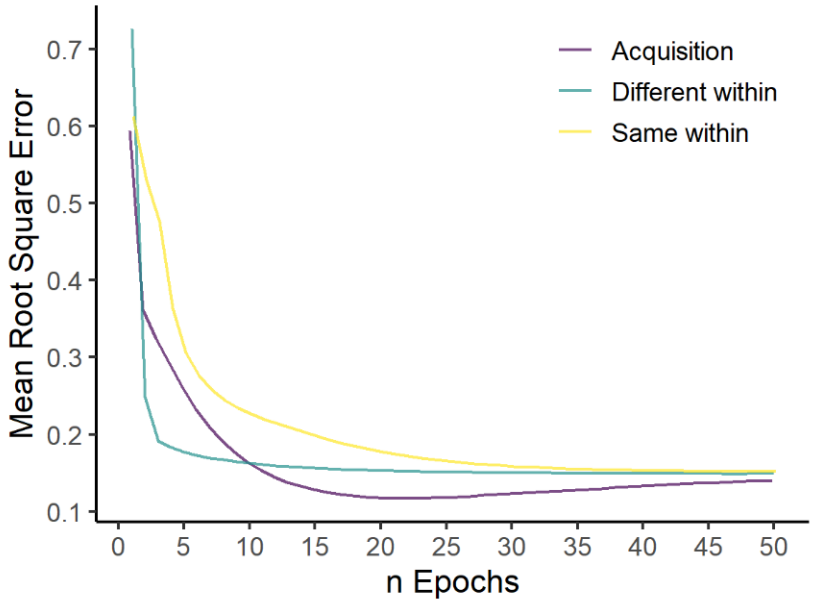
Note. Simulation of the acquisition data (identical for both sets of networks) and the reversal data with the same outcomes within input modality (e.g., L, R → R, L) and different outcomes within input modality (e.g., L, R → U, D) in a simulation with 12 inputs and a common element across all inputs. The error rate declines as the network learns. The decline is very sharp in the group with different outcomes within modality. However, the error rate is continuously low in the simulation with the same outcomes within input modality.

4.4.3 Simulating Delamater (1998) Experiment 3: Audio-visual discrimination

The simulation of Experiment 10 proceeded exactly like Experiment 9, with the exception that we assumed common elements shared between inputs coding for stimuli of a given modality, but no common shared elements between inputs from different modalities. Thus, it could help us understand why the simulation of Experiment 9 failed to show the anticipated pattern of activation. The initial acquisition data, summarised over the 50 epochs of revaluation in **Figure 45**, clearly shows a progressive decreased in the average root square error as time progressed. That is, the network learned the initial discrimination. Over the 50 epochs of initial training, the mean error rate for acquisition data was .159, 90% CI [.139, .178], with a standard deviation = .08. Of more interest is how the network performed during simulated reversals. Further inspection of **Figure 45** suggests that the simulation with Different outcomes within dimension started with a higher mean root squared error as compared to the simulation with the Same outcomes within dimension. However, whilst the simulation of the reversal for group Different had a disadvantaged start, the mean root square error dropped to levels close to zero clearly faster than the simulation of the reversal for group Same. Overall, the average root square error for the simulation of group Different was .168 (*SD* = .08), 90% CI [.148, .187] and the average root square error for the simulation of group Same was .204 (*SD* = *.10*), 90% CI [.181, .227]. That is, a difference of .036, 90% CI [.034, .037]. A paired *t*-test confirmed the advantage for the simulation with Different outcomes within dimension, $t(999) = 41.95$, $p < .001$, in accordance with Delamater's (1998) findings.

Figure 45

Simulations of the Acquisition Data and a Reversal with the Same Outcomes Within Dimensions and Different Outcomes Within Dimensions – analogous to Experiment 10



Note. Simulation of the acquisition data (identical for both sets of networks) and the reversal data with the same outcomes within input modality (e.g., L, R → R, L) and different outcomes within input modality (e.g., L, R → U, D) in a simulation with 12 inputs but no common element across all inputs. The error rate declines as the network learns. Once again the decline in error rate occurs faster in the simulation with different outcomes within input modality.

4.5 General Discussion

The purpose of Chapter 4 was to test the instantiation of Honey's model (Robinson et al., 2019) against some of the experimental data presented throughout this thesis. Specifically, we simulated data analogous to our 2-Stages configural acquired equivalence task, assessed the network performance against a revaluation stage with either the same or different outcomes, and simulated various non-configural Delamater-like discriminations. The simulation of our usual configural acquired equivalence task, with intermixed revaluation and test trials, demonstrated the model's ability to accommodate extant experimental data by generating the correct pattern of output unit activation without explicit training. The model generated activity in the correct output units in simulations with the same outcomes and different outcomes across training and revaluation. A direct comparison of the absolute levels of activation to the correct output units, analogous to Experiment 5, confirmed that simulations with different outcomes across stages produced an enhanced acquired equivalence effect, as demonstrated by reliably stronger levels of absolute activation to the correct output units. These simulated data offered computational support to our experimental data, which showed that participants performed reliably better when they experienced different outcomes across stages.

In some ways, this advantage for the simulation with different outcomes across stages might seem paradoxical: we might suppose that presenting the same outcomes during training and revaluation should facilitate performance. However, it is worth noting that the activation relies upon the network's ability to adjust the critical connections between input and hidden units and between hidden and output units. The key to understanding why this results in an enhanced acquired equivalence in group Different, as reflected by the greater level of absolute activation to the

correct output units, relies on the conditional principal component analysis (CPCA) feature of the network. The CPCA is a learning algorithm that calculates the conditional probability that a sending unit is active, given that the receiving unit is also active, and drives the changes in the connections' strength across the network in the appropriate direction. In this model, the CPCA determines that when a receiving unit (e.g., a hidden unit) is inactive, no changes in the weight strengths will occur. When a receiving unit is active, the strength in the connections will move in the direction of the sending unit. That is, if the receiving unit is active and the sending unit (e.g., an input unit) is active, the strength in the connection between these two will increase. If the receiving unit is active and the sending unit is inactive, the strength of their connection will decrease. Following the simulation of Experiment 5, we should expect the connections between inputs and their corresponding hidden units (*acw+*, *acx-*, *bdw-*, *bdx+*) to be at asymptote by the end of training. When A is revalued in group Same, the occurrence of output + will ensure that both the reciprocal connection between the output and hidden unit *acw+* and the connection between input A and hidden unit *acw+* remain strong. However, the strong activation of hidden unit *acw+* will have consequences for the rest of the connections in the network, critical to acquired equivalence. The asymptotic activation of *acw+* when sending unit C is inactive will result in a proportionally strong reduction in the strength of the connection between these two, by dint of the CPCA. The reduction in the strength of this connection will not be as drastic in group Different. When A is revalued but outcome *, instead of +, occurs in that trial, the reciprocal connections between *acw+* and outcome + will decrease, because sending unit + will be inactive in the presence of the active receiving unit *acw+*.

Because of the reduced activation to *acw+*, we would expect the link between input unit C and hidden unit *acw+* to take longer to extinguish.

Simulated data accorded with experimental data from Experiment 5, which showed that participants who experienced different outcomes across training and revaluation performed reliably better than participants who experienced the same outcomes across stages. This is important because our systematic counterbalancing within the task allows for direct comparison between the groups, unlike previous experiments that have used either the same or different reinforcers across stages (e.g., Iordanova et al., 2007; Ward-Robinson & Honey, 2000).

