



The University of  
**Nottingham**

UNITED KINGDOM · CHINA · MALAYSIA

# The specific Pavlovian-to- instrumental transfer (PIT) effect in humans

Daniel Alarcón, BSc, MSc

Thesis submitted to the University of Nottingham for the degree of  
Doctor of Philosophy

February 2016

## Publications

The Experiments presented in Chapter 3 of this thesis have been published: Alarcón, D., & Bonardi, C. (2016). The effect of conditioned inhibition on the specific Pavlovian-instrumental transfer effect. *Journal of Experimental Psychology: Animal Learning and Cognition*, 42(1), 82. For reasons of consistency, a different exclusion criterion was used in the mentioned article, which affected the statistical analysis. However, the main results and conclusions of these Experiments were exactly the same as presented in this thesis. Similarly, some of the Experiments reported in Chapter 4 of this thesis are currently being prepared to be submitted with a slightly different statistical analysis.

## Abstract

In Pavlovian conditioning subjects learn the predictive relation between a conditioned stimulus (CS) and a motivationally significant unconditioned stimulus (US), while in instrumental conditioning subjects learn the predictive relation between their responses and a motivationally significant outcome. Both types of associative learning interact in the phenomenon known as the *Pavlovian-to-instrumental transfer* (PIT) effect. In a PIT procedure subjects received Pavlovian conditioning, in which different CSs are paired with different outcomes (CS<sub>1</sub>->O<sub>1</sub>; CS<sub>2</sub>->O<sub>2</sub>, etc), and instrumental training, in which each of different responses are paired with these outcomes (R<sub>1</sub>->O<sub>1</sub>; R<sub>2</sub>->O<sub>2</sub>, etc). After this training the CSs are presented while subjects have the opportunity to perform the instrumental responses. Studies have found that the CS presentations affect instrumental performance by elevating the rate of responding, and this effect can take two different forms: general and specific. In *general PIT*, Pavlovian cues elevate performance of any instrumental responses that have been trained with a reinforcer of a similar motivational valence to the US. But in the *specific PIT* effect a CS paired with a particular outcome selectively elevates instrumental responses that produce that outcome, compared to its effect on responses producing different outcomes, i.e. CS<sub>1</sub>: R<sub>1</sub>>R<sub>2</sub>; CS<sub>2</sub>: R<sub>2</sub>>R<sub>1</sub>.

Different mechanisms have been proposed to explain the specific form of the PIT effect but none of these accounts can explain all the evidence that has been found. Some of these mechanisms propose that at test a CS evokes a representation of the outcome that, in turn, elicits those responses trained with that outcome. In contrast, other accounts suggest that the CS elicits responding via a direct association formed during training. The experiments reported in this thesis were conducted to provide further evidence on this phenomenon in order to distinguish between these mechanisms.

The experiments presented here used a standard PIT task with humans as participants. In the Pavlovian phase participants received presentations of different neutral fractal images (CSs), which were paired with presentations of drink and food images (outcomes). In the instrumental phase participants had to press two keys on a computer keyboard (instrumental responses), which were reinforced with the outcomes. The specific PIT effect was measured in a test in which participants could perform both instrumental responses in the presence and absence of the Pavlovian cues. The experiments reported in Chapter 2 and 3 made use of a procedure known as *conditioned inhibition*, in which a conditioned inhibitor (CI) is trained to signal the absence of an expected outcome; it has been proposed that presentations of a CI suppress the activation of an outcome representation. In the experiments presented in Chapter 2 two CIs were established, one for each of the outcomes, while in those reported in Chapter 3 only one CI was trained. In the studies of both

chapters the effect of the CIs, both alone and in compound with excitatory CSs, on the specific PIT effect was assessed. The findings revealed that the CIs did not exert any measurable effect when they were presented alone, but they reduced the specific PIT effect produced by the excitatory CSs. In Chapter 4 CSs were trained in either a forward or backward relation with the outcomes and their effect on instrumental performance was also measured. In some of the experiments the CSs trained in a backward relation with the outcome produced the specific PIT effect, while in others they did not. The contributions of both backward and forward associations were also assessed, and the results suggest that only the forward association supported the specific PIT effect. Overall, the findings suggest that the specific PIT effect is mediated by the activation of an outcome representation, although some assumptions are needed in order to explain the data with extant accounts of PIT.

## List of contents

Chapter I.....	1
General introduction .....	1
1.1 Pavlovian and instrumental conditioning.....	1
1.2 Pavlovian-to-instrumental transfer (PIT) .....	2
1.3 The specific and general PIT effect .....	3
1.4 The importance of the specific PIT effect.....	6
1.5 Pavlovian conditioning.....	9
1.6 Instrumental conditioning .....	10
1.7 Theoretical explanations of the specific PIT effect.....	15
1.8 The two-process theory.....	16
1.9 Expectancy version of the two-process theory.....	18
1.10 Stimulus-response (S-R) account.....	20
1.11 Evidence on the specific PIT effect.....	23
1.12 The importance of S-O associations to the specific PIT effect .....	25
1.13 The strength of the R-O associations in the specific PIT effect.....	34
1.14 The importance of the outcome value to the specific PIT effect.....	37
1.15 The importance of S-R associations in the specific PIT effect .....	45
1.16 Summary.....	47
Chapter II.....	51
Conditioned inhibition in the specific PIT effect .....	51
2.1 Overview .....	51
2.2 Introduction .....	52
2.3 Conditioned inhibition .....	52
2.4 Measuring conditioned inhibition .....	53
2.5 The nature of conditioned inhibition.....	55
2.6 Specificity of conditioned inhibition .....	59
2.7 Specific PIT effect and conditioned inhibition .....	63
2.8 Experiment 1 .....	66
2.8.1 Method .....	71
2.8.2 Results .....	77
2.8.3 Discussion .....	84
2.9 Experiment 2.....	85
2.9.1 Method .....	87

2.9.2 Results .....	89
2.9.3 Discussion .....	95
2.10 Experiment 3 .....	96
2.10.1 Method.....	98
2.10.2 Results.....	99
2.10.3 Discussion .....	106
2.11 General discussion.....	107
Chapter III.....	111
Outcome-specificity of conditioned inhibition in the specific PIT effect. ....	111
3.1 Overview .....	111
3.2 Introduction .....	112
3.3 Experiment 4 .....	117
3.3.1 Method .....	119
3.3.2 Results .....	121
3.3.3 Discussion .....	127
3.4 Experiment 5 .....	128
3.4.1 Method .....	129
3.4.2 Results .....	130
3.4.3 Discussion .....	135
3.5 Experiment 6 .....	136
3.5.1 Method .....	138
3.5.2 Results .....	139
3.5.3 Discussion .....	147
3.6 Experiment 7 .....	147
3.6.2 Method .....	149
3.6.3 Discussion.....	155
3.7 Experiment 8 .....	155
3.7.1 Method .....	157
3.7.2 Results .....	158
3.7.3 Discussion .....	162
3.8 General Discussion .....	162
Chapter IV .....	171
Backward conditioning in the specific PIT effect .....	171
4.1 Overview .....	171
4.2 Introduction .....	172
4.3 Backward conditioning .....	172

4.4 Inhibitory associations in backward conditioning .....	172
4.5 Excitatory associations in backward conditioning .....	174
4.6 Excitatory and inhibitory associations in backward conditioning .....	175
4.7 The specific PIT effect and backward conditioning .....	178
4.8 Experiment 9 .....	182
4.8.1 Method .....	183
4.8.2 Results .....	185
4.8.3 Discussion .....	187
4.9 Experiment 10 .....	188
4.9.1 Results .....	189
4.9.2 Discussion .....	190
4.10 Experiment 11 .....	191
4.10.1 Method.....	191
4.10.2 Results.....	192
4.10.3 Discussion .....	195
4.11 Experiment 12 .....	197
4.11.1 Method.....	199
4.11.2 Results.....	202
4.11.3 Discussion .....	206
4.12 Experiment 13 .....	207
4.12.1 Method.....	209
4.12.2 Results.....	210
4.12.3 Discussion .....	213
4.13 Experiment 14 .....	214
4.13.1 Method.....	216
4.13.2 Results.....	217
4.13.3 Discussion .....	222
4.14 General Discussion .....	223
Chapter V .....	226
General Discussion.....	226
5.1 Summary of results .....	226
5.1.1 Chapter II: Conditioned inhibition in the specific PIT effect.....	226
5.1.2 Chapter III: Outcome-specificity of conditioned inhibition in the specific PIT effect .....	228
5.1.3 Chapter IV: Backward conditioning in the specific PIT effect.....	231
5.2 Theoretical implications.....	232

5.3 Limitations and future research .....	244
5.4 Conclusion .....	247
References .....	249
Appendix A: Task instructions.....	A-1
Appendix B: Counterbalancing tables .....	B-25

## List of Figures

<i>Figure 1.</i> Sample images of the pictures used as CSs and outcomes. Top panel: Fractal images used as CSs. Bottom panel: food and drink pictures used as outcomes. ....	68
<i>Figure 2.</i> Summation test in Experiment 1. Top panel: Mean ratings of the likelihood of O1 occurrence during FX, FC and FN1, and O2 occurrence during GY, GH and GN2. Bottom panel: Mean ratings of the likelihood of O1 occurrence during X, C and N1, and O2 during Y, H and N2. ....	79
<i>Figure 3.</i> Response rate during the pre-CS period of the PIT test in Experiment 1. F/G: excitatory CS; X/Y: inhibitory stimuli; C/H: pre-exposed control stimuli. ....	81
<i>Figure 4.</i> Mean rate of congruent and incongruent responding for each type of stimulus in the CS period of the PIT test of Experiment 1. Top panel: responses during the excitatory cues F/G. Middle panel: responses during the inhibitory stimuli X/Y. Bottom panel: responses during the pre-exposed control stimuli C/H.....	83
<i>Figure 5.</i> Mean scores grouped by CS and block for responses made in the Pavlovian phase of Experiment 2. ....	90
<i>Figure 6.</i> Mean ratings of the likelihood of O1 occurrence during FX, FC, FN1 and FY; and O2 occurrence during GY, GH, GN2 and GX in the summation test of Experiment 2. ....	92

<i>Figure 7.</i> Mean responses grouped by congruence for each of the compounds in the PIT test of Experiment 2. Top panel: PIT scores. Bottom panel: responses during the preCS period. ....	94
<i>Figure 8.</i> Mean scores grouped by CS and block in the Pavlovian phase of Experiment 3. ....	100
<i>Figure 9.</i> Mean ratings of the likelihood of the occurrence of O1 during FX, FC, FN1 and FY, and of O2 during GY, GH, GN2 and GX in the summation test of Experiment 3. ....	101
<i>Figure 10.</i> PIT scores in the Test A and B of Experiment 3. Top panel: FC/GH, FX/GY and FN1/GN2 in Test A. Bottom panel: F/G, FY/GX and X/Y in Test B. ....	104
<i>Figure 11.</i> Mean rates of congruent and incongruent responding for each of the compounds in the pre-CS period of Test A and B of Experiment 3. Top panel: responses during the pre-CS of Test A. Bottom panel: responses during the pre-CS of Test B. ....	105
<i>Figure 12.</i> Mean scores grouped by CS and block in the Pavlovian phase of Experiment 4. Top panel: the mean scores to the CSs in the inhibitory training. Bottom panel: the mean scores to the CSs in the test excitor training. ....	123
<i>Figure 13.</i> Mean ratings of the likelihood of O1 occurrence during FC and FX, and O2 occurrence during GH and GX in the summation test of Experiment 4. ....	124
<i>Figure 14.</i> Mean rate of congruent and incongruent responses for each type of CS in the PIT Test of Experiment 4. Top panel: PIT scores for FC and FX. Bottom panel: PIT scores for GH and GX....	126

*Figure 15.* Mean scores grouped by CS and block in the Pavlovian phase of Experiment 5. Top panel: the mean scores to the CSs in the inhibitory training. Bottom panel: the mean scores to the CSs in the test excitor training. .... 131

*Figure 16.* Mean ratings of the likelihood of O1 occurrence during FC and FX, and O2 occurrence during GC and GX in the summation test of Experiment 5. .... 132

*Figure 17.* Mean rate of congruent and incongruent responses for each type of CS in the PIT Test of Experiment 5. Top panel: PIT scores for FC and FX. Bottom panel: PIT scores for GC and GX.... 134

*Figure 18.* Mean scores grouped by CS and block in the Pavlovian phase of Experiment 6. Top panel: the mean scores to the CSs in the test excitor training. Bottom panel: the mean scores to the CSs in the inhibitory training. .... 141

*Figure 19.* Mean ratings of the likelihood of O1 occurrence during FC and FX, and O2 occurrence during GC and GX in the summation test of Experiment 6. .... 142

*Figure 20.* Mean rate of congruent and incongruent responses for each type of CS in the PIT Test of Experiment 6. Top panel: PIT scores for FC and FX. Bottom panel: PIT scores for GC and GX.... 144

*Figure 21.* Mean rate of congruent and incongruent responses for each type of CS in the pre-CS period of the PIT Test of Experiment 6. Top panel: responses during FC and FX. Bottom panel: responses for GC and GX..... 146

<i>Figure 22.</i> Mean scores grouped by CS and block in the Pavlovian phase of Experiment 7. Top panel: the mean scores to the CSs in the inhibitory training. Bottom panel: the mean scores to the CSs in the test excitor training. ....	151
<i>Figure 23.</i> Mean ratings of the likelihood of O1 occurrence during FC and FX, and O2 occurrence during GC and GX in the summation test of Experiment 7. Top panel: ratings before participants' exclusion. Bottom panel: ratings after participants' exclusion.....	153
<i>Figure 24.</i> Mean response rate for X and C in the PIT Test of Experiment 7. ....	154
<i>Figure 25.</i> Mean scores grouped by CS and block in the Pavlovian phase of Experiment 8. Top panel: the mean scores to A, AB, CH and DX. Bottom panel: the mean scores to F and G. ....	159
<i>Figure 26.</i> Mean rate of congruent and incongruent responses for each type of CS in the PIT Test of Experiment 8. Top panel: PIT scores for FC and FX. Bottom panel: PIT scores for GC and GX....	161
<i>Figure 27.</i> Mean rate of congruent and incongruent responses for forward and backward trials in the PIT Test of Experiment 9. ....	186
<i>Figure 28.</i> Mean rate of congruent and incongruent responses for forward and backward trials in the PIT Test of Experiment 10. ....	190
<i>Figure 29.</i> Mean rating scores grouped by type of trial (forward; backward) in the Pavlovian tests of Experiment 11. ....	193
<i>Figure 30.</i> Mean rate of congruent (c) and incongruent (i) responses for forward and backward trials in the PIT Test of Experiment 11. ...	194

*Figure 31.* Responses in the forward and backward trials of the Pavlovian phase in Experiment 12. Top panel: the mean percentage of correct responses. Bottom panel: the mean reaction times..... 203

*Figure 32.* Mean rate of congruent and incongruent responses for forward and backward trials in the PIT Test of Experiment 12. .... 204

*Figure 33.* Mean percentage of correct and incorrect outcome responses in the assessment questionnaire of Experiment 12..... 206

*Figure 34.* Responses in the forward and backward trials of the Pavlovian phase in Experiment 13. Top panel: the mean percentage of correct responses. Bottom panel: the mean reaction times..... 211

*Figure 35.* Mean rate of congruent and incongruent responses for forward and backward trials in the PIT Test of Experiment 13. .... 212

*Figure 36.* Mean percentage of correct and incorrect outcome responses in the assessment questionnaire of Experiment 13..... 213

*Figure 37.* Responses in the forward and backward trials of the Pavlovian phase in the experimental and control group of Experiment 14. Top panel: the mean percentage of correct responses. Bottom panel: the mean reaction times. .... 219

*Figure 38.* Group mean rates of congruent and incongruent responses for forward and backward trials in the PIT test of Experiment 14..... 220

*Figure 39.* Mean percentage of correct and incorrect outcome responses in the assessment questionnaire of the experimental and control group of Experiment 14..... 221

## List of Tables

<i>Table 1.</i> Design of Experiment 1. ....	67
<i>Table 2.</i> Mean number of R1 and R2 responses and mean number of O1 and O2 deliveries in the instrumental phase of Experiments 1, 2 and 3. ....	77
<i>Table 3.</i> Design of Experiment 2. ....	86
<i>Table 4.</i> Design of Experiment 3. ....	98
<i>Table 5.</i> Design of Experiment 4. ....	118
<i>Table 6.</i> Mean number of R1 and R2 responses and mean number of O1 and O2 deliveries in the instrumental phases of Experiments 4, 5, 6 and 8. ....	125
<i>Table 7.</i> Mean preCS response rates in each block of the PIT test of Experiment 4. ....	127
<i>Table 8.</i> Mean preCS response rates in each block of the PIT test of Experiments 5, 6 and 8.....	135
<i>Table 9.</i> Design of Experiment 6. ....	138
<i>Table 10.</i> Design of Experiment 7. ....	149
<i>Table 11.</i> Design of Experiment 8. ....	157
<i>Table 12.</i> Design of Experiment 9. ....	183
<i>Table 13.</i> Mean number of R1 and R2 responses and mean number of O1 and O2 deliveries in the instrumental phase of Experiments 9, 10, 11, 12 and 13. ....	185
<i>Table 14.</i> Mean preCS response rates in each block of the PIT test of Experiments 9, 10, 11, 12 and 13.....	187

*Table 15.* Design of Experiment 14. .... 216

*Table 16.* Mean number of R1 and R2 responses and mean number of O1 and O2 deliveries for the experimental and control group in the instrumental phase of Experiment 14. .... 218

*Table 17.* Group mean preCS response rates in each block of the PIT test of Experiment 14..... 221

# Chapter I

## General introduction

### 1.1 Pavlovian and instrumental conditioning

In simple terms, Pavlovian conditioning is the process by which an organism learns the predictive relationship between two stimuli, usually a neutral cue (conditioned stimulus or CS) and a motivationally relevant event (unconditioned stimulus or US) (e.g. Rescorla, 1988; Wasserman & Miller, 1997). When the CS predicts deliveries of the US it becomes excitatory (CS+), eliciting conditioned responses (CRs) that prepare the animal to receive the US. In instrumental conditioning, an animal learns the relationship between its behaviour (R) and the consequence or outcome (O) of this behaviour (e.g. Dickinson & Balleine, 1994). When a response is followed by a reinforcer (appetitive O), subjects will be more likely to perform the same response in the future.

Research in associative learning has shown that Pavlovian and instrumental conditioning interact. For instance, presentations of a CS+ can affect instrumental performance by either invigorating or suppressing behaviour. This phenomenon has been extensively studied by using a procedure known as *Pavlovian-to-instrumental transfer* (PIT), in which the Pavlovian and instrumental relations are

trained independently, and the effect of CSs on instrumental behaviour is measured.

## **1.2 Pavlovian-to-instrumental transfer (PIT)**

In a standard PIT procedure subjects receive Pavlovian conditioning, in which different CSs (e.g. CS<sub>1</sub>, CS<sub>2</sub>) are paired with delivery of different outcomes (e.g. O<sub>1</sub>, O<sub>2</sub>; i.e. CS<sub>1</sub>->O<sub>1</sub>, CS<sub>2</sub>->O<sub>2</sub>). These outcomes may be presented either during, or on termination of, the CS presentations. In a separate instrumental conditioning phase subjects are trained to perform one or more responses (e.g. R<sub>1</sub>, R<sub>2</sub>), and these responses are reinforced with the outcomes (e.g. R<sub>1</sub>->O<sub>1</sub>, R<sub>2</sub>->O<sub>2</sub>). Then in the PIT test subjects have the chance to perform the instrumental responses again. This test is usually conducted in extinction, i.e. no outcomes are delivered, and the CSs are presented while instrumental performance is measured. The results reported using this procedure seem to be unaffected by the order of the phases (Holmes, Marchand & Coutureau, 2010).

The use of this task has consistently found that CS+s affect instrumental responding, a phenomenon known as the *PIT effect*. Estes (1943) was the first to report direct evidence of the effect of a CS+ on instrumental behaviour using appetitive rewards. In his experiment, two groups of rats were trained to press a lever (R) to obtain a food delivery (O). Then the experimental group received presentations of a tone (CS) followed by food (O). At test, both groups

had the opportunity to perform R in the absence of O deliveries, but only the experimental group received presentations of the CS. The CS presentations increased the rate of responding, which was taken as evidence of the invigorating effect of the Pavlovian stimulus on instrumental behaviour.

### **1.3 The specific and general PIT effect**

Research over the years has found that the PIT effect can take two distinct forms. In the *specific* form a CS+ will elevate performance of a response trained with the same outcome as the CS+ more than a response trained with a different outcome. For instance, if one response is reinforced with food pellets ( $R_1 \rightarrow O_1$ ) and another with sucrose solution ( $R_2 \rightarrow O_2$ ), a CS+ paired with food ( $CS_1 \rightarrow O_1$ ) will increase performance of the response reinforced with food more than the response reinforced with sucrose ( $CS_1: R_1 > R_2$ ), and a CS+ paired with sucrose ( $CS_2 \rightarrow O_2$ ) will increase performance of the response trained with sucrose more than the response trained with food ( $CS_2: R_2 > R_1$ ) (e.g. Kruse, Overmier, Konz & Rokke, 1983). But in the *non-specific* or *general* form a CS+ will elevate instrumental responding even if it signals an outcome different from the instrumental reinforcer, as long as the outcome predicted by the CS+ and the reinforcer are of the same motivational valence (e.g. both appetitive). For instance, if a CS is paired with an outcome different to that used in the instrumental training (e.g.  $CS_3 \rightarrow O_3$ ), this CS will elevate performance of both  $R_1$

and  $R_2$  similarly (e.g. Corbit, Janak & Balleine, 2007; Nadler, Delgado & Delamater, 2011).

It has been proposed that the specific and general forms of the PIT effect are determined by different contributions of the CSs to performance. The general PIT reflects the CSs increasing subjects' arousal, indiscriminately facilitating instrumental behaviour, but in the specific PIT effect the role of the CS is to signal the outcome delivery, providing detailed information about this outcome which facilitates performance of the responses reinforced with that outcome (e.g. Dickinson & Balleine, 2002).

Regardless of the role of the CS in the general and specific PIT effect, the idea that both forms of PIT are determined by different mechanisms has found support in research on the neural mechanisms of the PIT effect (e.g. Blundell, Hall & Killcross, 2001; Corbit & Balleine, 2005; Holland & Gallagher, 2003). For instance, Corbit and Balleine (2005) assessed the contribution of different components of the amygdala to the PIT effect by conducting an experiment in which both forms could be detected in the same task. Rats were initially divided in three groups. One of them received lesions in the basolateral amygdala (BLA), another in the amygdala central nucleus (CN) and a third control group received sham surgery. After this procedure, all subjects were trained to perform two responses, each of them reinforced with a different outcome ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ), and then received training in which each of two CSs was paired with one of these outcomes ( $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ), and a third CS with a novel

outcome (CS<sub>3</sub>->O<sub>3</sub>). In the PIT test subjects could perform both instrumental responses while each of the CS+s was presented. In the control group presentations of CS<sub>1</sub> increased R<sub>1</sub> and CS<sub>2</sub> increased R<sub>2</sub>, i.e. specific PIT, while presentations of CS<sub>3</sub> increased both R<sub>1</sub> and R<sub>2</sub>, i.e. general PIT. In the group with BLA lesions the specific PIT effect was abolished (neither CS<sub>1</sub> nor CS<sub>2</sub> elevated responding) while the general PIT effect remained intact (CS<sub>3</sub> elevated both R<sub>1</sub> and R<sub>2</sub>). The reverse was found in the group with CN lesions: they found no evidence of the general PIT effect (CS<sub>3</sub> failed to elevate responding) but the same specific PIT effect as the control group.

Similar results have been found in humans using an fMRI technique (Prévost, Liljeholm, Tyszka & O'Doherty, 2012). Prévost and colleagues conducted an experiment in which participants were trained to perform three responses, each of them reinforced with a different food outcome (R<sub>1</sub>->O<sub>1</sub>; R<sub>2</sub>->O<sub>2</sub>; R<sub>3</sub>->O<sub>3</sub>). Then three CSs were each paired with one of these outcomes and a fourth CS was nonreinforced (CS<sub>1</sub>->O<sub>1</sub>; CS<sub>2</sub>->O<sub>2</sub>; CS<sub>3</sub>->O<sub>3</sub>; CS<sub>4</sub>-). In the PIT test participants could perform two of the three trained responses while each of the CSs was presented. The results of the test showed evidence of the specific PIT effect, as CS<sub>1</sub> and CS<sub>2</sub> selectively increased R<sub>1</sub> and R<sub>2</sub> performance respectively; there was also evidence for the general effect, in that CS<sub>3</sub> elevated performance of both R<sub>1</sub> and R<sub>2</sub> more than CS<sub>4</sub>. Importantly, Prévost et al. (2012) found a correlation between the size of the specific PIT effect and the activity

in the basolateral amygdala region, while the general PIT effect correlated with activity in the centromedial nucleus of the amygdala.

Taken together, these results strongly suggest that the general and specific forms of the PIT effect are distinct processes with a different neural basis (Corbit & Balleine, 2011; Corbit, Janak & Balleine, 2007). This makes it possible to dissociate the contributions of the Pavlovian stimuli to instrumental performance and to study each of the forms of the PIT effect separately. The present thesis focusses on one of these forms: the specific PIT effect.