Simulated data of a configural and non-configural acquired equivalence task, analogous to Experiment 7, showed that the model was capable of generating the correct pattern of output activation with the same and different outcomes across stages. However, with the addition of non-configural inputs, the simulation with different outcomes across stages generated notably lower levels of activation to the correct output units, and increased levels of residual activation to the incorrect outputs. This could be taken as a challenge to the instantiation of Honey's model, especially after arguing how having different outcomes across stages enhanced acquired equivalence in the previous simulation. However, this does not have to be the case if we consider that configural and non-configural inputs will be affected by the weightings differently. With the same outputs across stages, we should expect the network in a simulation of Experiment 7 to behave just like that of Experiment 5. In addition to the formation of hidden units *acw+*, *acx-*, *bdw-* and *bdx+*, inputs S1 and S2 will become linked to hidden units *s1* and s2$* through training. Upon revaluation of A+, B-, S1* and S2$, all connections will be maintained at similar asymptotic levels of activation. Just like in Experiment 5, the asymptotic activation

of hidden units should result in a proportionally strong extinction of the connections between critical inputs C/D and their corresponding hidden units. This extinction will not apply to S1 and S2, because these inputs will continue to activate hidden units *s1** and *s2$*, respectively. By the end of training, the weight matrices will be identical for group Different. When A*, instead of A+, occurs on a revaluation trial, A will partially activate hidden units *acw+* and *acw-*. However, the actual trial outcome will feed back to hidden unit *s1**, which was exclusively and unequivocally activated by non-configural input S1 during training. In that struggle to readjust the weight connections upon partial activation to *acw+* and *acx-*, and asymptotic activation to *s1**, the winner takes all hidden unit selection mechanism is likely to "stick" with the wrong hidden unit, at least for some time. This being the case, the network could need extra revaluation to allow for all the weight changes necessary to solve this discrimination to occur.

Indeed, it is worth noting that this disparity in overall level of activation to correct output units between the simulations was reduced, and eventually disappeared, as the number of epochs of revaluation increased. For example, a simulation of configural and non-configural inputs with different outcomes across stages *and* six epochs of revaluation, instead of two, generated levels of activity comparable to those of the simulation with the same outcomes. These simulated data are qualitatively similar to our experimental data. Participants in group Different showed a more progressive performance during revaluation trials compared to participants in group Same. They also showed marginally worse acquired equivalence, although these differences in performance were not reliable and not as drastic as the differences in absolute levels of activation in the simulated data would suggest. In any case, these simulated data helped test the model, insofar as

suggesting that adding non-configural inputs results in an overall low level of output activation.

Simulated data of Experiments 8, 9 and 10 tested the model specifically on non-configural discriminations of the kind reported in Delamater (1998). These are important because they anticipate an enhanced reversal acquisition when *different* outcomes are used across stimulus modalities not explicable in mediated conditioning terms, and should be captured by a model of discrimination learning. It is worth reminding that our behavioural data failed to obtain any group differences in reversal acquisition. However, Chapter 3 discussed shortcomings that might have caused the observed results in detail. The model captured the anticipated group differences in reversal acquisition when simulating data from Experiment 8 (four exemplars from two visual dimensions) and Experiment 10 (12 exemplars from distinct visual and auditory modalities). Simulated data from Experiment 9 (12 exemplars from two visual dimensions) departed from the anticipated results and showed a marginal group difference in output activation (.006) in favour of the simulation with the same outcomes. This failure to obtain the expected group differences in the simulation of Experiment 9 suggests that the number of shared common elements becomes critical as the number of non-configural input increases.

These results demonstrate that the model is capable of capturing group differences in non-configural discriminations, empirical facts that should be captured by any model. The instantiation was sensitive to these discriminations without a need to resort to modality-specific hidden units, even when more shared elements were added across all inputs, which makes it a more parsimonious account than Delamater's (2012). However, the simulated data suggest that the number of common shared elements becomes critical as the number of inputs signalling the

same output increases, with the only difference between the simulated data from Experiment 9 and Experiment 10 being a common element shared across inputs from both modalities.

4.5.1 Conclusion

The simulated data in Chapter 4 tested the sensitivity of a formal instantiation of Honey's model to various forms of configural and non-configural discrimination learning, not explicable in terms of mediated conditioning. Specifically, it captured acquired equivalence in our adapted 2-stages task perfectly, supported an enhanced acquired equivalence when different outcomes are used across training and revaluation, and offered valuable insights about the consequences of intermixing configural and non-configural trials in a discrimination. The simulations presented also helped qualify the model further. They showed that, when non-configural inputs are added, some adjustments might be needed to obtain equivalent levels of output activation. They also showed that adjustments might be needed as the number of shared common elements in non-configural discriminations increases.

Chapter 5:

Overall discussion

The aim of the current thesis was to: (i) investigate the correlations between performance in configural acquired equivalence and attentional set tasks and to interpret these findings from the perspective of a connectionist network of the characteristics described by Honey and colleagues (Honey, 2000; Honey et al., 2010; Honey & Ward-Robinson, 2002). (ii) To investigate outcome manipulations in configural and non-configural acquired equivalence tasks by comparing performance between participants who received the same vs participants who received different outcomes across the training and revaluation stages of our acquired equivalence task and (iii) to test the ability of a formal instantiation of Honey's network (Robinson et al., 2019) to account for the experimental findings presented in this thesis.

Chapter 2 explicitly measured the relationship between performance at test in a configural acquired equivalence task and two forms of attentional set tasks: IDS/EDS and optional-shift. Experiments in this chapter demonstrated the acquired equivalence effect, as shown by an increased generalisation between stimulus A and C – and B and D – after an initial configural training designed to render them equivalent (Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Cx+). Experiments in this chapter also demonstrated a clear bias towards the predictive over the nonpredictive dimensions of compound stimuli. In the IDS/EDS this bias was evidenced by an increased number of errors-to-criterion in EDS compared to IDS trials. In the optional-shift task, participants responded based on the stimulus dimension that had previously been established as relevant to the solution of the discrimination, even when the elements of the compound were objectively equally relevant. These results are consistent with previous findings and reflect the role of prior training history in performance. Given the assumption that Honey et al.'s network could be applied to both forms of learning, we anticipated a positive correlation between test

performance in both tasks. However, performance was only found to correlate positively when attentional set was measured using an optional-shift task (overall BF = 25.33), which was matched with the acquired equivalence task on number of stages (training and intermixed revaluation and test trials), number of revaluation and test trials, stimuli and way of administering the task. This positive correlation in performance between both tasks is important because, to the best of our knowledge, no previous research has compared performance in both tasks using a single, directly comparable, within-subjects experiment. For example, Robinson and Owens (2013) reported a selective impairment in configural acquired equivalence performance, but not in the initial conditional learning, in a group of elderly participants. Similarly, Owen et al. (1991) found that elderly adults were selectively impaired when performing EDS, but not IDS. Whilst these two experiments could be taken as preliminary evidence for a positive correlation between performance in both tasks, they were run over 20 years apart on different participants and, presumably, under different laboratory conditions and, as such, they do not allow for a direct comparison of the effects.

Results in Chapter 2 revealed some important implications. It was found that performance in configural acquired equivalence and IDS/EDS, as measured by CANTAB IDS/EDS task, did not correlate. This is an interesting finding because it could imply that the IDS/EDS and optional-shift tasks are not measuring the same construct, or at least not to the same level of specificity. Despite having made a number of important contributions to our understanding of learning and attention and being regularly used in clinical settings (e.g., Bünger et al., 2019; Lawrence et al., 1996; Shamay-Tsoory et al., 2007), the CANTAB IDS/EDS task has some limitations. For example, it has been considered to be a rather non-specific marker of

executive function and problem solving ability, with the EDS resting on multiple cognitive components, rather than a specific test of set-shifting ability (Hampshire & Owen, 2010). First, since the IDS stage always occurs prior to the EDS, and only one dimension (e.g., colour) is relevant during IDS, the participants will have to identify the second dimension (e.g. shape) as an actual alternative, which they might have not considered at all until confronted with the EDS. Even if participants fail to identify that switching stimulus dimension is the correct strategy during EDS, they could still be partially correct. That is, because stimuli presented on any given trial belong to two separate dimensions – coloured shapes and lines – participants' selection of the incorrect stimulus dimension could incorrectly lead to positive feedback during EDS. Imagine a participant makes a selection according to the rule that shapes are relevant, even if in fact the selection should be now based on the line dimension. The participant might still guess a trial correctly without initially being aware of the required shift by sticking with the shapes but selecting the rectangle that happens to contain the now relevant line. Additionally, without full counterbalancing and with IDS always occurring prior to EDS, the task could show effects of stimulus generalisation – rather than specific attentional-shift – and reflect time-related artefacts. Some of these issues were addressed in our pilot intra/extra dimensional set-shifting task. Specifically, our task was carefully counterbalanced to rule out stimulus generalisation and presented IDS and EDS trials on average at the same point in time. Unfortunately, we did not obtain the anticipated IDS superiority. This failure to observe a positive correlation between performance in configural acquired equivalence and IDS/EDS could imply that Honey et al. (2010) are wrong in their attempt to apply their connectionist learning network to both forms of learning and that it might be necessary to appeal to a specific process of attentional modulation to

explain IDS/EDS. However, findings in Chapter 2 could not unambiguously confirm or refute Honey et al.'s (2010) claims.