#### **1.4 The importance of the specific PIT effect**

Both human and animal studies have contributed to the understanding of Pavlovian and instrumental conditioning, and the underlying processes seem to be fundamentally the same (e.g. Garcia, Garcia y Robertson, 1985; Moore, 2004; Shank, 1994). Thus it is no surprise that specific PIT has not only been consistently found in the animal literature (e.g., Baxter & Zamble, 1982; Lovibond, 1983; Rescorla, 1994a); in the last decade research on this issue has been extended to human studies, and many examples of this effect have been reported (e.g., Geurts, Huys, den Ouden & Cools, 2013; Hogarth & Chase, 2011; Lewis, Niznikiewicz, Delamater & Delgado, 2013; Nadler, Delgado & Delamater, 2011; Prevost, Liljeholm, Tyszka & O'Doherty, 2012; Watson, Wiers & de Wit, 2014). In addition to the importance of understanding how these fundamental learning processes interact, research into PIT may also have clinical relevance.

It has been found that CS+s increase subjects' responding to obtain food or drugs, even if these outcomes are no longer desirable. Watson and colleagues (2014) used a PIT task in which they trained participants to perform one response to obtain chocolate (e.g.  $R_1 \rightarrow O_1$ ) and a different one to obtain popcorn (e.g.  $R_2 \rightarrow O_2$ ). Then four different images were either paired with one of these outcomes ( $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ), a third novel outcome ( $CS_3 \rightarrow O_3$ ) or no outcome delivery ( $CS_4^-$ ). Before the PIT test participants were divided into 3 groups: one had further access to  $O_1$ , another to  $O_2$ , and a third group had no manipulation. In the PIT test,  $CS_1$  and  $CS_2$  presentations selectively increased participants' performance of the response associated with the same outcome ( $CS_1$ :  $R_1 > R_2$ ;  $CS_2$ :  $R_2 > R_1$ ) -- the specific PIT effect -- while  $CS_3$  presentations increased performance of both instrumental responses more than  $CS_4^-$  -- the general PIT effect. Importantly, these effects were still found in those participants that were satiated with one of the outcomes. Based on these results, Watson and colleagues suggested that the specific PIT effect may be one of the contributors to obesity and binge-eating disorders (see Lovibond & Colagiuri, 2013 and Colagiuri & Lovibond, 2015 for a similar example in the general PIT effect).

In the case of drug abuse, CSs associated with drugs not only elicit craving and contribute to relapse (Everitt, Dickinson & Robbins, 2001; Tiffany, 1990), but they also increase drug-seeking behaviour (Hogarth, 2012; LeBlanc, Ostlund & Maidment, 2012). For instance Hogarth (2012) trained participants to perform one response to obtain

chocolate and a different response to obtain cigarettes. Before the PIT test one of the outcomes was devaluated by giving participants the chance to either consume chocolate or to use a nicotine nasal spray, aiming to reduce participants' desire for each of these outcomes (satiation). Then in the test participants could perform both responses while a chocolate or cigarette image was presented. The authors found that the nicotine picture increased the response reinforced with nicotine, while the picture of chocolate increased the response trained with chocolate, regardless of whether the participants had had previous access to these outcomes. The specific PIT effect has also been tested using ethanol as outcome (Corbit & Janak, 2007; Garbusow et al., 2014; Glasner, Overmier & Balleine, 2005; Krank, 2003; Troisi II, 2006; Martinovic et al., 2014; Milton et al., 2012), as well as cocaine (LeBlanc, Ostlund & Maidment, 2012; Saddoris, Stamatakis & Carelli, 2011) and heroin (Di Ciano & Everitt, 2003).

Although these experiments clearly indicate the clinical relevance of the specific PIT effect, there is still a debate regarding the mechanisms responsible for it. Several accounts have been formulated but none of them can fully explain this phenomenon. These accounts make use of different conceptualizations of the elements in the associative learning process, so before describing them it is necessary to describe both Pavlovian and instrumental conditioning in more detail.

## 1.5 Pavlovian conditioning

A huge amount of research has been conducted to understand the mechanisms underlying Pavlovian conditioning (Rescorla, 1988; Wasserman & Miller, 1997). One of the most important questions refers to what is learned during conditioning and how this learning is expressed in behaviour. It was initially thought that a CS+ elicits the CR through a direct stimulus-response (S-R) association formed during conditioning. An alternative is based on the idea that the US elicits the unconditioned response (UR) via a pre-existent US-UR association. It has been proposed that the US consists of different features, and each of these features elicits different URs. During Pavlovian training new associations are formed between the CS and the internal US representations corresponding to these different features, which in turn elicit a CR similar to the URs elicited by the US (Konorski, 1967). For instance, after Pavlovian conditioning a CS can activate a sensory US representation that encodes specific information about the US and also a more general US representation that reflects the motivational properties of the US category (e.g. appetitive or aversive). Konorski (1967) made this distinction in order to explain the difference between consummatory CRs, which depend on specific attributes of the US, and preparatory CRs, which are more diffuse responses that depend on the motivational category of the US. The first type is thought to reflect a CS activating a sensory US representation, and the second the activation of a motivational US representation.

Both interpretations of Pavlovian conditioning make different predictions that can be tested. For instance, if the CR is elicited by a representation of the US, then manipulations of the US after conditioning should affect responding. But if the CS directly elicits the CR, then post-training manipulations of the US should have no impact on performance. Holland and Rescorla (1975) conducted a series of experiments aiming to discriminate between these two possibilities. Two groups of rats received pairings of a tone and food deliveries (CS-US) and then for the experimental group the US was devalued by either pairing it with induced illness (Experiment 1) or by allowing rats to consume food freely before the test (Experiment 2). Finally at test, both groups received presentations of the CS while the CRs were measured. If the CS elicits the CR directly, the same level of CR should be found in both groups because the CS value is not manipulated. However, the CRs were only reduced in the experimental group, which supports the idea that Pavlovian conditioning is mediated by the US representation. Yet, it has been pointed out that outcome devaluation procedures usually produce small differences between the experimental and control groups, that it consistently found residual responding, and that it seems to be dependent on the experimental parameter (Holland, 2008). Furthermore, it is possible to think that during the outcome devaluation procedure a new association was formed between the now aversive outcome and a representation of the response (S-R), which affected the CRs elicited by the CSs later at

test. Nevertheless, nowadays it is common to conceptualize Pavlovian conditioning as the formation of CS-US associations.

## **1.6 Instrumental conditioning**

Research on instrumental conditioning started with the ideas of Thorndike (1898) on animal intelligence. Thorndike placed cats inside a puzzle box, in which the animals had to press a lever in order to get out of the box. Thorndike noticed that the animals improved their ability to escape through trial and error, which led him to propose the *law of effect*. According to this principle, a link was formed between the contextual cues within the box (S) and the response required to solve the puzzle (R), a link that was strengthened by the satisfactory outcome (O) of escaping. This stimulus-response (S-R) mechanism has the virtue of explaining behaviour in simple terms by stating that the mere presentation of contextual stimuli elicits responses.

The main characteristic of this account is that the outcome serves only as a reinforcer of the S-R link and is not encoded in the associative chain. But an alternative conceptualization of instrumental conditioning assumes that an association is formed between the response and a representation of the outcome (R->O) (Adams & Dickinson, 1981; Balleine & Dickinson, 1998; Colwill & Rescorla, 1986; Mackintosh & Dickinson, 1979). As in Pavlovian conditioning, one way to assess this idea is to manipulate the value of the outcome after training and to test if this affects instrumental performance. If instrumental responding depends on a S-R association, then

modifying the outcome value after training should have less effect on subjects' performance than if this performance depends on an R-O association. This idea has been tested in several studies; for example Colwill and Rescorla (1985) trained rats to perform two responses, each of them reinforced either with a sucrose solution or a food pellet ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). In the first two experiments, after this instrumental training one of the outcomes (e.g.  $O_1$ ) was devalued by pairing it with an injection of lithium chloride (LiCl), and then the rats had the chance to perform both responses in extinction. Although in this test subjects still performed  $R_1$  that was trained with the devalued outcome, this performance was greatly reduced compared to the performance of the other response,  $R_2$ . These results support the idea that responding occurs to obtain the outcome that is encoded in the R-O association formed in training, so when the outcome loses its value responding declines.

However, the fact that outcome devaluation does not completely abolish performance has led some researchers to suggest that not only are R-O associations formed during conditioning, but also outcome-response ( $O \rightarrow R$ ) associations. Although this distinction might seem unnecessary, it has been proposed that both types of association encode different information about the outcome and affect instrumental performance in different manners. Furthermore, in standard instrumental conditioning only one response is reinforced at a time, so the only events occurring are the response and the outcome delivery. This ensures that the response is not only followed but also

preceded by the outcome delivery, potentially resulting in both R-O and O-R associations. Balleine and Ostlund (2007) proposed that the R-O associations encode information about the sensory and motivational properties of the outcome, but the O-R associations only the sensory aspects. They also argued that activation of the outcome representation initially has a signalling role, similar to the role of the stimulus in the S-R account proposed by Thorndike (1989), directly eliciting the response in the O-R association. However, the degree to which this response is performed depends on the value of the outcome, which is encoded in the R-O association.

Ostlund and Balleine (2007) explored the possible contribution of both types of association by using a phenomenon known as *reinstatement*. In instrumental extinction subjects are allowed to perform a previously reinforced response in the absence of reinforcement, which results in a significant decrease in performance. But it has been found that non-contingent deliveries of the training outcome after extinction produces a response recovery, termed reinstatement (Bouton & Bolles, 1979; Rescorla & Heth, 1975). In one of the experiments reported by Ostlund and Balleine (2007), two groups of rats were trained to perform two responses, each of them reinforced by either a food pellet or a sucrose solution. For the *congruent* group each response was preceded and followed by the same outcome ( $O_1: R_1 \rightarrow O_1; O_2: R_2 \rightarrow O_2$ ) but for the *incongruent* group the responses were preceded and followed by different outcomes ( $O_2: R_1 \rightarrow O_1; O_1: R_2 \rightarrow O_2$ ). After this, all subjects received an extinction

session ( $R_1$ -;  $R_2$ -) and at the end of this session one of the outcomes was delivered (e.g.  $O_1$ ) to produce reinstatement. The question of interest was if the outcome delivery would reinstate the response encoded in the R-O or the O-R association in which that outcome appeared. In the instrumental training of the congruent group,  $O_1$  preceded and followed  $R_1$ , resulting in both  $R_1$ - $O_1$  and  $O_1$ - $R_1$  associations. Thus, presentations of  $O_1$  should result in the recovery of  $R_1$  performance. But in the incongruent group  $O_1$  served as a reinforcer of  $R_1$  but as an antecedent of  $R_2$ , resulting in an  $R_1$ - $O_1$  and an  $O_1$ - $R_2$  association. If instrumental performance is solely dependent on an R-O link then  $O_1$  should also reinstate  $R_1$  in the incongruent group ( $R_1$ - $O_1$ ), but if it is the O-R association that is responsible for action selection, then  $O_1$  should reinstate  $R_2$  ( $O_1$ - $R_2$ ). The results favoured this latter suggestion:  $O_1$  reinstated  $R_1$  in the congruent group but  $R_2$  in the incongruent group. In the same experiment and after the extinction session, one of the outcomes was devalued by giving the subjects free access to it (e.g.  $O_1$ ). After this the instrumental responses were measured again. If the R-O but not the O-R associations encode the motivational value of the outcome, then devaluing  $O_1$  should result in a reduction of the response previously followed, not preceded, by that outcome - that is,  $R_1$  in both congruent and incongruent groups. This is exactly what Ostlund and Balleine (2007) found and is consistent with previous results of outcome devaluation studies described above (e.g. Colwill & Rescorla, 1985). These results were taken as evidence that both R-O and O-R

associations are formed in training and that both contribute to instrumental performance. The action selection might be caused by an O-R association, but how much this response is performed depends on the outcome value encoded in a R-O association.

### **1.7 Theoretical explanations of the specific PIT effect**

Overall, the evidence supports the idea that CSs and instrumental responses become linked with outcome representations, in Pavlovian and instrumental conditioning respectively. The fact that an outcome devaluation procedure reduces performance of CRs and instrumental responses suggests that both Pavlovian (S-O) and instrumental (R-O) associations encode information about the motivational valence of the outcomes. In other words, manipulating the value of an outcome encoded in a Pavlovian association affects the ability of a CS (that has been paired with that outcome) to elicit CRs, and also affects performance of those instrumental responses that have been associated with that outcome. Furthermore, outcome devaluation procedures seem to affect responding selectively, that is, they mostly affect the CRs elicited by a CS that have been paired with the devalued, and responses that have been reinforced with that outcome compared to other responses in the same motivational class. This suggests that S-O and R-O associations also encode specific sensory information about the outcomes. In addition, recent research suggests that instrumental conditioning might also produce O-R associations, in which a sensory representation of the outcome directly

elicits responding. Overall, these conceptualizations of Pavlovian and instrumental conditioning have been used to develop different theories and accounts that may be used to explain the specific PIT effect. In the following section the most influential of these accounts are described.