Chapter 2 also measured eye-gaze during our optional-shift task to investigate the role of predictiveness in overt attention, since eye movements and attentional shifts are accepted to be tightly coupled (Deubel & Schneider, 1996). Eye tracking data in Experiment 3 revealed that participants directed their eye-gaze towards the stimulus dimension that was predictive of the outcome. More interestingly, participants continued to direct their gaze towards the stimulus dimension that had initially been established as relevant to solve the discrimination even in a subsequent stage, where both stimulus dimensions were objectively equally predictive (Aw+, Ax+, Bw-, Bx- → Cy+, Dz-, Cz ?, Dy ?). This finding adds to the body of experimental evidence that has documented the strong relationship between a cue's predictiveness and eye-gaze (Aristizabal et al., 2016; Hogarth et al., 2010; Le Pelley et al., 2011). However, unlike a number of previous studies that have measured eye-gaze in experimental procedures where stimuli remain identical but outcomes change across stages (e.g., learned predictiveness: Av-O1, Aw-O1, Cx-O2, Cy-O2 → Ax-O3, Cv-O4), we measured eye-gaze in a task where the stimuli – but not the outcomes – changed across stages. Additionally, Experiment 3 explicitly looked at the relationship between the bias for predictive over nonpredictive stimulus dimensions, calculated as the difference in dwell time to the relevant minus the dwell time to the irrelevant dimension early in Stage 2 of the optional-shift task and subsequent accuracy keyboard performance in the late trials of Stage 2. Results showed that a greater early bias for the predictive dimension correlated positively with critical test trial accuracy later in the task (r = .55).

Taken together, Experiments 2, 3 and 4 accorded well with the suggestion of a common mechanism underlying acquired equivalence and at least some forms of attentional set and are explicable in terms of the connectionist network proposed by Honey and colleagues (Honey, 2000; Honey et al., 2010). The network is able to explain the generalisation of responding to test trials observed after revaluation in our configural acquired equivalence task (Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+ → A+, B-, C ?, D ?) by allowing hidden units to be shared by similar stimuli that have been trained to signal the same outcomes (*acw+*, *acx-*, *bdw-*, *bdx+*). These hidden units, shared by patterns of stimulation that have not necessarily been presented in the same trial but that are mediated by the same outcome, are crucial to anticipate this and other forms of configural acquired equivalence. They are a departure from other connectionist networks that only contemplate changes to the activation of hidden units that represent the specific pattern of stimulation present in any one trial (e.g., Pearce, 1994). Similarly, the network is also able to accommodate the finding that participants in our optional-shift task responded based on the stimulus dimension that had been relevant during training. In brief, the initial discrimination would result in hidden units *awx+* and *bwx-*, shared by the compounds that have signalled the same outcomes (Aw+, Ax+ and Bw-, Bx-). Here, it should be clear that input A will have a stronger link to hidden unit *awx+* – and input B to hidden unit *bwx-* – than will w or x, because A and B will always be unequivocally followed by the outcome correctly anticipated by their corresponding hidden unit (A → *awx+* and B → *bwx-*). On the contrary, inputs w and x will signal one outcome and the other half of the time, reducing their efficacy to activate the correct hidden unit (w → *awx+/bwx-* and x → *awx+/bwx-*). These differences in input-to-hidden layer connections, governed by Hebbian and anti-Hebbian processes

(later the CPCA algorithm in the formal instantiation of the model), will be responsible for later allowing new stimuli from the initially relevant dimension to exert greater activation to the hidden units than new stimuli from the previously irrelevant dimension.

Alternatively, the configural discriminations described in Chapter 2 could also be explicable in occasion setting terms (Bonardi & Jennings, 2009; Bonardi et al., 2017). For example, in a configural discrimination of the form Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+, stimulus w and x will have two separate associations, one with each of the outcomes (e.g., w → + and x → -). Occasion setters –stimuli A-D – are said to operate on the entire stimulus-outcome association, enabling its activation. If we assume this to be the case, when stimulus A is presented it will act as an occasion setter for the w+ and x- associations. Conversely, stimulus B will act as an occasion setter for the complimentary w- and x+ associations. The revaluation of A with a footshock will result in mediated conditioning to the entire w+ and x- associations. Stimulus C, which enables the same now fear eliciting w+ and x- associations, will elicit more fear than stimulus D, which does not. However, although revaluation acquired equivalence procedures might be explicable in terms of occasion setting and mediated conditioning, other forms of configural acquired equivalence described in this thesis pose a challenge to this account. For example, it is not evident how an occasion setting account could anticipate the finding that a congruent discrimination (Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+ → Az+, Ay-, Bz-, By+, Cz+, Cy-, Dz-, Dy+) is acquired more readily than an incongruent one (Aw+, Ax-, Bw-, Bx+, Cw+, Cx-, Dw-, Dx+ → Az+, Ay-, Bz-, By+, Cz-, Cy+, Dz+, Dy-) (e.g., Honey & Ward-Robinson, 2001). Overall, Honey's connectionist network

appears to be a more adequate model, capable of accommodating a broader range of acquired equivalence procedures.

Chapter 3 investigated different outcome manipulations in configural and non-configural acquired equivalence. The first set of experiments served as exploratory investigations into the effect of presenting the same or different outcomes across training and revaluation in our usual configural acquired equivalence task. To that end, Experiment 5 used a factorial manipulation of our usual outcomes to produce a task with either *the same* set of outcomes across training and revaluation (Bite/Sting → Bite/Sting and Poison/Suffocate → Poison/Suffocate) or *different* outcomes across stages (Bite/Sting → Poison/Suffocate and Poison/Suffocate → Bite/Sting). Stage 2 results demonstrated that participants that experienced different outcomes across stages showed an enhanced performance during revaluation trials and a stronger acquired equivalence effect than participants who experienced the same outcomes across training and revaluation.

This is an experimentally interesting finding for two main reasons. First, with both groups receiving differential outcomes, these group differences cannot be attributed to a simple DOE, in which the expectancy of a given outcome acts as an additional cue to aid discrimination learning. This being the case, the expectancies generated for both groups during training (e.g., *Aw – O1 – R1, Ax – O2 – R2, Bw – O2 – R2, Bx – O1 – R1, Cw – O1 – R1, Cx – O2 – R2, Dw – O2 – R2,* and *Dx – O1 – R1)* should facilitate further learning in group Same (*A – O1 – R1* and *B – O2 – R2*) over group Different, which would have to update their now inaccurate expectancies (*A – O3 – R1* and *B – O4 – R2*). Second, prior research has measured acquired equivalence using the same appetitive outcomes across stages (e.g., Coutureau et al.,

2002; Iordanova et al., 2007) and using appetitive outcomes during training followed by a footshock during revaluation (e.g., Honey & Watt, 1998; Ward-Robinson & Honey, 2000). A direct comparison of the effect sizes in previous experiments would be an inadequate approach to assessing the effect of having the same or different differential outcomes across stages in performance, because the finding of an enhanced performance (e.g., in an experiment where rats received appetitive training followed by a mild footshock) could simply reflect differences in the intrinsic properties of the outcomes (e.g., footshock is a more potent reinforcer than a food pellet) rather than differences in the group treatments per se. To our knowledge, experiments in this chapter were the first to factorially manipulate the outcomes in a single between-subjects task to control for any intrinsic differences and directly compare the effects of receiving the same or a different set of outcomes in performance. Results from Experiment 5 suggest that subsequent research could benefit from using different outcomes across stages when using a revaluation configural acquired equivalence task. However, the interpretation of these findings is somewhat limited by the fact that participants in group Same failed to show an acquired equivalence effect. This is particularly unfortunate because group Same completed the exact same task used in Experiments 1, 2, 3 and 4 in the previous chapter, the results of all of which demonstrated the acquired equivalence effect.