### **1.8 The two-process theory**

Rescorla and Solomon (1967) published an influential review in which they examined the evidence for Pavlovian and instrumental conditioning and the interaction between them. The evidence led them to conclude that the two types of conditioning are distinct processes, establishing the *two-process theory*. Instrumental conditioning was viewed as resulting in an S-R association, similar to that proposed by Thorndike (1898), in which the stimuli of training and the response become associated because they are followed by a motivationally significant outcome. But these stimuli also enter into a Pavlovian association with the motivational state elicited by the outcomes. Thus, after conditioning, the Pavlovian stimuli will elicit a *central motivational state* in the subjects similar to that of training, contributing the motivational basis for performance.

According to the two-process theory, Pavlovian stimuli will either facilitate or hinder instrumental performance depending on the motivational state they elicit. For instance, stimuli that elicit a positive motivational state, such as that produced by appetitive outcomes, will facilitate performance of those responses also trained with appetitive

outcomes (e.g. Zamble, 1969). But if the Pavlovian stimuli elicit an aversive motivational state, such as fear, they will reduce responses reinforced with appetitive outcomes, e.g. lever pressing to obtain food, but the same stimuli will facilitate responses that were trained to avoid or escape aversive outcomes. As an example, Solomon and Turner (1962) trained dogs in an avoidance task in which they had to press a panel to avoid the delivery of electrical shocks. Subsequently, the dogs were immobilized and they received presentations of two tones, one of them paired with the electrical shock (CS+) but the other not (CS-). In a final test, the dogs received the opportunity to perform the avoidance responses while the tones were presented; in accordance with the theory, the CS+ increased the panel pressing, while few responses were performed during the presentations of the CS-.

The predictions of this theory are consistent with the general form of the PIT effect. Modern authors seem to agree that in general PIT the Pavlovian stimuli elicit a generalised elevation of the subjects' arousal, which results in an increase in performance of instrumental responses paired with outcome(s) of the same motivational modality as that signalled by the CS (Corbit & Balleine, 2005; 2011; Dickinson & Balleine, 2002; Holland, 2004). However, this theory predicts that two or more stimuli paired with outcomes similar in their motivational value but with different sensory properties, e.g. food pellets and sucrose solution, will elicit a similar central motivational state, enhancing performance of those instrumental responses trained with appetitive outcomes. However, the evidence on the specific PIT effect

has shown that a CS predictor of an outcome will produce more responses paired with the same outcome than responses paired with a different outcome, despite the outcomes having the same motivational value (e.g. Kruse, Overmier, Konz & Rokke, 1983). This means that the two-process account, as proposed by Rescorla and Solomon (1967), cannot explain the specific PIT effect.

### **1.9 Expectancy version of the two-process theory**

A few years after Rescorla and Solomon's publication, Trapold and Overmier (1972) proposed an alternative version of the two-process theory, according to which the role of Pavlovian stimuli is to activate an expectancy of the outcome rather than to provide a motivational state to support instrumental performance. In this version, the Pavlovian stimuli become associated with the outcome (S-O) so they can activate a specific representation of O, which in turn produces the response via a stimulus-outcome-response (S-O-R) chain. To test this idea, Trapold and Overmier (1972) trained two groups of rats with pairings of two cues and two outcomes. The *facilitation* group received pairings of CS<sub>1</sub>->O<sub>1</sub> and CS<sub>2</sub>->O<sub>2</sub>, while the *interference* group received presentations of CS<sub>1</sub>->O<sub>2</sub> and CS<sub>2</sub>->O<sub>1</sub>. Subsequently, both groups were trained to perform two responses, each of them reinforced with an outcome in the presence of the Pavlovian cues, that is, CS<sub>1</sub>: R<sub>1</sub>->O<sub>1</sub> and CS<sub>2</sub>: R<sub>2</sub>->O<sub>2</sub>. According to this theory responding can be elicited by the outcomes, e.g. O<sub>1</sub>->R<sub>1</sub>, so if the CSs evoke a representation of the same outcome that the

response is reinforced with, then performance will be enhanced. Consistent with this idea, the results showed that the facilitation group acquired the discrimination faster than the interference group.

According to this S-O-R account, each Pavlovian CS provides detailed information about the specific outcome it predicts. Unlike Rescorla and Solomon's (1967) theory, in which the central motivational state elicited by the CSs was the main factor eliciting responding, this version relies on the CSs eliciting a specific expectation of the outcome, which allows it to successfully predict the specific PIT effect. As an example, in a PIT procedure in which two responses are trained separately, each of them with one outcome ( $O_1$ :  $R_1 \rightarrow O_1$ ;  $O_2$ :  $R_2 \rightarrow O_2$ ), and two CSs are paired with these outcomes ( $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ),  $CS_1$  will activate a representation of  $O_1$  that will produce  $R_1$ , but  $CS_2$  will produce  $R_2$  performance through the activation of  $O_2$ .

As in the original two-process theory (Rescorla & Solomon, 1967), this expectancy version also assumes that the structure underlying instrumental conditioning is an S-R association. That is, this account assumes that a CS elicits the outcome expectancy which directly produces responding. This is similar to the O-R associations described above (Balleine & Ostlund, 2007), which implies that this expectancy does not encode the motivational value of the outcome, but only its sensory aspects. The problem with this is that it is inconsistent with the evidence from outcome devaluation procedures, which indicates that instrumental conditioning results in R-O

associations that encode both sensory and motivational information of the outcomes (e.g. Adams & Dickinson, 1981; Colwill & Rescorla, 1985). However, one way to integrate the modern evidence with the expectancy version of the two-process theory is to assume that both R-O and O-R associations are formed, and that both contribute to the specific PIT effect. An outcome representation evoked by a CS might directly elicit responding through an O-R link (as in an S-R association), but the degree to which this response is performed will depend on the current value of the outcome encoded in an R-O association (Balleine & Ostlund, 2007). An alternative is to propose that R-O associations are bidirectional in nature. In this sense, a CS might activate an outcome representation, which retrieves the response encoded in this association in a backward manner (O $\leftarrow$ R) (Asratyan, 1974; Pavlov, 1932; Rescorla, 1994b).

Both of these adaptations of the S-O-R mechanism, based on the expectancy version of the two-process theory (Trapold & Overmier, 1972), can be used to explain the specific PIT effect. If a CS activates an outcome representation then this can selectively increase responding either via a bidirectional R-O or by an O-R association, as described above.

### **1.10 Stimulus-response (S-R) account**

An alternative explanation of the specific PIT effect has been proposed recently: the stimulus-response (S-R) account (Cohen-Hatton, Haddon, George & Honey, 2013). This account assumes that

both S-O and R-O associations are bidirectional in nature and that during training associations are formed between the Pavlovian stimuli and the instrumental responses, even if they are trained separately. For example, in a PIT task in which the instrumental training is conducted first (e.g.  $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ), during the Pavlovian phase that follows every time an outcome is presented during or after a CS (e.g.  $CS_1 \rightarrow O_1$ ) it activates a representation of the response learned in the instrumental phase, i.e.  $O_1 \rightarrow [R_1]$ . This allows the formation of a link between the CS that precedes the outcome, i.e.  $CS_1$ , and the response representation activated by this outcome, i.e.  $CS_1 \rightarrow R_1$ . Thus the S-R association formed during training provides the CS with the ability to directly elicit the response at test. The S-R account is also capable of explaining the specific PIT effect when the order of the phases is reversed. When Pavlovian conditioning is conducted first (e.g.  $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ), in the instrumental phase (e.g.  $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ) the outcome delivery activates a representation of the CS (e.g.  $O_1 \rightarrow [CS_1]$ ), allowing the formation of an association between the response and the CS representation activated by the outcome (i.e.  $R_1 \rightarrow [CS_1]$ ). Because this account also assumes that the S-R association is bidirectional, then at test S also elicits R directly.

This idea that associations between different events can be established even if one of these events is not present is supported by research on *mediated conditioning*. In this phenomenon, active representations of stimuli or responses can form associations with other events during training (Holland, 1981; Honey & Hall, 1989). For

instance, in one of the experiments reported by Holland (1981), rats first received pairings of two CSs, each of them with a distinct outcome ( $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ) and then in a second phase one of the CSs was paired with an aversive consequence (injections of LiCl) in the absence of the outcomes. After this, the consumption of both outcomes was measured in the subjects' home cages. Although none of the outcomes was directly paired with the injections of LiCl, subjects were less likely to consume the outcome that, in training, was signalled by the CS paired with the aversive consequence. The explanation of these results is that during the second phase the CS evoked a US representation, which became associated with the aversive consequence.

The main difference between the S-R mechanisms and the S-O-R accounts is that the former does not need to appeal to the activation of the outcome representation at the moment of test. This is because according to the S-R account the CSs are responsible for eliciting the instrumental responses directly. In contrast, the activation of the outcome representation at test is crucial for the S-O-R mechanisms. This key feature makes it possible to discriminate between both types of mechanisms of the specific PIT effect. For instance, different strategies can be adopted to manipulate the value of the outcome before the test. If these manipulations affect the specific PIT effect then it is more likely that the activation of the outcome representation is involved in the mechanism underlying the specific PIT effect. In contrast, if manipulating the outcome does not affect the

specific PIT effect, then it is more likely that this effect is governed by a S-R mechanism.

### **1.11 Evidence on the specific PIT effect**

The aim of the next section is to review the literature that provides evidence to either support or refute the accounts of the specific PIT effect described above. I will confine this review to those experiments in which the effect of CSs is assessed in a PIT task, instead of the studies that have either used different tasks or stimuli other than CSs. One of the virtues of the PIT procedure is that it involves training the Pavlovian stimuli in the absence of instrumental responses, which allows us to reject some more complex possible explanations of this phenomenon. Several studies have assessed the effect of discriminative stimuli ( $S^d$ s), which are cues that signal that a response will be reinforced, in instrumental performance. However, these  $S^d$ s are trained differently to CSs; that is, instead of being cues paired with outcomes independently of subjects' behaviour, these stimuli are reinforced or not depending on subjects' performance of the instrumental responses. Because of this, the effect of  $S^d$ s on behaviour is likely to be governed by different rules, and their effect on the specific PIT might be quite different to that of Pavlovian cues, which has found some empirical support.

For instance Colwill and Rescorla (1988) compared the effect of an  $S^d$  with a CS using a PIT task. In one of the experiments reported, in the presence of one stimulus ( $S^d$ ) a response (nose-poking;  $R_0$ ) was

reinforced with an outcome delivery, e.g. food pellet, and a second stimulus (CS) was paired with a different outcome delivery, e.g. a sucrose solution ( $S^d$ :  $R_0 \rightarrow O_1$ ;  $CS \rightarrow O_2$ ). After this, rats were trained to perform two responses (press a lever or pull a chain), each of them reinforced with one of the outcomes presented in the previous phase ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). After this, subjects had access to both the lever and the chain, and the responses were measured while the  $S^d$  and the CS were presented. The results showed that the  $S^d$  increased performance of the response trained with the same outcome,  $R_1$ , but had no effect on the response trained with the other outcome,  $R_2$ . However, the CS produced the opposite effect: that is, it had no impact on the response reinforced with the same outcome,  $R_2$ , but it diminished performance of the response trained with the other outcome,  $R_1$ . One of the explanations proposed by Colwill and Rescorla (1988) was that both stimuli encode specific information about the outcome, but they affect behaviour in different manners. The  $S^d$  helps the response to activate the outcome encoded in the R-O association ( $S^d$ :  $R_1 \rightarrow O_1$ ), increasing the performance of that response (cf. Holman & Mackintosh, 1981). In contrast, the CS directly activates the outcome representation, e.g.  $CS \rightarrow [O_2]$ , and this might interfere with the ability of  $R_1$  to activate  $O_1$ , resulting in a reduction of performance of that response.

The fact that both types of stimulus produced opposite effects on responding is enough to demonstrate that they might be governed by different mechanisms. Because the purpose of this thesis is to

investigate the effect of CSs on the specific PIT effect, those experiments that used  $S^d$ s are not described.

### **1.12 The importance of S-O associations to the specific PIT effect**

According to the S-O-R accounts the specific PIT effect must be determined by the ability of the CS to evoke a representation of the outcome. In this sense it is possible to argue that the strength of the S-O association might affect the degree in which the PIT effect is observed. Studies have used different strategies to manipulate strength of the Pavlovian associations, either during training or after the S-O associations are formed, and then measured the effect of these manipulations on specific PIT. However, it is still not clear to what extent the strength of the S-O associations affect the size of the specific PIT effect.