We argued that a network with the characteristics of Honey's model could explain findings from Experiment 5 by appealing to the critical role of the outcome in mediating the strength of the input-to-hidden connections. By the end of training, the connections between each combination of inputs and their corresponding hidden unit should be at asymptote. That is, Aw will strongly activate hidden unit *acw+*, Bw will strongly activate hidden unit *bdw-* and the same will be true for the rest of the

configurations. In group Same, that asymptotic activation will result in revaluation trials A+ and B- eliciting a very strong activation to their respective *acw+* and *bdx-* hidden units. This strong activation to the hidden unit will set in motion the activity that will propagate across adjacent layers of the network. That is, the strong activation to the hidden unit (e.g., *acw*) will elicit great activation in the correct output unit (e.g., +), which will feedback again to the hidden unit and strengthen the connection between the two. The connection between input A and hidden unit *acw+* will consequently be strengthen during these trials. However, the strong reciprocal activity in the correct hidden and output units, and the corresponding strengthening of the A → *acw+* connection in the absence of input C will result in a proportionate reduction in the ability of input C to activate hidden unit *acw+*, which could manifest as a more rapid reduction in the acquired equivalence effect. In group Different, the presentation of A* and B$ during revaluation should not result in the same reciprocal asymptotic activation of the *acw+* and *bdx-* links and the connection between C and hidden unit *acw* might extinguish more slowly, which would result in a comparably stronger acquired equivalence effect.

Experiment 6 and later Experiment 7 in Chapter 3 incorporated additional non-configural cues to our usual initial discrimination. We reasoned that by presenting all possible trial outcomes from the onset of the task the potential differences in arousal due to the differential group treatments during revaluation trials would be reduced, and that this would provide more decisive evidence of the effect of changing outcomes across stages in performance. Both groups demonstrated the acquired equivalence effect. However, the enhanced performance during revaluation and test trials previously observed in group Different disappeared. These results could mean that the differences in performance observed in

Experiment 5 might have indeed been driven by the differential levels of arousal or interest in the task between groups Same and Different. Even if this were the case, it could still be worth it to use different outcomes across training and revaluation in the future to help increase the effect sizes that are often reported to be small in these sort of tasks (e.g., Honey & Hall, 1989).

The second set of experiments in Chapter 3 were intended to replicate and extend the generality of the effect reported in Experiment 3 in Delamater (1998), which found that rats that received a reversal with different outcomes within stimulus modality (A1+, A2-, V1-, V2* → A1-, A2*, V1+, V2-) acquired the discrimination more readily than rats that received a reversal with the same outcomes within stimulus modality (A1+, A2-, V1-, V2* → A1-, A2+, V1*, V2-). These results are theoretically significant because the group differences in reversal acquisition cannot be attributed to DOE nor be accommodated by a mediated conditioning explanation of the sort used by Honey and Hall (1989) to explain results in a non-configural acquired equivalence task. For example, at the start of the first reversal stage, presenting A1 in both groups will activate the mental representation of food, which will no longer be delivered. Presenting A2 should initially activate the representation of no food. Instead, a food pellet and sucrose will be delivered in the group with the same outcomes within stimulus modality and the group with different outcomes, respectively. That is, with each stimulus and outcome representation equally reinforced and nonreinforced during the reversal stages of both groups, there are no grounds for mediated conditioning to anticipate group differences in reversal acquisition.

Experiments 8, 9 and 10 in this chapter failed to replicate Delamater's (1998) findings in our participants. Data from these three experiments showed that

participants were performing at asymptote after the first half of training, even when we attempted to increase the difficulty of the task by increasing the number of exemplar. This asymptotic performance may have diminished the overall sensitivity of the task to group differences. Participants showed a small numerical advantage in reversals with different outcomes within stimulus modality, more evident in Experiment 10, but these numerical differences were marginal and participants in both groups were overall comparably good at learning the discrimination during the reversal stages. These results could indicate that humans may use mental processes very different to the ones used by rats when approaching these tasks by, for example, assigning verbal labels to stimuli. Even in Experiment 10, which used black and white checkboards generated randomly, participants unofficially reported having identified patterns in the checkboards and having focused on particular features to solve the discrimination (e.g., "The one with the white T" or "The one with the long Z"). An alternative explanation is that our tasks were not ideal to replicate Delamater's (1998) findings. For example, our experiments used four completely differentiated outcomes (e.g., keys T, D, V and H) as opposed to two differential outcomes (a food pellet and sucrose) and the absence of these two outcomes. Delamater's findings relied on rats discriminating stimuli from the auditory and visual modalities. In contrast, participants in Experiments 8 and 9 learned about sets of stimuli that were distinct (i.e., bears and snakes) but belonged to the same visual modality.

Delamater (1998) favoured a connectionist interpretation of these findings, by appealing to a model that would allow for the internal representations of stimuli to converge or diverge based on the outcome they have signalled. Subsequently, Delamater (2012) proposed a connectionist network with dedicated sets of modality-

specific and multimodal hidden units mediating the connections between the input and output units. However, Honey's network offers a more parsimonious network structure that does not necessitate of stimulus-modality-specific pathways connecting inputs and outputs. Instead, Honey's model could accommodate findings from Delamater (1998) by assuming that common elements (e.g., *x* and *y*) drive generalisation between stimuli belonging to a given modality (e.g., visual or auditory, respectively), but not across modalities. In any case, experiments in Chapter 3 cannot unequivocally confirm or refute Honey's network because their interpretation is based on an informal description of the model. However, they are useful in providing experimental evidence to be compared against a formal implementation of the network described by Honey and colleagues.

Chapter 4 presented a recently published version of a formal computational instantiation of Honey's connectionist network and investigated its ability to accommodate some of the key experimental findings presented during Chapters 2 and 3 in this thesis. Although Honey and colleagues (Honey, 2000; Honey et al., 2010; Honey & Ward-Robinson, 2002) have carefully described the characteristics of their network, verbal descriptions can be prone to unforeseen errors. By formalising an instantiation of the model, Robinson et al. (2019) were forced to explicitly define the theoretical assumptions and implications of Honey's model, writing them down in the form of programming code that can be amended if necessary and that can help generate new research hypotheses based on the simulated data (Guest & Martin, 2020).

Robinson et al. (2019) showed that their learning network could successfully simulate training, revaluation and test data from a revaluation configural acquired equivalence task. However, we reasoned that an inability for the network to simulate

learning and acquired equivalence in a task with an intermixed revaluation and test stage would undermine it, in its departure from the experimental data presented in Experiments 1, 2, 3, 4, 5 (partially) and 7, which demonstrated the acquired equivalence effect in a 2-Stage procedure. This simulation in Chapter 4 showed the correct pattern of output unit activation when simulating training and intermixed revaluation and test trials with either the same outputs or different outputs across stages. The simulations showed that the level of activation to the output units was undifferentiated when presenting stimulus A-D after training, reflecting the configural nature of the discrimination. After the intermixed revaluation (A+, B-) and test (C, D) simulated trials, the pattern of output unit activation elicited by A transfer correctly to C without explicit feedback. Conversely, the correct pattern of output activation generated by stimulus B transfer to stimulus D. That is, the acquired equivalence effect. Simulations in Chapter 4 also showed that the levels of output activation generated in these two simulations (one with the same and one with different outcomes across stages) were reliably different. A direct comparison between the levels of absolute correct output activation, measured as the activation to the correct output unit minus the activation the incorrect output unit, revealed an enhanced level of output activation during revaluation trials and an enhanced acquired equivalence effect in the simulation with different outcomes across stages. That is, the findings yielded by the simulated data were analogous to our experimental findings in Experiment 5 of Chapter 3.