For instance, Colwill and Motzkin (1994) provided evidence that S-O associations are required to observe the specific PIT effect. In one of the experiments they reported, rats received instrumental training ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ) followed by Pavlovian conditioning ( $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ), but in the Pavlovian phase one of the outcomes was also delivered in the absence of the CSs (e.g.  $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ;  $O_2$ ). This arrangement should have resulted in strong  $CS_1 \rightarrow O_1$  and weak  $CS_2 \rightarrow O_2$  associations, because  $CS_1$  was a reliable predictor of  $O_1$  but  $CS_2$  did not provide more information about  $O_2$  than the context (e.g. Rescorla, 1967). The results of the PIT test, in which the CSs were presented, showed that  $CS_1$  but not  $CS_2$  produced the specific PIT

effect, suggesting that the strength of the S-O association is an important factor for producing the specific PIT effect.

Similarly, Delamater (1995) conducted a series of experiments in which the formation of one of the S-O relationships was interfered with by unsignalled deliveries of one outcome. The first experiment was equivalent to that described above (Colwill & Motzkin, 1994): rats received training of two responses ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ) followed by pairings of two CSs, each of them with one of the outcomes. But in this Pavlovian phase one of the outcomes was also presented in the absence of the CSs (e.g.  $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ;  $O_2$ ). By the end of this phase subjects performed more magazine approaches (CRs) during presentations of  $CS_1$  than in the absence of any stimulus (ITI), confirming that it became a CS+. In contrast, there was no difference in the number of CRs during  $CS_2$  and during the ITI. These results confirmed that the non-contingent presentations of  $O_2$  undermined the formation of the  $CS_2 \rightarrow O_2$  association. After training subjects received two identical PIT tests in which  $CS_1$  and  $CS_2$  were presented while performance of  $R_1$  and  $R_2$  was measured. In these tests  $CS_1$  produced the specific PIT effect, i.e.  $CS_1: R_1 > R_2$ , but not  $CS_2$ , i.e.  $CS_2: R_1 = R_2$ , which is consistent with the results reported by Colwill and Motzkin (1994).

A different approach to assess the contribution of the S-O associations to the specific PIT effect is, first, to allow them to form normally and then manipulate these associations before test. For instance, in a second experiment Delamater (1995) explored the effect

of non-contingent presentations of one outcome *after* the S-O associations were established. The design was identical to the experiment described above (Experiment 1 in Delamater, 1995), except that the Pavlovian phase was divided into three parts: first subjects received 16 sessions in which they received presentations of CS<sub>1</sub>->O<sub>1</sub> and CS<sub>2</sub>->O<sub>2</sub>. After this, rats received 28 similar sessions in which one of the outcomes was also delivered during the ITI (CS<sub>1</sub>->O<sub>1</sub>; CS<sub>2</sub>->O<sub>2</sub>; O<sub>2</sub>), followed by a PIT test. Then rats received 8 additional sessions of the S-O training and unsignalled deliveries of O<sub>2</sub> (identical to previous training phase) before two additional PIT tests were conducted. By the end of the first part of the Pavlovian phase both CSs elicited similar number of CRs, but the inclusion of the unsignalled O<sub>2</sub> presentations reduced the CRs elicited by CS<sub>2</sub> over the course of the phase. As in the first experiment, the second set of PIT tests showed that only CS<sub>1</sub> produced the specific PIT effect, suggesting that the strength of the Pavlovian associations is important for the specific PIT effect to be found.

However, Delamater (1996) found opposite results in a series of experiments in which the strength of the S-O associations was also manipulated in a different way before the test. In all the experiments rats initially received instrumental training (R<sub>1</sub>->O<sub>1</sub>; R<sub>2</sub>->O<sub>2</sub>), followed by Pavlovian conditioning (CS<sub>1</sub>->O<sub>1</sub>; CS<sub>2</sub>->O<sub>2</sub>) and then a PIT test, in which CS<sub>1</sub> and CS<sub>2</sub> were presented. In the first experiment, an extinction procedure was conducted before the test: subjects received presentations of one of the CSs in the absence of the outcomes, e.g.

CS<sub>1</sub>. The results of the PIT test showed that both CS<sub>1</sub> and CS<sub>2</sub> produced a similar specific PIT effect, regardless of the non-reinforced presentations of CS<sub>1</sub> in the extinction phase. The results also showed the CRs (magazine approach) were greatly diminished during extinction, suggesting that the results of the PIT test were not caused by an ineffective extinction procedure. In a second experiment, two groups received additional pairings of both CSs with a novel common outcome (CS<sub>1</sub>->O<sub>3</sub>; CS<sub>2</sub>->O<sub>3</sub>). One group received this training after the Pavlovian phase, i.e. after CS<sub>1</sub>->O<sub>1</sub>; CS<sub>2</sub>->O<sub>2</sub>, and the other group received it before. Conducting this training after the Pavlovian phase was thought to be a stronger extinction procedure than simple extinction because instead of pairing the CSs with the absence of the outcomes they were paired with another US. However, the results of the PIT test showed that the CSs produced a similar PIT effect in both groups. In the third experiment CS<sub>1</sub>->O<sub>1</sub> and CS<sub>2</sub>->O<sub>2</sub> were trained in different sessions. Following the Pavlovian phase, animals were divided into two groups. One of them received identical additional Pavlovian sessions, except that in each of these sessions the corresponding outcome was also presented in the absence of the CS (e.g. CS<sub>1</sub>->O<sub>1</sub>; O<sub>1</sub> and CS<sub>2</sub>->O<sub>2</sub>; O<sub>2</sub>). To ensure that any detrimental effect in this group was caused by the non-contingent presentations of the CSs and the outcomes, another group also received additional sessions but only the outcomes were presented (e.g. O<sub>1</sub>; O<sub>2</sub>). The fourth experiment was identical, except that in the second part of the Pavlovian conditioning (O<sub>1</sub>; CS<sub>1</sub>->O<sub>1</sub>; O<sub>2</sub>; CS<sub>2</sub>->O<sub>2</sub>), the non-

contingent outcomes were not delivered immediately before, after or during the CS presentations. Yet still the results of the PIT tests of all experiments showed that the CSs produced a comparable specific PIT effect in both groups.

It is interesting that these experiments found opposite results to those found by Delamater (1995), considering the similarities between these studies. For instance, the second experiment reported by Delamater (1995) was almost identical to the third reported in 1996. Both of them aimed to weaken the S-O associations by delivering non-contingent outcome presentations after these associations were established. However, in 1995 this procedure eliminated the ability of the CSs to produce the specific PIT effect while in 1996 it did not. One of the differences between these two studies is that in the experiment reported in 1995 subjects received extensive post-training conditioning in which one of the S-O associations was weakened by delivering the corresponding outcome of training. In contrast, in the experiment reported in 1996 subjects had the same number of Pavlovian and post-conditioning training (10 sessions each), and both S-O associations were weakened by delivering  $O_1$  and  $O_2$ . It is possible that the extensive post-training given in Delamater's (1995) study was more effective in disrupting the S-O association than in the studies reported by Delamater (1996), resulting in a reduction of the specific PIT effect in the first but not in the second experiment.

Experiments in humans have also explored the effect of extinction using PIT tasks (Hogarth et al., 2014; Rosas, Paredes-Olay,

Garcia-Gutierrez, Espinosa & Abad, 2010). For instance, Rosas and colleagues (2010) used a cover story to train participants to perform two responses to obtain two outcomes ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ) followed by pairing of each of two CSs with one of the outcomes ( $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ). After this training, extinction was conducted in which one of the CSs was presented without the outcome, e.g.  $CS_1$ . After this, participants could again perform both responses but in the presence and absence of the Pavlovian stimuli. The results of this PIT test showed that both CSs produced a similar specific PIT effect, i.e. increased the response trained with the same outcome as the CS, regardless of whether the S-O association was extinguished or not.

Overall, these results seem to suggest that the formation of an S-O association is critical for the specific PIT to be observed. Manipulations aimed to interfere in the formation of these associations abolished the CS's ability to produce the specific PIT effect (Colwill & Motzkin, 1994; Delamater, 1995), which is consistent with both S-O-R and S-R accounts. For instance, non-contingent outcome deliveries during training should interfere with the formation of the S-O association. According to the S-O-R account the CS activates an outcome representation at test encoded in an S-O associations, thus if this association is not established then the CS cannot produce the specific PIT effect. In the case of the S-R account, the non-contingent outcome presentations should have activated a representation of the response paired with that outcome in the instrumental phase but in the absence of the CS, undermining the formation of the S-R associations.

However, it seems that, once formed, the strength of the S-O associations does not affect the size of the specific PIT effect. When these associations were weakened *after* they were established, the CSs produced the specific PIT effect. Importantly, the size of this effect was not different to that produced by a CS whose association with the outcome was not manipulated (Delamater, 1996). If this is correct, then the results reported by Delamater (1996) suggest that the CSs can produce the specific PIT effect as long as they are associated with an outcome, regardless of the strength of the S-O association.

The S-O-R account cannot easily explain these results. If the specific PIT effect depends on the CSs activating an outcome representation, then extinction of this association should diminish the PIT effect produced by these cues (cf. Cohen-Hatton et al., 2013). In the case of the S-R account, non-contingent outcome presentations after training should have an impact on the ability of the CS to elicit responding. Each time the outcome is presented in the absence of the CS activates a representation of the response trained previously with that outcome, which should weaken the S-R association. Similarly, presenting the CS in an extinction procedure should also weaken the S-R associations. However, if the S-R associations are more resistant to these manipulations, e.g. extinction, than the S-O associations (Cohen-Hatton et al., 2013), then this account could explain how the CS still produces the specific PIT effect after extinction. Otherwise it would also have difficulties in explaining these results.

Delamater and Oakeshott (2007) further explored the idea that the strength of the S-O associations is not critical for the specific PIT effect. Six groups of rats received instrumental training ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ) followed by Pavlovian conditioning ( $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ). The groups differed in the amount of Pavlovian training that they received (4, 8, 16, 24, 56 or 112 presentations of each S-O pairings). In the PIT test instrumental performance and CRs (magazine approach) were measured in the presence of  $CS_1$  and  $CS_2$ . The findings revealed that the CSs elicited fewer magazine approaches in the groups with less Pavlovian training, indicating different levels of associative strength in each of the groups, but in all the groups these CSs produced a similar specific PIT effect. These results suggest that as soon as the S-O associations are established, even earlier in training, these CSs can produce the specific PIT effect, regardless of the strength of these associations. These findings are problematic for the S-R account because less Pavlovian training implies fewer trials in which the response representation is evoked (by the outcome in the Pavlovian phase) contiguous to the CS. In this sense it is possible to argue that fewer of these trials should produce weaker S-R associations, and thus result in a smaller specific PIT effect.

In summary, these results indicate that relatively small amounts of training are sufficient for the formation of an S-O association, which allows the CSs to produce the specific PIT effect. Moreover, manipulations of these associations, once they are formed, do not seem to affect the size of the specific PIT effect. According to the S-O-

R accounts, it is the ability of the CS to activate the outcome representation that is responsible for the PIT effect; thus the fact that extinction reduces the CRs elicited by the CS but leaves the specific PIT effect intact seems to be in contradiction with these accounts. An alternative interpretation of these results can be made based on the ideas proposed by Konorski (1948; 1967). As was described above, Konorski argued that a CS representation participates in independent associations with both a sensory and a motivational US representation. Activation of the motivational US representation results in preparatory CRs, such as approach responses, but the specific PIT effect must reflect the CS activating a sensory US representation, otherwise it would not affect behaviour selectively (Delamater, 2007). If these associations are independent, then it is possible to argue that extinction might affect them differentially (Delamater, 2004). If the association between the CS and a sensory US representation is relatively resistant to extinction compared to that with a motivational representation, then this would explain how extinction reduced the CRs (magazine approach) but did not affect the specific PIT effect in the experiments reported by Delamater (1996). However, there is not enough evidence to confirm that extinction has a differential effect on the associations between the CS and the different representations of the US. Without further assumptions, the results described in this section seem to favour the S-R account, which can explain the results of the extinction experiments but only by assuming that the S-R associations are more resistant to extinction than the S-O

associations. However, this account cannot explain how a specific PIT effect is observed after only a small amount of Pavlovian training because this should also undermine the formation of S-R associations.

### **1.13 The strength of the R-O associations in the specific PIT effect**

For the S-O-R accounts, the R-O associations are as important as the S-O associations described above. In this sense, it might be argued that an outcome representation, activated by a CS at test, should elicit responding more easily when the R-O association is strong rather than weak. If this is correct, then procedures that aim to interfere with the formation of the R-O associations, e.g. non-contingent outcome deliveries during training, or to weaken these associations after they are established, e.g. extinction, should diminish the size of the specific PIT effect.