This computational support of our findings is important because, unlike human participants, simulations in the Hebbian learning network are completely agnostic as to the nature of the outcome. That is, for the network outcomes are either the same (i.e., the same output elements are turned on in the Stage 1 and Stage 2

script matrices) or different (i.e., the output elements turned on in the script during Stage 1 are different to those turned on during Stage 2). With the same learning rate parameters in both simulations, an enhanced performance in one of them must therefore be down to the fact that different outcomes were activated across stages. Simulating these data also yielded interesting additional information. In the simulation with the same outcomes across stages, over half of the individual networks generated the *wrong* pattern of output activation in test trials. However, when simulating the task with different outcomes across stages, no individual network yielded the incorrect pattern of activation. This confirms that presenting the same outcomes over two epochs of revaluation in the absence of feedback was enough for the connections between inputs C and D and their corresponding hidden and output units to be disrupted.

Adding non-configural inputs to the simulation of configural acquired equivalence, analogous to Experiment 7 in the previous chapter, revealed that both simulations (with either same or different outcomes across stages) showed the acquired equivalence effect, with the correct pattern of activation generalising from inputs A and B to inputs C and D, respectively. These simulations were interesting because they showed that, whilst both generated the correct pattern of activation, adding non-configural inputs to the simulation with different outcomes resulted in an overall decrease in the levels of activity and an increase in residual output activity (i.e., the activity to the outputs used during Stage 1). Adding non-configural inputs to the simulation with the same outputs across stages did not seem to have much of an effect. The levels of activation resulting from this simulation were very similar to those obtained in the purely configural simulation. Also in line with the configural simulation, almost half of the networks generated the wrong pattern of activation.

However, the notable disparity in levels of output activity resulted in group Same showing an advantage over group Different in this simulation. Taken together with the fact that increasing the number of epochs of revaluation (in simulations not formally reported in this thesis) improved overall levels of output activation, these simulations suggest that mixing configural and non-configural trials, at least when different outcomes are presented across stages, makes it harder for the network to readjust the connections across layers.

These findings are qualitatively comparable to our findings in Experiment 7. Both groups showed the acquired equivalence effect. However, the advantage for group Different observed in Experiment 5 disappeared and group Same showed a small numerical advantage, albeit non-reliable, during test performance. Also in line with findings from Experiment 7, group Different showed more progressive learning during revaluation trials compared to group Same.

Honey (2000) used configural acquired equivalence discriminations as a means to illustrate the need for a theoretical account to incorporate the fact that stimuli that have been presented on separate trials can come to activate a common associate. Robinson et al., (2019) subsequently demonstrated that a formal implementation of Honey's model could successfully simulate a number of configural acquired equivalence procedures. However, and given the fact that adding non-configural inputs to our simulations resulted in non-anticipated notably lower levels of output activation, we reasoned it would be instructive to examine whether this instantiation of Honey's model could also anticipate Delamater's (1998) findings, which are inexplicable in terms of mediated conditioning despite being a simple non-configural form of acquired equivalence.

The network replicated Delamater's (1998) group differences in a simulation analogous to Experiment 8, which presented four exemplars from two distinct visual dimensions, and Experiment 10, which presented 12 exemplars belonging to a visual and auditory modality, of this thesis. That is, the simulated group that received a reversal with different outcomes within stimulus modality showed an enhanced performance, reflected as a significant reduction in mean root square error, compared to a simulated group that received the same outcomes within stimulus modality.

These findings are theoretically important for two main reasons. On the one hand, they confirm that the formal implementation of Honey's connectionist network can accommodate findings from experimental designs other than configural acquired equivalence, which any model that claims to be a general model of discrimination learning should do. On the other hand, they also confirm that this simple connectionist network can accommodate these findings without needing to invoke a more complex modality-specific and multimodal hidden unit structure (Delamater, 2012). The implementation of the model does this by assuming that inputs that code for stimuli that belong to a specific modality (e.g., auditory) share a common element between them (e.g., *A1x, A2x*), but that no common element is shared across modalities (e.g., *A1x, A2x, V1y, V2y*), reducing the scope for stimulus generalisation. This implementation of Honey's model was able to replicate Delamater's (1998) group differences even when a common element was added to all inputs, to reflect that stimuli were distinct – bears and snakes – but all belonged to the visual modality in Experiment 8 (i.e., *B1xa, B2xa, S1ya, S2ya*). Contrary to the successful simulations of Experiments 8 and 10, the model failed to replicate the advantage for group Different in a simulation analogous to that of Experiment 9, which involved the presentation of 12 visual exemplars. From the simulation of Experiment 8, it is

clear that the network can solve a simple discrimination where all inputs share a common element. The simulation of Experiment 10 also demonstrated that the current instantiation of the network can solve a discrimination where 12 inputs share fewer common elements. This failure to obtain group differences in the simulation of Experiment 9, which combined 12 inputs and a common element shared across all inputs, helps qualify the present instantiation of Honey's model. It shows that the model is currently unable to simulate group differences when an increased number of inputs share common elements between and within modalities, which should guide any subsequent amendments to the network.

5.1 Future research

5.1.1 Configural acquired equivalence and attentional set

Experiments in Chapter 2 offered overall substantial support for a positive correlation between performance at test in a configural acquired equivalence and optional-shift task. However, the fact that the measure of attentional set derived from IDS/EDS did not correlate with our acquired equivalence task is a challenge to Honey's model, which used IDS/EDS specifically to illustrate how the same network structure used to solve a configural discrimination could accommodate discriminations in which some elements of a compound are predictive and some are nonpredictive of the outcome (Honey et al., 2010). Instead, these results suggest the need to supplement associative processes with explicit attentional processes to explain how stimuli gain and lose distinctiveness in IDS/EDS. To help confirm or refute the findings presented in Chapter 2, future research should assess performance

in configural acquired equivalence, optional-shift and IDS/EDS in a single task. If performance once again correlated positively between test trials in the configural acquired equivalence and optional-shift tasks, but not between configural acquired equivalence and IDS/EDS or between optional-shift and IDS/EDS, this could be the confirmation that these tasks may not be measuring the same construct.

In the current thesis, we presented a pilot IDS/EDS task that attempted to address some of the shortcoming of the CANTAB IDS/EDS. Namely, our task counterbalanced stimuli and outcomes and presented intermixed IDS and EDS trials to account for possible time artefacts. The task failed to replicate the usual IDS superiority. However, there are indications that improving this pilot task might still be an enterprise worth pursuing. The analysis of the IDS and EDS data per relevant dimension showed a reliable increase in performance when the colour dimension underwent an IDS compared to when it underwent an EDS. That is, a discrimination where the colour dimension was relevant during training and new colour exemplars were also relevant during IDS trials was more readily acquired than a discrimination where the colour dimension went from being relevant during training to being irrelevant during EDS trials. However, the differences in performance between IDS and EDS trials were not significant for any of the other dimensions. This finding is important because it shows that the task has the potential to detect differences in performance between IDS and EDS. Previous research has shown that discriminations involving colour are easier to perform for animals that those involving shapes or line orientations (e.g., Mackintosh & Little, 1969). A different choice of the stimulus dimensions accompanying colour could improve the sensitivity of the task in the future. An IDS superiority in this task would be inexplicable in terms of stimulus generalisation and would not reflect any time-

related effects. Additionally, it would naturally provide a single attentional set datum per participant, which could be easily correlated with our measure of acquired equivalence and with performance in optional-shift.

5.1.2 Does this instantiation of Honey's model simulate attentional data?

Chapter 2 in this thesis investigated the relationship between configural acquired equivalence and attentional set using a correlation approach, which yielded substantial support in favour of a positive correlation between performance in both when attentional set was measured with an optional-shift task. The ability for the current implementation of Honey's network to accommodate different forms of acquired equivalence was well documented by Robinson et al., (2019) and in Chapter 4 of this thesis. However, the question of whether this instantiation of Honey's model is capable of simulating attentional tasks still remains.

There are existing examples of connectionist networks capable of formally accounting for the relationship between attention and associative learning. For example, George and Pearce (2012) presented a formal extension of Pearce's (1994) connectionist network to reflect the changes in attention to stimuli that occur during conditioning. To that end, they incorporated two attentional parameters – α and σ – to Pearce's configural network. The former served the purpose of altering the effective salience of a stimulus by enhancing the salience of stimuli that signal events of significance and lowering the salience of stimuli that are irrelevant. The latter focused on uncertain stimuli, altering the associability at the hidden unit level so that it would increase on trials where the outcome is surprising and decrease on trials where the outcome is expected. George and Pearce demonstrated that this

network could accommodate attentional changes in conditioning. For example, the network correctly anticipated an IDS superiority in a simulation of IDS/EDS and correctly anticipated the differences in associability in a discrimination of the form Ax+ Bx-, which showed a decline in α for stimulus x and an increase for A and B.

In the present model, which does not contemplate specific attentional parameters, successfully simulating attentional set tasks could confirm that there is no need to appeal to anything more than associative processes to accommodate these findings. A failure of this implementation of the model in simulating attentional set data would mean that Honey et al. (2010) are wrong in their suggestion of applying the network to both forms of learning and that this simple connectionist network is not an adequate model to accommodate these data. Or that, at least, the current description of the model needs revisiting. Furthermore, cloning the networks and running a single simulation of configural acquired equivalence and IDS/EDS or optional-shift would yield a within-networks simulation of performance in these two tasks, which would be of most interest to this thesis in confirming or refuting Honey et al.'s (2010) claims.

5.1.3 General considerations for the current implementation of the model

The simulations reported in Chapter 4 did a reasonable job accommodating experimental data in the only formal instantiation of Honey's model available to date. However, it is important to remember that this is one formal implementation of Honey and colleagues' verbal description of the network and that, in light of additional experimental data, some amendments might be needed. The simulation of Experiment 7, which intermixed configural and non-configural trials, showed that

although the pattern of activation was correct, overall levels of output unit activation could be very low. Similarly, Robinson et al. (2019) briefly noted how the network could simulate Honey and Hall's (1989) acquired equivalence task, albeit with low levels of correct output unit activation. Empirical effect sizes can be small in acquired equivalence tasks (e.g., Honey & Hall, 1989). However, they are similarly small in configural and non-configural acquired equivalence. In the simulations reported in Chapter 4 however, the levels of output activation were very substantially different when simulating configural and non-configural tasks. These substantial differences differ from empirical findings and suggest that Robinson et al. (2019) should consider amending the learning algorithms involved to optimise activation levels when the model is presented with non-configural inputs.

Simulations have confirmed that the current implementation of Honey's model can accommodate a number of forms of acquired equivalence. Robinson et al. (2019) showed that the model can accommodate findings from revaluation configural acquired equivalence, congruent and incongruent context discriminations, whole vs. partial reversal discrimination acquisition and congruent vs. incongruent discrimination acquisitions. They also briefly mentioned the successful simulation of Honey and Hall's (1989) simple acquired equivalence procedure. In this thesis, we have added to those demonstrations the ability for the network to deal with intermixed revaluation and test stages or to simulate group differences in reversal acquisition of the kind reported in Delamater (1998). However, it would be instructive to investigate whether the network can accommodate other procedures in its current form. For example, animals can perform feature negative discriminations in which two stimuli signal one outcome when presented separately (e.g., food) and a different outcome (e.g., no food) when presented together (i.e., A+, B+, AB-). If

stimuli that anticipate the same outcome (A and B in the discrimination above) come to activate the same hidden unit (e.g., *ab+*), which is a critical feature of the current model, it is unclear what mechanism would allow the network as it currently stands to develop a connection to a different hidden and output unit when the same stimuli are presented as a compound. Simulating this and other types of discrimination learning tasks could help qualify the current instantiation of the model further.

Finally, formal computational modelling is a way of promoting transparent science. It can help advance theories by offering formal implementations that can be shared, replicated or amended to better reflect the assumptions of any given theory. To help encourage open science, it could be beneficial to replicate Robinson et al.'s (2019) instantiation of the model using open-source software, which would facilitate access to the network to researches that may otherwise have no access to licensed software, broadening its reach.

5.2 Concluding comments

Configural learning discriminations challenge unique cue accounts of associative learning (e.g., Rescorla & Wagner, 1972) and lend themselves to configural explanations (e.g., Pearce, 1994). However, alternative models are needed to accommodate the finding that following a configural discrimination, stimuli are treated as more similar when they have shared a common training history. That is, the acquired equivalence effect. Honey et al.'s (2000, 2010) connectionist network offered the highest explanatory power to accommodate extant experimental findings for different forms of configural acquired equivalence.

Experiments in Chapter 2 in this thesis found evidence for the acquired equivalence, optional-shift and IDS superiority effects in our human participants.

Performance was found to correlate positively between test trials in our configural acquired equivalence and optional-shift tasks. However, no correlation between performance in acquired equivalence and IDS/EDS was found. These results could be seen as a challenge to Honey's network, which used IDS/EDS as a means to explain how a simple connectionist network could also anticipate performance in an attentional set task. However, experiments in this chapter were not able to unambiguously support or refute Honey et al.'s (2010) claims.

Experiments in Chapter 3 showed an enhanced acquired equivalence effect in participants who had received different outcomes across training and revaluation compared to participants who received the same outcomes across stages. This chapter also showed that this enhanced acquired equivalence effect disappeared when configural and non-configural trials were presented during training. The second set of experiments in this chapter attempted to extend the generality of the finding that presenting different outcomes within stimulus modality result in an enhanced reversal acquisition compared to presenting the same findings within stimulus modality (Delamater, 1998). However, participants in our experiments did not show reliable differences in reversal acquisition. Findings in this chapter were still inconclusive in supporting or challenging the network because our interpretation of the results was based on an informal description of Honey's model.

Chapter 4 tested a formal implementation of Honey's network (Robinson et al., 2019) on its ability to accommodate experimental findings presented throughout this thesis. The network was successful in simulating (i) performance in our 2-Stages configural acquired equivalence task. (ii) An enhanced revaluation performance and acquired equivalence effect in a simulation with different outcomes across training and revaluation, which disappeared once configural and non-configural stimuli were

presented during training. (iii) Delamater's (1998) finding of an enhanced reversal acquisition in a simulation with different outcomes within stimulus modality, suggesting that our failure to replicate the results may have been down to the notable experimental differences between the two tasks.

Overall, this thesis offered support for the ability of Honey's network to account for performance on a number of discrimination learning tasks. The formal implementation of the model, which became available later during this thesis, offered important computational support for Honey's learning network in its ability to simulate findings that, we reasoned, should be captured by the network as it was described. Additionally, simulating the experimental data presented in this thesis offered valuable insights that could inform future research and a potential optimisation of the model.

References

Aristizabal, J. A., Ramos-Álvarez, M. M., Callejas-Aguilera, J. E., & Rosas, J. M. (2016). Attention to irrelevant contexts decreases as training increases: Evidence from eye-fixations in a human predictive learning task. *Behavioural Processes, 124*, 66–73. https://doi.org/10.1016/j.beproc.2015.12.008

Audacity Team, (2019). Audacity [Computer software]. Retrieved from http://audacityteam.org.

Baron-cohen, S., Leslie, A. M., & Frith, U. T. (1985). the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Baxter, M. G., & Gaffan, D. (2007). Asymmetry of attentional set in rhesus monkeys learning colour and shape discriminations. *Quarterly Journal of Experimental Psychology*, *60*(1), 1–8. https://doi.org/10.1080/17470210600971485

Bergvall, Å. H., Wessely, H., Forsman, A., & Hansen, S. (2001). A deficit in attentional set-shifting of violent offenders. *Psychological Medicine*, *31*(6), 1095–1105.