One of the problems of this approach is that is not entirely clear that these procedures allow us to discriminate between the S-O-R and the S-R accounts because both of them make similar predictions. For instance, in a typical PIT design ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$  followed by  $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ), non-contingent  $O_1$  deliveries during instrumental training should result in weaker  $R_1-O_1$  than  $R_2-O_2$  associations. However, according to the S-R account, weaker  $R_1-O_1$  associations should affect the ability of  $O_1$  to evoke a representation of  $R_1$  in the Pavlovian phase that follows, resulting in a weaker  $CS_1-R_1$  association relative to the  $CS_2-R_2$  associations. According to both accounts, this procedure

should result in  $CS_1$  producing a smaller specific PIT effect than  $CS_2$ . In the case of weakening an R-O association after it is established, extinction of a response before test, e.g.  $R_1$ , should, in principle, reduce the associative strength between that response and the outcome of training; that is, the  $R_1-O_1$  link should be weaker than  $R_2-O_2$  association. But from the perspective of the S-R account, instrumental extinction should also reduce the associative strength of the S-R association. Each time subjects perform  $R_1$  during the extinction phase it should not only activate a representation of  $O_1$  but also a representation of  $CS_1$  that was encoded in the  $CS_1-R_1$  association formed during training. However, if it is assumed that these S-R associations are more resistant to extinction than the R-O associations, as stated by Cohen-Hatton and colleagues (2013), then instrumental extinction should not affect the ability of the CSs to produce the specific PIT effect.

Nevertheless, the problem with this procedure is the evidence indicating that R-O associations are highly resistant to extinction. In an extensive series of experiments Rescorla (1991, 1992b, 1993a, 1995) established R-O associations and then he manipulated these associations by using either extinction, reinforcing the responses with a different outcome, and presenting non-contingent outcome deliveries. In all these experiments Rescorla found that the original R-O associations remained present after these manipulations. For instance, in one of his experiments Rescorla (1993) trained rats to perform four responses, two of them reinforced with  $O_1$  and the other

two with  $O_2$  ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ;  $R_3 \rightarrow O_1$ ;  $R_4 \rightarrow O_2$ ). After this training, two of the R-O associations were extinguished ( $R_1^-$ ;  $R_2^-$ ), and then one of the outcomes was devalued, e.g.  $O_1$ . Following this, subjects had the chance to perform each pair of responses in extinction ( $R_1$  vs  $R_2$ ;  $R_3$  vs  $R_4$ ). As has been described above, outcome devaluation selectively affects the response that was paired with the devalued outcome; thus it was expected that devaluation of  $O_1$  would reduce performance of  $R_3$  relative to  $R_4$ , which were the responses that were not extinguished. However, if the R-O associations are no longer present after extinction, then this selective effect should not be found in the case of  $R_1$  and  $R_2$ . Contrary to this idea, Rescorla found a selective effect of outcome devaluation in these extinguished responses comparable to that found in the non-extinguished responses, i.e.  $R_1 < R_2$ , suggesting that the R-O associations were still present after extinction.

Although the experiments conducted by Rescorla did not assess the effect of CSs on the performance of extinguished R-O associations, it is not uncommon to conduct an extinction procedure before the PIT test (e.g. Delamater, 1995, 1996; Holmes et al., 2010; Lovibond, 1981). The logic behind this is that the general reduction of performance caused by extinction facilitates the detection of any elevation produced by the CS presentations. The fact that the specific PIT effect has been found in these experiments also suggest that the R-O associations are relatively resistant to extinction.

## **1.14 The importance of the outcome value to the specific PIT effect**

Experiments on Pavlovian and instrumental conditioning have consistently found that reducing the value of an outcome after training reduces performance of the responses previously related to that outcome, leaving the performance of responses trained with different outcomes relatively intact (e.g. Adams & Dickinson, 1981; Colwill & Rescorla, 1985; Dickinson, Nicholas & Adams, 1983; Holland & Rescorla, 1975). This suggests that outcome devaluation necessarily affects a sensory representation of the outcome; otherwise these procedures should result in a generalised reduction of performance. In this sense, the studies that have used outcome devaluation procedures in PIT tasks provide critical evidence about the mechanisms behind the specific PIT effect. Because according to the S-O-R accounts the specific PIT effect occurs only if the CSs activate a sensory-outcome representation, a CS that activates a representation of a devalued outcome should produce a smaller specific PIT effect than a CS that evokes a representation of a still valued outcome. In contrast, the S-R account does not require the outcome to be valuable at test because according to this account responding is elicited directly by the CS.

Different outcome devaluation techniques have been used in PIT tasks, and one of them is to induce satiety before test. For instance, in the first of two experiments reported by Corbit, Janak and Balleine (2007) hungry rats received presentations of three stimuli,

each of them paired with one type of food outcome ( $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ;  $CS_3 \rightarrow O_3$ ). In the following instrumental phase, two of these outcomes served as reinforcers of two instrumental responses ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). After this training the performance of  $R_1$  and  $R_2$  was measured separately in the presence and absence of the CSs. These tests were conducted in the same motivational state that rats were during training (hungry). As expected,  $CS_1$  and  $CS_2$  produced the specific PIT effect, i.e.  $CS_1: R_1 > R_2$ ;  $CS_2: R_2 > R_1$ , while  $CS_3$  produced a general effect by elevating both responses. Then the PIT tests were conducted again but after rats had free access to their maintenance food in their home cages. Corbit and colleagues (2007) argued that because rats were no longer hungry, the value of the outcomes was reduced. According to the S-O-R accounts the CSs might activate a representation of the outcomes, but because these are not desirable no increase in performance should be found. In contrast, the S-R account asserts that, at least initially, the CSs should elicit the response that is encoded in the S-R association. The results of these PIT tests favoured the S-R account:  $CS_1$  and  $CS_2$  produced the specific PIT effect (cf. Cohen-Hatton et al., 2013). The general PIT effect produced by  $CS_3$  in the first tests was absent in the second tests, which is consistent with the idea that it is caused by a motivational effect on subjects.

Another technique of outcome devaluation is to pair an outcome with an aversive consequence immediately before test. For instance, in the second experiment reported by Holland (2004), three groups of

rats were trained with either one or two outcomes in the Pavlovian and/or instrumental phases. The critical group for the present discussion received pairings of two CSs, each of them with one outcome ( $CS_1 \rightarrow O_1$ ;  $CS_2 \rightarrow O_2$ ) and two instrumental responses, each with one of the outcomes ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). After training, a PIT test was conducted in which  $CS_1$  and  $CS_2$  were presented followed by the outcome devaluation procedure. Subjects received pairing of one of the outcomes with injections of LiCl, e.g.  $O_1$ , and then an additional PIT test. According to the S-O-R accounts, in the second test the  $CS_1$  should activate a representation of an  $O_1$  which now evokes an aversive state. This should diminish the specific PIT effect produced by  $CS_1$  compared to that produced by  $CS_2$ , which was paired with an outcome that was not devalued. But because outcome representations do not mediate the specific PIT effect, according to the S-R account,  $CS_1$  should elicit  $R_1$  as effectively as  $CS_2$  elicits  $R_2$ . As in Corbit and colleagues' (2007) experiments, the results of both tests (before and after outcome devaluation) showed that the CSs produced the specific PIT effect, i.e.  $CS_1: R_1 > R_2$  and  $CS_2: R_2 > R_1$ . Although the outcome devaluation reduced performance of response  $R_1$  reinforced with that outcome in the absence of the CSs,  $CS_1$  paired with that outcome still elevated performance of that response, e.g.  $CS_1: R_1 > R_2$ . Furthermore, this specific PIT effect was comparable to that produced by the CS paired with the non-devalued outcome, e.g.  $CS_2: R_2 > R_1$ , consistent with the S-R account.

Experiments with humans have used a slightly different procedure in which participants are explicitly told that one of the outcomes is no longer valuable (Allman, DeLeon, Cataldo & Johnson, 2010; Hogarth & Chase, 2011; Hogarth, Field & Rose, 2013). For example, in one of the experiments reported by Hogarth and Chase (2011), participants were trained to perform two responses to obtain either chocolate or cigarette points ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). Four of each of these points represented a real outcome (one cigarette or one chocolate bar), and the accumulation of these points gave participants access to the outcomes, so they could touch and move the outcomes, but they could not consume them. After this training, one of the outcomes was devalued by showing the participants a health warning, e.g. smoking kills or chocolate induces obesity. A devaluation test was conducted next to assess if the outcome devaluation procedure was effective. In this test participants could perform  $R_1$  and  $R_2$  in the absence of any cue, followed by a PIT test in which chocolate and cigarette pictures were presented as CSs while participants performed both responses. The results of the devaluation test showed lower performance of the response previously trained with the devalued outcome, confirming that outcome devaluation affects instrumental performance selectively. Yet in the PIT test, although responding was generally reduced, each of the CSs produced the specific PIT effect, regardless of whether the outcome was devalued or not; i.e. chocolate and cigarette pictures elevated performance of the response trained with chocolate and cigarette respectively. These results are difficult to

explain for the S-O-R accounts. If devaluation of an outcome reduces performance of the response paired with that outcome, then a CS that activates that outcome representation should not produce the specific PIT effect. But these results can be explained by the S-R account because although instrumental performance is determined by the value of the outcome encoded in the R-O association, in the PIT test it is the CS that directly elicits responding.

However there is one study that reported evidence of a reduction in the specific PIT when one of the outcomes was devalued before the PIT test. Allman and colleagues (2010) used a stock market paradigm in which participants learned the relationship between symbols (CSs) and different currencies (outcomes). In the Pavlovian phase three CSs ( $CS_1$ ,  $CS_2$  and  $CS_3$ ) were paired with one outcome ( $O_1$ ), another three ( $CS_4$ ,  $CS_5$  and  $CS_6$ ) with a different outcome ( $O_2$ ), and two more ( $CS_7$  and  $CS_8$ ) were non-reinforced, (i.e.  $CS_{123} \rightarrow O_1$ ;  $CS_{456} \rightarrow O_2$ ;  $CS_{78}^-$ ). In the instrumental phase two responses were trained, each of them with one of the outcomes, and a third response was non-reinforced ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ,  $R_3^-$ ). After this training participants received a PIT test in which they could perform any of the instrumental responses while the CSs were presented, and following this one of the outcomes was devalued by informing the participants that one of the currencies had lost its value in the market. Then an additional PIT test was conducted. The results of the first PIT test showed that  $CS_{123}$  and  $CS_{456}$  produced the specific PIT effect, i.e. the CS+s elevated performance of the response trained with the same

outcome that they signalled more than the response trained with a different outcome. But in the second test only the CS paired with the non-devalued outcome produced this effect. The CSs that signalled the devalued outcome did not elevate performance of the response trained with that outcome more than the response trained with a different outcome.

These results are in contradiction with the rest of the literature, but there are several procedural differences that make a direct comparison difficult. One possibility is that this procedure did not reduce the outcome value but directly instructed participants to ignore one of the outcomes. In this sense, it might be that instructions given to the participants are responsible for the different results. This non-associative explanation is supported by one of the experiments reported by Hogarth et al. (2014). In that study two groups of participants were trained to perform two responses to obtain either beer or chocolate rewards ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ) and then in the PIT test participants could perform both responses while pictures of beer and chocolate were presented. Critically the experimental group received additional instructions before the PIT test indicating to them that the beer and chocolate pictures did not provide information about which response was being reinforced (in capital letters). A control group did not receive these instructions. As expected, the control group showed the specific PIT effect - i.e. beer images elevated the response reinforced with beer rewards and chocolate images increased the response reinforced with chocolate -- but these images produced no

effect in the experimental group. This evidence supports the idea that in humans direct instructions can reduce the specific PIT effect, but this cannot be replicated in animal research.

Aside from the report of Allman and colleagues (2010) the literature is consistent in suggesting that reducing the value of the outcome after training, either by pairing with an aversive consequence or through satiety, does not affect the specific PIT effect. As was described above, the S-O-R accounts cannot easily explain these results because it is a representation of the outcome that mediates the specific PIT effect. If this outcome loses its value before the test, then it should lose its ability to elicit instrumental responding. But according to the S-R account it is not an outcome representation but rather the CSs that are responsible for the specific PIT effect. This S-R association does not encode any outcome value so this account has no problem in explaining the results of outcome devaluation procedures.

Nevertheless, these results can also be explained by the S-O-R accounts by assuming a mechanism similar to that of the S-R account. The S-R account states that during training an association is formed between the CS and the response, so at test the CS presentations directly elicit responding. Analogously, it is possible that during training an association is formed between a *sensory*-outcome representation and the instrumental response, and that this association is independent of other possible associations that involve the motivational value of the outcome. Then in the PIT test the CSs

retrieve this sensory-outcome representation that *directly* elicits the responses encoded in this R-O association. Here the effect of the outcome representation on performance is not mediated by the motivational properties of the outcome; instead this representation works as a stimulus trained in an S-R association in which the outcome value is not encoded (Balleine & Ostlund, 2007).

The main problem with this interpretation is that when performance is measured in the absence of the CSs, as in an extinction test, outcome devaluation reduces behaviour. Moreover, this reduction affects responding *selectively*, suggesting that the degree to which a response is performed is determined by the current value of the outcome encoded in a specific R-O association. Then, it is necessary to further assume that instrumental performance in the presence and absence of the CSs (PIT and extinction tests, respectively) is governed by different rules.

This idea is supported by some evidence, indicating that instrumental performance in extinction and PIT tests is determined by dissociable mechanisms (de Borchgrave, Rawlins, Dickinson & Balleine, 2002; Dickinson, Smith & Mirenowicz, 2000). For instance, Corbit, Muir and Balleine (2001) explored the effect of lesions in the shell and core of the nucleus accumbens (NAc) on performance using rats as subjects. They found that shell-lesioned rats showed the selective effect of outcome devaluation on performance when it was measured in extinction, e.g.  $R_1 < R_2$  when  $O_1$  was devalued. However, these animals showed no evidence of selectivity when a CS paired

with the devalued outcome was presented in a PIT test, e.g.  $R_1=R_2$  when  $CS_1$  was presented. In contrast, core-lesioned subjects showed the opposite pattern: no selective effect of outcome devaluation in an extinction test by an evident specific PIT effect. This evidence suggests that different neural structures determined different aspects of the expression of instrumental associations.

Overall the research described here suggests that devaluing an outcome reduces performance of a response trained with that outcome. However, a CS predictor of a devalued outcome can still elevate performance of a response trained with that outcome. The S-R account can explain these results based on the central idea that the outcome representation does not mediate the PIT effect. Thus, manipulating the value of the outcome after training should not affect the specific PIT effect. In contrast, these results cannot be explained by the S-O-R accounts unless further assumptions are made. It must be assumed that the expression of the R-O associations in an extinction test depends strongly on the motivational value of the outcome, while the effect produced by the CSs on performance in a PIT test is mainly determined by a sensory outcome representation.

### **1.15 The importance of S-R associations in the specific PIT effect**

Cohen-Hatton and colleagues (2013) conducted a series of experiment to assess the idea that S-R associations are also formed during training, and that these associations contribute to the specific PIT effect. In one of their experiments rats were trained to press two

levers, one of them reinforced with food pellets and the other with a sucrose solution ( $R_1 \rightarrow O_1$  and  $R_2 \rightarrow O_2$ ) and then received pairings of two CSs with each of the outcomes, either in a forward ( $CS_1 \rightarrow O_1$ ) or backward relationship ( $O_2 \rightarrow CS_2$ ). Then in the PIT test half of the subjects had the opportunity to perform one response during  $CS_1$  and  $CS_2$ , and the other half received the same treatment but for the other response. The same experiment was replicated but subjects could perform  $R_1$  and  $R_2$  in the same test. The results of the PIT test in both experiments showed that  $CS_2$  but not  $CS_1$  produced the specific PIT effect.

These results were interpreted in terms of the S-R account proposed by Cohen-Hatton et al. (2013). According to this account, during Pavlovian conditioning each outcome activated a response representation, allowing the formation of S-R associations. Thus in the forward trials an association was formed between  $CS_1$  and  $R_1$  via  $O_1$ , and between  $CS_2$  and  $R_2$  through  $O_2$ . But the contiguity between the CSs and the activation of the response representation was not the same in both types of trials. In the forward trials  $R_1$  was activated when  $CS_1$  was no longer present, but in the backward trials  $R_2$  was activated closer in time to the presentation of  $CS_2$ . The fact that  $CS_2$  and the  $R_2$  representation were temporally closer than  $CS_1$  and  $R_1$  should result in a stronger S-R association in the case of  $CS_2$ - $R_2$  than in  $CS_1$ - $R_1$ . Thus at test  $CS_2$  produced the specific PIT effect by directly eliciting  $R_2$ , and thus more effectively than  $CS_1$  could by eliciting  $R_1$ .

Although the fact that a CS trained in a backward procedure produced the specific PIT effect supports the S-R account, there is also evidence that CSs trained in this manner produce the opposite effect, that is they selectively reduce performance of a response trained with the same outcome as the CSs (Delamater, LoLordo & Sosa, 2003; Laurent, Wong & Balleine, 2014). For instance, in one of the experiments conducted by Delamater et al. (2003) rats first received instrumental training ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ) followed by Pavlovian conditioning, in which each of two CSs was paired with one outcome but in a backward relationship ( $O_1 \rightarrow CS_1$ ;  $O_2 \rightarrow CS_2$ ). Then in the PIT test subjects could perform both responses during the presentations of the CSs. The results of the test showed that the CS presentations increased the response trained with a different outcome more than the response trained with the same outcome that the CSs.

### **1.16 Summary**

The specific PIT effect is a phenomenon that is well established in the associative learning literature and certainly has an important impact in maladaptive behaviour. This makes understanding the mechanisms that underlie this phenomenon important. The literature reviewed here suggests that this effect might be mediated by the activation of an outcome representation at test, i.e. consistent with the S-O-R account; but there is also compelling evidence that it may be caused by a direct effect of the CSs without any mediation of the outcome, i.e. as predicted by the S-R account. However, neither of

these accounts can fully explain the evidence found in the literature without further assumptions.

In the case of the S-O-R mechanism, it is necessary to assume that the specific PIT effect is mediated mainly by the activation of a sensory outcome representation that does not encode the motivational valence of the outcome. Otherwise this account cannot explain the evidence indicating that the specific PIT effect is immune to outcome devaluation procedures (e.g. Corbit, Janak & Balleine, 2007; Holland, 2004). A second assumption is that the association between the CS and the sensory outcome representation is formed early in training, in order to explain the fact that the specific PIT effect is still found with small amounts of Pavlovian conditioning (Delamater & Oakeshott, 2007). In addition, this type of association must be resistant to extinction (Delamater, 1996). Fewer assumptions are needed to explain the data with the S-R account. One of them is that the S-R associations are resistant to extinction relative to the S-O associations. If this is the case this mechanism can successfully explain the results of the experiments employing extinction. However, it must also assume that the S-R associations require small numbers of trials to be formed, otherwise it cannot explain the results of the studies reported by Delamater and Oakeshott (2007). Importantly, this account does not have problems in explaining the results of outcome devaluation procedures.

The aim of this thesis is to provide further evidence on specific PIT, to discriminate between these different accounts of the effect. The

techniques used in the experiments described in the previous sections do not allow us to discriminate between the S-O-R and S-R accounts. For this reason different strategies were adopted to achieve this. In the experiments reported in Chapter 2 different CSs were trained to signal the absence of the outcomes, i.e. conditioned inhibitors, and their effect on instrumental performance was assessed using a PIT task. This strategy is similar to extinction of S-O associations, in the sense that non-reinforced presentations of a CS+ should undermine its ability to evoke a representation of the US. However, while extinction aims to reduce the excitatory association between a CS and an outcome, conditioned inhibition produces a CS that has the opposite effect on behaviour to a CS+. It has been argued that this effect is due to the conditioned inhibitor's ability to suppress the outcome representation (e.g. Rescorla & Holland, 1977). If the activation of an outcome representation by a CS+ at test is critical for the specific PIT effect, then a conditioned inhibitor that suppresses such a representation should reduce specific PIT. The aim of Chapter 3 was to provide additional evidence on the effect of conditioned inhibition in the specific PIT effect, and to try to elucidate if a conditioned inhibitor indeed suppresses a sensory outcome representation or if it affects a more general motivational system associated with the outcomes. In Chapter 4 the idea that backward conditioning, in which a CS is consistently presented *after* an outcome delivery, results in inhibitory conditioning was explored and the effect of a CS trained in a backward procedure was assessed using a PIT task. Although some authors

have found that backward conditioning endows the CSs with excitatory properties (e.g. Burkhardt, 1980; Mahoney & Ayres, 1976), there is compelling evidence suggesting that this procedure results in inhibitory CSs (e.g. Siegel & Domjan, 1971, 1974). Moreover, as was described above, some authors have found that a CS trained in a backward relation with an outcome produces the opposite to the specific PIT effect, reducing rather than elevating performance of a response trained with the same outcome (Delamater et al., 2003; Laurent et al., 2014), which is consistent with the predictions of the S-O-R accounts. However, there is also evidence indicating that these CSs produce a standard specific PIT effect (Cohen-Hatton et al., 2013), which supports the S-R account. It is expected that the experiments reported in this thesis will contribute to a better understanding of the specific PIT effect and the mechanism that underlies this phenomenon.

## Chapter II

### Conditioned inhibition in the specific PIT effect

#### 2.1 Overview

The main difference between the various versions of the *two-process theory* and the *stimulus-response* account is that only according to the former the activation of the outcome representation at the moment of test to explain the specific PIT effect. One logical approach to compare both types of accounts is to manipulate the activation of the outcome representation at test. As was described above, procedures aiming to weaken the S-O associations, e.g. extinction, or to reduce the value of the outcome before the PIT test do not seem to affect specific PIT. However, an alternative is to suppress the activation of the outcome representation by using *conditioned inhibitors*, which are cues that predict the absence of the outcomes. In this chapter three experiments were conducted to assess the effect of conditioned inhibitors on instrumental responding in a specific PIT study. In the instrumental training phase of all the experiments two responses were trained, each with one outcome ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). In Experiment 1 the effect of excitatory cues on responding at test was compared with that of conditioned inhibitors presented alone, and in Experiment 2 excitatory CSs were presented in compound with the conditioned inhibitors. Experiment 3 aimed to replicate the results of both Experiment 1 and 2 in the same task.

## **2.2 Introduction**

The experiments reported in this chapter made use of a conditioned inhibition procedure to compare the predictions of the different PIT accounts. But in order to explain the rationale of these experiments it is necessary to briefly review what conditioned inhibition is, and the mechanisms that are thought to explain this phenomenon.

## **2.3 Conditioned inhibition**

In an excitatory conditioning procedure a CS is trained to consistently predict US presentations. Due to the positive contingency with the US, the CS becomes excitatory (CS+), eliciting conditioned responses (CRs) when presented. But if a stimulus is trained to signal the absence of an expected US, this stimulus becomes a conditioned inhibitor (CI) that produces the opposite tendency to a CS+ (Rescorla, 1969). In experimental settings, a CI is a stimulus trained to predict the absence of a US in circumstances where this outcome would be otherwise delivered. Although different procedures for establishing inhibition have been designed, in all of them the putative CI must be paired with the omission of an expected outcome. For example, in the standard procedure of conditioned inhibition designed by Pavlov (1927), a CS is paired with a US (e.g. A->US) and the same cue is also presented with the putative CI but in the absence of the US (AX->nothing). During training, A becomes a CS+ so that when AX is presented A provides the expectation of the US, which endows X with inhibitory properties (Rescorla, 1969; Williams, Travis and Overmier,

1986; Williams, Overmier and Lolordo, 1992). In a different procedure known as differential inhibition, a CS is paired with the US (A->US) and on different trials X is presented but without the US (X<sup>-</sup>). In this procedure, the expectation of the US is provided by the context (Hearst & Franklin, 1977; Miller, Hallam, Hong & Dufore, 1991).

## **2.4 Measuring conditioned inhibition**

While a CS+ elicits CRs that can be observed directly, this is not always the case for the CIs. For example, a CS+ predictor of food elicits salivation in dogs, but a CI that predicts the absence of food does not reduce salivation below the baseline level (Rescorla, 1969). Although in some cases a CS+ and a CI elicit opposite responses that are relatively easy to detect, e.g. heart rate (Cunningham, Fitzgerald & Francisco, 1977; Hoffman & Fitzgerald, 1982), in most instances it is necessary to conduct further tests. One of them is the *summation test*, in which a CS+ different to the one used to establish the CI is presented alone (e.g. B) and in compound with the CI (e.g. BX). If X is a CI then it should counteract the excitatory properties of B, resulting in fewer CRs in the presence of BX than B alone. But as Rescorla (1969) pointed out, in this version of the test lower responding to BX does not necessarily mean that X acquired inhibitory properties. For instance, BX is a different stimulus to B alone, so lower responding to BX might reflect that B suffers generalisation decrement when presented with B.

However, appropriate controls should solve this apparent problem. For example Pavlov (1927) used a summation test in which dogs received presentations of a CS+ with a CI (e.g. BX) and with a novel stimulus (e.g. BN). Thus lower responding to BX than to BN cannot be explained by generalisation decrement because it should affect both compounds. Although initially the novelty of N might attract subjects' attention, reducing CRs to BN, this reduction should be temporary and smaller than the effect of the CI (Baker, 1977; Mackintosh, 1983).

But Rescorla (1969) also noted that the treatments that a CS+ and a CI receive during training might affect subjects' attention to these stimuli differentially. For instance, in a A+/AX- procedure, presenting X together with A in training might attract subjects' attention to X over A. Then in a summation test in which a CS+ is presented with X and with a novel cue (BX and BN), X might attract more attention than B compared to the attention received by N in the BN compounds, which should result in BX eliciting fewer CRs than BN. For this reason Rescorla (1969) proposed that in addition to the summation test a *retardation test* should also be conducted. In this test a CI is paired with a new US and the rate of acquisition is compared to that of a neutral novel stimulus trained in the same manner. If the CI has inhibitory strength, then it should require more trials to acquire the learning. Thus, when a CI reduces responding to a CS+ relative to a novel cue in a summation test and also requires more time than a neutral stimulus to acquire an excitatory association, then the most

likely explanation is that it has become a CI (Rescorla, 1969). This is because attentional explanations cannot account for both results. For example, if training increases participants' attention to the CI then this should facilitate a new learning instead of retarding it.