Bonardi, C., Rey, V., Richmond, M., & Hall, G. (1993). Acquired Equivalence of Cues in Pigeon Autoshaping: Effects of Training With Common Consequences and With Common Antecedents. *Animal Learning and Behaviour*, *21*(4), 369–376. https://doi.org/10.3758/BF03198003

Bonardi, C., & Jennings, D. (2009). Journal of Experimental Psychology: Animal Behavior Processes: Editors. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(3), 440–445. https://doi.org/10.1037/0097-7403.29.1.c2

Bonardi, C., Robinson, J., & Jennings, D. (2017). Can existing associative principles

explain occasion setting? Some old ideas and some new data. *Behavioural Processes, 137*, 5–18.

Bünger, A., Urfer-Maurer, N., & Grob, A. (2019). Multimethod assessment of attention, executive functions, and motor skills in children with and without ADHD: Children's performance and parents' perceptions. *Journal of Attention Disorders*.

Cambridge Cognition, (2019). CANTAB [Cognitive assessment software]. Retrieved from http://.cantab.com.

Carlson, J. G., & Wielkiewicz, R. M. (1972). Delay of reinforcement in instrumental discrimination learning of rats. *Journal of Comparative and Physiological Psychology*, *81*(2), 365.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

Coutureau, E., Killcross, A. S., Good, M., Marshall, V. J., Ward-robinson, J., & Honey, R. C. (2002). Acquired Equivalence and Distinctiveness of Cues : II . Neural Manipulations and Their Implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(4), 388–396. https://doi.org/10.1037//0097-7403.28.4.388

Delamater, A. R. (1998). Associative mediational processes in the acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(4), 467–482. https://doi.org/10.1037/0097-7403.24.4.467

Delamater, A. R. (2012). On the nature of CS and US representations in Pavlovian learning. *Learning and Behavior*, *40*(1), 1–23. https://doi.org/10.3758/s13420-011-0036-4

Delamater, A. R., & Joseph, P. (2000). Common coding in symbolic matching tasks

with humans: Training with a common consequence or antecedent. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *53*(3), 255–273. https://doi.org/10.1080/027249900411182

Delamater, A. R., Kranjec, A., & Fein, M. I. (2010). *Differential outcome effects in Pavlovian biconditional and ambiguous occasion setting tasks*. *36*(4), 471–481. https://doi.org/10.1037/a0019136.Differential

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837.

Duffaud, A. M., Killcross, S., & George, D. N. (2007). Optional-shift behaviour in rats : A novel procedure for assessing attentional processes in discrimination learning. *The Quarterly Journal of Experimental Psychology*, *60*(4), 37–41. https://doi.org/10.1080/17470210601154487

Edwards, C. A., Jagielo, J. A., Zentall, T. R., & Hogan, D. E. (1982). Acquired equivalence and distinctiveness in matching to sample by pigeons: Mediation by reinforcer-specific expectancies. *Journal of Experimental Psychology: Animal Behavior Processes*, *8*(3), 244–259. https://doi.org/10.1037/0097-7403.8.3.244

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

Estévez, A. F., Fuentes, L. J., Marı-Bêffa, P., González, C., & Alvarez, D. (2001). The differential outcome effect as a useful tool to improve conditional discrimination learning in children. *Learning and Motivation*, *32*(1), 48–64.

Garner, J. P., Thogerson, C. M., Wurbel, H., Murray, J. D., & Mench, J. A. (2006).

Animal neuropsychology : Validation of the Intra-Dimensional Extra-Dimensional set shifting task for mice. *Behavioural Brain Research*, *173*(1), 53–61. https://doi.org/10.1016/j.bbr.2006.06.002

George, D. N. (2018). *HebbianNN [electronic resource: Matlab source code]*. https://github.com/DavidNGeorge/HebbianNN.git

George, D. N., & Pearce, J. M. (2012). A configural theory of attention and associative learning. *Learning and Behavior*, *40*(3), 241–254. https://doi.org/10.3758/s13420-012-0078-2

Grice, G. R., & Davis, J. D. (1958). Mediated stimulus equivalence and distinctiveness in human conditioning. *Journal of Experimental Psychology*, *55*(6), 565–571.

Guest, O., & Martin, A. E. (2020). *How computational modeling can force theory building in psychological science.*

Hall, G., Ray, E., & Bonardi, C. (1993). Acquired equivalence between cues trained with a common antecedent. *Journal of Experimental Psychology: Animal Behavior Processes*, *19*(4), 391.

Hampshire, A., & Owen, A. M. (2010). Clinical studies of attention and learning. In *Attention and Associative Learning From Brain to Behaviour* (pp. 385–406). Oxford University Press Oxford.

Harvey, P. O., Fossati, P., Pochon, J. B., Levy, R., LeBastard, G., Lehéricy, S., Allilaire, J. F., & Dubois, B. (2005). Cognitive control and brain resources in major depression: an fMRI study using the n-back task. *Neuroimage*, *26*(3), 860–869.

Haselgrove, M., Le Pelley, M. E., Singh, N. K., Teow, H. Q., Morris, R. W., Green, M. J., Griffiths, O., & Killcross, S. (2016). Disrupted attentional learning in

high schizotypy: Evidence of aberrant salience. *British Journal of Psychology*, *107*(4), 601–624. https://doi.org/10.1111/bjop.12175

Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall.

Hodder, K. I., George, D. N., Killcross, S., & Honey, R. C. (2003). Representational blending in human conditional learning: Implications for associative theory. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *56 B*(2), 223–238. https://doi.org/10.1080/02724990244000269

Hogarth, L., Dickinson, A., & Duka, T. (2010). The associative basis of cue-elicited drug taking in humans. *Psychopharmacology*, *208*(3), 337–351.

Holland, P. C. (1981). Acquisition of representation-mediated conditioned food aversions. *Learning and Motivation*, *12*(1), 1–18. https://doi.org/10.1016/0023-9690(81)90022-9

Honey, R. C. (2000). The experimental psychology society prize lecture: Associative priming in Pavlovian conditioning. *Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *53*(1), 1–23. https://doi.org/10.1080/027249900392977

Honey, R. C., Close, J., & Lin, T. E. (2010). Acquired distinctiveness and equivalence: A synthesis. *Attention and Associative Learning: From Brain to Behaviour*, 159–186.

Honey, R. C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology. Animal Behavior Processes*, *15*(4), 338–346. https://doi.org/10.1037/0097-7403.15.4.338

Honey, R. C., & Ward-Robinson, J. (2001). Transfer Between Contextual

Conditional Discriminations An Examination of How Stimulus Conjunctions Are Represented. *Journal of Experimental Psychology: Animal Behavior Processes*, *27*(3), 196–205.

Honey, R. C., & Ward-Robinson, J. (2002). Acquired Equivalence and Distinctiveness of Cues : I . Exploring a Neural Network Approach. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(4), 378–387. https://doi.org/10.1037//0097-7403.28.4.378

Honey, R. C., & Watt, A. (1998). Acquired relational equivalence between contexts and features. *Journal of Experimental Psychology: Animal Behavior Processes*, *25*(3), 324–333. https://doi.org/10.1037/0097-7403.25.3.324

Iordanova, M. D., Killcross, A. S., & Honey, R. C. (2007). Role of the Medial Prefrontal Cortex in Acquired Distinctiveness and Equivalence of Cues. *Behavioral Neuroscience*, *121*(6), 1431–1436. https://doi.org/10.1037/0735-7044.121.6.1431

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394–412.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

JASP Team, (2019). JASP (Version 0.11.1) [Computer software]. Retrieved from https://jasp-stats.org.

Jazbec, S., Pantelis, C., Robbins, T., Weickert, T., Weinberger, D. R., & Goldberg, T. E. (2007). Intra-dimensional/extra-dimensional set-shifting performance in schizophrenia: impact of distractors. *Schizophrenia Research*, *89*(1), 339–349.

Jeffreys, H. (1961). *Theory of probability, 3rd edn oxford: Oxford university press*.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329.

Kamin, L. J. (1968). Attention-like processes in classical conditioning. *Miami Symposium on the Prediction of Behavior: Aversive Stimulation*, 9–31.

Kempton, S., Vance, A., Maruff, P., Luk, E., Costin, J., & Pantelis, C. (1999). Executive function and attention deficit hyperactivity disorder: stimulant medication and better executive function performance in children. *Psychological Medicine*, *29*(3), 527–538.

Kéri, S., Nagy, O., Kelemen, O., Myers, C. E., & Gluck, M. A. (2005). Dissociation between medial temporal lobe and basal ganglia memory systems in schizophrenia. *Schizophrenia Research*, *77*(2–3), 321–328.

Lai, M. L., Tsai, M. J., Yang, F. Y., Hsu, C. Y., Liu, T. C., Lee, S. W. Y., Lee, M. H., Chiou, G. L., Liang, J. C., & Tsai, C. C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, *10*, 90–115.

Lawrence, A. D., Sahakian, B. J., Hodges, J. R., Rosser, A. E., Lange, K. W., & Robbins, T. W. (1996). Executive and mnemonic functions in early Huntington's disease. *Brain*, *119*(5), 1633–1645.

Lawrence, D. H. (1949). Acquired distinctiveness of cues: I. Transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology*, *39*(6), 770–784. https://doi.org/10.1037/h0058097

Lawrence, D. H. (1950). Acquired distinctiveness of cues: II. Selective association in a constant stimulus situation. *Journal of Experimental Psychology*, *40*(2), 175.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *The Quarterly Journal of*

Experimental Psychology Section B, *57*(3b), 193–243.

Le Pelley, M. E., Beesley, T., & Griffiths, O. (2011). Overt Attention and Predictiveness in Human Contingency Learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(2), 220–229. https://doi.org/10.1037/a0021384

Le Pelley, M. E., Mitchell, C. J., & Johnson, A. M. (2013). Outcome value influences attentional biases in human associative learning: Dissociable effects of training and instruction. *Journal of Experimental Psychology: Animal Behavior Processes*, *39*(1), 39–55. https://doi.org/10.1037/a0031230

Le Pelley, M. E. (2010). Attention and human associative learning. *Attention and Associative Learning: From Brain to Behaviour*, 187–215.

Lee, D. K. (2016). Alternatives to P value: Confidence interval and effect size. *Korean Journal of Anesthesiology*, *69*(6), 555–562. https://doi.org/10.4097/kjae.2016.69.6.555

Mackintosh, N. J. (1974). The psychology of animal learning. In *The psychology of animal learning*. Academic Press.

Mackintosh, N. J., & Little, L. (1969). Intradimensional and extradimensional shift learning by pigeons. *Psychonomic Science*, *14*(1), 5–6.

McCormack, J. C., Elliffe, D., & Virués-Ortega, J. (2019). Quantifying the effects of the differential outcomes procedure in humans: A systematic review and a meta-analysis. *Journal of Applied Behavior Analysis*, *52*(3), 870–892. https://doi.org/10.1002/jaba.578

Melchers, K. G., Shanks, D. R., & Lachnit, H. (2008). Stimulus coding in human associative learning: Flexible representations of parts and wholes. *Behavioural Processes*, *77*(3), 413–427. https://doi.org/10.1016/j.beproc.2007.09.013

Miller, N. E., & Dollard, J. (1941). *Social learning and imitation.* Yale University Press.

Oswald, C. J. P., Yee, B. K., Rawlins, J. N. P., Bannerman, D. B., Good, M., & Honey, R. C. (2001). Involvement of the Entorhinal Cortex in a Process of Attentional Modulation: Evidence From a Novel Variant of an IDS/EDS Procedure. *Behavioral Neuroscience, 115*(4), 841–849.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25*(1), 46–59. https://doi.org/10.1002/hbm.20131

Owen, A. M., Roberts, A. C., Polkey, C. E., Sahakian, B. J., & Robbins, T. W. (1991). Extra-dimensional versus intra-dimensional set shifting performance following frontal lobe excisions, temporal lobe excisions or amygdalo-hippocampectomy in man. *Neuropsychologia*, *29*(10), 993–1006. https://doi.org/10.1016/0028-3932(91)90063-E

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning: variation in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review*, *94*(1), 61–73. https://doi.org/10.1037/0033-295X.94.1.61

Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychological Review*, *101*(4), 587–607.

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior*, *30*(2), 73–95.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological*

Review, *87*(6), 532.

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

Poremba, A., Saunders, R. C., Crane, A. M., Cook, M., Sokoloff, L., & Mishkin, M. (2003). Functional mapping of the primate auditory system. *Science*, *299*(5606), 568–572.

Quintana, D. S., & Williams, D. R. (2018). Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using JASP. *BMC Psychiatry*, *18*(1), 1–8. https://doi.org/10.1186/s12888-018-1761-4

Rac-Lubashevsky, R., & Kessler, Y. (2016). Decomposing the n-back task: An individual differences study using the reference-back paradigm. *Neuropsychologia*, *90*, 190–199.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372.

Rescorla, R. A. (1991). Associative relations in instrumental learning: The eighteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology*, *43*(1), 1–23.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.

Rizley, R. C., & Rescorla, R. A. (1972). Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology*, *81*(1), 1.

Roberts, A. C., Robbins, T. W., & Everitt, B. J. (1988). The Effects of Intradimensional and Extradimensional Shifts on Visual Discrimination

Learning in Humans and Non-Human Primates. *The Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, *4*(40B), 321–341. https://doi.org/10.1080/14640748808402328

Robinson, J., George, D. N., & Heinke, D. (2019). A Computational Implementation of a Hebbian Learning Network and its Application to Configural Forms of Acquired Equivalence. *Journal of Experimental Psychology: Animal Learning and Cognition*, *45*(3), 356.

Robinson, J., & Owens, E. (2013). Diminished acquired equivalence yet good discrimination performance in older participants. *Frontiers in Psychology*, *4*, 1–8. https://doi.org/10.3389/fpsyg.2013.00726

RStudio Team, (2016). RStudio (Version 1.1.463) [Computer software]. Retrieved from http://www.rstudio.com/.

Sahakian, B. J., & Owen, A. M. (1992). Computerized assessment in neuropsychiatry using CANTAB: discussion paper. *Journal of the Royal Society of Medicine*, *85*(7), 399–402.

Shamay-Tsoory, S. G., Shur, S., Harari, H., & Levkovitz, Y. (2007). Neurocognitive basis of impaired empathy in schizophrenia. *Neuropsychology*, *21*(4), 431.

Simon, H. A. (1954). Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, *49*(267), 467–479.

Simon, J. R., & Gluck, M. A. (2013). Adult age differences in learning and generalization of feedback-based associations. *Psychology and Aging*, *28*(4), 937–947. https://doi.org/10.1037/a0033844

Trapold, M. A. (1970). Are expectancies based upon different positive reinforcing events discriminably different? *Learning and Motivation*, *1*(2), 129–140. https://doi.org/10.1016/0023-9690(70)90079-2

Trapold, M. A., & Overmier, J. B. (1972). The second learning process in instrumental learning. In *Classical conditioning: II. Current research and theory.* (pp. 427–452). New York: Appleton-Century-Crofts.

Urcuioli, P. J. (2005). Behavioral and associative effects of differential outcomes in discrimination learning. *Learning and Behavior*, *33*(1), 1–21. https://doi.org/10.3758/bf03196047

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. *Information Processing in Animals: Memory Mechanisms*, *85*, 5–47.

Wagner, A. R., & Rescorla, R. (1972). *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement.* (Issue X). Appleton-Century- Crofts.

Ward-Robinson, J., & Hall, G. (1999). The role of mediated conditioning in acquired equivalence. *The Quarterly Journal of Experimental Psychology: Section B*, *52*(4), 335–350.

Ward-Robinson, J., & Hall, G. (1999). The role of mediated conditioning in acquired equivalence. *The Quarterly Journal of Experimental Psychology. B, Comparative and Physiological Psychology*, *52*(4), 335–350. https://doi.org/10.1080/713932712

Ward-Robinson, J., & Honey, R. C. (2000). A Novel Contextual Dimension for Use With an Operant Chamber : From Simple to Hierarchical Forms of Learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *26*(3), 358–363. https://doi.org/10.1037//0097-7403.26.3.358