There are different alternative approaches to solve these problems. For instance, an alternative is to conduct a summation test in which a CS+ is presented with the CI, e.g. BX, and also with a stimulus that has been trained in a similar manner to the CI. If during training a neutral stimulus (e.g. C) is also presented in non-reinforced trials but without explicitly signalling the absence of an expected US (i.e. A+/AX-/C-), it should attract subjects' attention a similar degree as the CI but without acquiring inhibitory properties. Then in the summation test (e.g. BX and BC) a reduction in the CRs to BX compared to BC can only be explained by the inhibitory strength of X (cf. Rescorla, 1969). However, the experiments reported in this chapter made use of a summation test in which both strategies were used. The inhibitory properties of the CIs were assessed by presenting them together with a CS+, e.g. BX, and their effect on behaviour was compared to that produced by a novel stimulus, e.g. BN, and also by a pre-exposed control cue, e.g. BC.

## **2.5 The nature of conditioned inhibition**

It has been proposed that the effect of the CI on behaviour is due to its ability to suppress the US representation (Rescorla & Holland, 1977). If this is correct then conditioned inhibition can be

useful to discriminate among the accounts of the specific PIT effect. For instance, the S-O-R accounts states that is activation of the US representation that produces responding. However, before assuming that the CI acts on such a representation it is necessary to review the evidence to support this idea.

The mechanism by which a CI reduces CRs is not yet clear. Pavlov (1927) suggested that CS presentations in the absence of the US were enough to endow the CS with inhibitory properties. According to Pavlov (1927) a CI produces a general suppression in behaviour which is the cause of the reduction in the CRs in the summation test. However, this perspective has been described as a non-associative account because it considers inhibition as a property of the CS regardless of its relationship with the US (e.g. LoLordo & Fairless, 1985). In contrast, Rescorla (1969) defined conditioned inhibition as an associative learning process parallel to excitatory conditioning. The effect of a CI on behaviour is opposite to that produce by a CS+ that has been trained with the same outcome, which implies that a CI is specific to the outcome of training.

It has also been proposed that conditioned inhibition is based on a stimulus-response (S-R) mechanism. According to these accounts, the effect of a CS+ is caused by the formation of an *excitatory* link between the CS+ and the CRs during training. In a similar way a CI suppresses responding because of an *inhibitory* association with these CRs. It has also been suggested that in training a CI forms associations with responses antagonistic to those elicited

by the CS+. In the former case a CI acts by suppressing behaviour, while in the latter a CI elicits competing responses, reducing the CRs elicited by the CS+. None of these interpretations consider representations of the US as part of the associative process but, as in excitatory conditioning, the ideas proposed by Konorski (1948, 1967) suggesting that US representations are involved in conditioned inhibition received more attention and empirical support.

Konorski (1948) initially proposed that a CS presentation activates an internal representation or *CS centre*, and the US delivery activates different *US centres*, one that encodes sensory information and another motivational value of the outcome. According to Konorski (1948) conditioning occurs when a link is formed between the CS centre and the rising in the activity of the US centres. If the CS is present while the activity on the US centre is rising, e.g. US presentation, the association between the CS and the US is strengthened. But if the activity on the US centre is decaying, then the CS-US association is weakened (Konorski, 1948; McLaren & Dickinson, 1990). In inhibitory conditioning the CS is presented while subjects have the expectancy of the US, but due to the absence of the US the activity on the US centre falls. This results in an *inhibitory* association between the CS centre and the US centres. Then when this CS is presented it inhibits or suppresses the activation of the US centres, producing the opposite effect on behaviour to a CS+.

Some of the strongest evidence supporting Konorski's ideas comes from the experiments reported by Rescorla and Holland (1977).

They aimed to discriminate between different mechanisms of conditioned inhibition, including Konorski's interpretation that a CI acts on the US representations but also the idea that the CI acts directly on the CRs, as suggested by the S-R account described above.

According to the S-R explanation, in an A+/AX- design an inhibitory association is formed between X and the CRs elicited by A. Thus when X is presented with a different CS+ in a summation test, e.g. BX, X will suppress responding to B but only if A and B elicit the same CRs, otherwise X will have no effect. This prediction differs from the mechanism proposed by Konorski. According to his initial proposal (Konorski, 1948) and using the previous example, during training X activates an X centre in the absence of the US, forming an inhibitory association with the US centres. Then when BX is presented at test, X will inhibit the activation of the US centres otherwise activated by B. Importantly, X will suppress responding to B even if A and B elicit different responses. What is important is that B activates the same US centres as A (which are the same as those that X inhibits).

In order to test these predictions, Rescorla and Holland (1977) used CSs that produce different CRs. For example, a clicker paired with food elicits startle and head jerking responses, while a light also trained with food elicits rearing and magazine behaviour (Holland, 1977). In one of the experiments rats received presentations of a clicker (C) and a light (L), each of them paired with food delivery (US). Simultaneously, one group of rats received non-reinforced presentations of C in compound with a tone (C->US; CT-) and another

group had trials of L with the tone (L->US; LT-). This should result in T becoming a CI in both groups. Then at test all subjects received presentations of C, L, CT and LT while the CRs specific to C and L were measured. If T suppressed the CR of training, as the S-R account states, then C should have elicited more CRs (startle and head jerking) than CT but only in the group trained with CT trials. Similarly, L should have elicited more rearing and magazine behaviour than LT but only in the group trained with LT presentations. However, both types of CRs were reduced by T in all the subjects. After the test, half of the subjects in each group received C-shock pairings, and the other half L-shock trials, and then all the subjects received C, L, CT and LT presentations. Again CT and LT elicited less responding than C and T respectively, but only when the CSs kept their original association with food intact; when the CSs+ were paired with shock they elicited fear responses that were not reduced by T. These results are consistent with Konorski's conceptualization of conditioned inhibition. Taken together these results support the idea that a CI suppresses the activation of the US representation. Moreover, it also suggests that the effect of a CI is specific to the US of training, or at least specific to its motivational value (e.g. appetitive vs. aversive).

## **2.6 Specificity of conditioned inhibition**

Later Konorski (1967) postulated a slightly different version of his account. In this new version, a CI does not directly suppress the activation of the US centres, but it activates *no-US centres* via an

excitatory association. These new centres are equivalent to the US centres (sensory and affective properties) but by nature their activation inhibits that of the US centres. For instance, a CS+ that predicts a shock activates a *shock-centre* that includes specific information about the shock (intensity, duration, etc) and a *fear-centre* responsible for the motivational state produced by the shock. Then a CI trained to predict the absence of this US turns on a *no-shock centre* and a *no-fear centre*, activation of which counteracts activity in the *shock* and *fear centres* respectively. In the case that a CI is trained to predict the absence of a different US (e.g. aversive loud noise), it will activate different US centres, which makes conditioned inhibition outcome-specific. Even if both USs (shock and loud noise) have the same motivational valence (aversive), their sensory aspects will differ. For this reason a CI that signals the absence of a noise will not suppress the CRs elicited by a CS+ predictor of shock, at least not to the same degree.

However, some authors have argued against the idea that conditioned inhibition is US-specific. Considering Konorski's assumptions (1967) it is easy to understand that a CS+ becomes associated with a representation of both sensory and affective properties of the US because of the contiguity between the CS and US. But this is not the same for a similar process between the CI and *no-US centres*. In training the CI is presented in the absence of the US so it is more difficult to conceptualise the CI forming an association with the sensory aspects of the US (e.g. flavour, intensity, etc),

considering that these aspects of the US are never contiguous to the CI. In contrast, the CI is always followed by the motivational state produced by the absence of expected US. A simpler idea then, is that the CI becomes linked to this motivational state, which reflects the motivational properties of the absence of the US (Mackintosh, 1983). In this vein, Dearing and Dickinson (1979) proposed that a CI forms an *excitatory* link with a motivational state that is opposite to that elicited by the US. For instance, a CI that signals the absence of shock will elicit a general motivational state of relief that has an opposite effect to the fear elicited by a shock. Thus a CI does not encode specific information about the omitted US but it elicits a motivational state incompatible with this US. This implies that a CI can suppress responding to a CS+ trained with a different US as long as both USs elicit the same motivational state.

The evidence to discriminate between the two positions is not entirely consistent. For example LoLordo (1967) conducted an experiment with dogs in which all the animals were trained to press a panel to avoid shock (US) deliveries. Then, for half of the subjects a CS+ and a CI were established for shock, while for the other half a CS+ and a CI were trained using an aversive loud noise instead. Then these stimuli were presented again and their effect on avoidance responding was measured. Although shock and noise have different sensory properties, both USs elicited the same motivational state (fear). Thus, if a CI acts by producing an opposite motivational state (e.g. relief) then the CIs should reduce responding even in the group

that was trained to predict the absence of the loud noise. However, both CSs+ increased avoidance but the CI effectively reduced responding only in the group in which predicted the absence of the shock.

But there is also evidence supporting the idea that a CI does not encode sensory information about the US (LoLordo and Fairless, 1985; Nieto, 1984; Pearce, Montgomery and Dickinson, 1981). For example Nieto (1984) used rats as subjects that were initially trained to press a lever to obtain food. Then a CS+ predicting a loud noise (A+) and a CI signalling the absence of the noise (AX-) were established. In parallel sessions, a second CS+ was trained to predict shock deliveries (B\*). At test the rate of responding was measured while subjects received presentations of AX and BX. If conditioned inhibition is specific to the US of training then X should only cancel the suppressive effect of A on responding because both A and X were trained with the same US (shock). But the results revealed that X reduced the effect of A and B similarly, supporting a general effect of inhibition. This sort of evidence strongly suggests that a CI does not encode specific information about the US but it does suppress the motivational state elicited by the CS+, either directly or by activating an opposite motivational system. However, there is no consensus in the literature about the US-specificity of conditioned inhibition and the debate still continues. But if a CI exerts its inhibitory properties based on the specific sensory properties of the US of training, then its effect on behaviour can be studied using a PIT task.

## 2.7 Specific PIT effect and conditioned inhibition

If a CI suppresses a US representation that encodes sensory information then it can be used to discriminate between the different accounts of the specific PIT effect. According to S-O-R accounts, specific PIT is observed at test if the CS+ activates a detailed representation of the US. This US representation selectively increases those responses that also encode information about the same US, either via an R-O or an O-R association (Balleine & Dickinson, 1998; Ostlund & Balleine, 2007). In this sense, the presence of a CI should counteract the ability of the CS+ to activate the US representation, reducing the specific PIT effect. But according to the S-R account, the US only contributes to the formation of a link between the CS+ and the responses during training, and then at test the CS+ produces these responses directly (Cohen-Hatton et al., 2013). Thus the US representation has no direct role in the manifestation of the specific PIT effect so even if a CI suppresses this representation it should not reduce the effect of a CS+ on responding.

Laurent and colleagues (2015) provided evidence on this issue with mice as subjects using a PIT procedure. In one of their experiments animals received presentations of two auditory stimuli, each of them paired with either grain or chocolate pellets (A->O<sub>1</sub>; B->O<sub>2</sub>) and a CI was established for O<sub>1</sub> (C) and another for O<sub>2</sub> (D) by presenting them in non-reinforced trials with the CS+ (AC-; BD-). In the instrumental training phase two responses were reinforced, each

of them with one of the outcomes ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). In the PIT test subjects could perform both responses while A and B were presented in compound with the CIs that either signalled the absence of the same outcome as the CS+ (AC; BD) or an alternative outcome (AD; BC). Because A and  $R_1$  were paired with  $O_1$ , and B and  $R_2$  with  $O_2$ ,  $R_1$  should be performed more than  $R_2$  in the presence of A and the reverse for B, i.e. the specific PIT effect. This is exactly what the authors found in the AD and BC trials. However, they found the opposite for AC and BD: AC produced more  $R_2$  than  $R_1$  responses and BD more  $R_1$  than  $R_2$  responses, demonstrating that a CIs can reverse the specific PIT effect.

These results can be interpreted using the S-O-R account and Konorski's (1967) conceptualization of conditioned inhibition. In the case of AC, C activated no- $O_1$  centres that counteracted the effect of A, resulting in a reduction of  $R_1$  performance. But in the case of AD, D activated no- $O_2$  centres which would have no effect on the activation of the  $O_1$ -centres produced by A, leaving the specific PIT effect intact. This evidence strongly supports the idea that conditioned inhibition is specific to the US of training. Because  $O_1$  and  $O_2$  were food pellets, both outcomes elicited the same motivational state; thus if the CIs act by activating a motivational state opposite to that of training (or directly suppressing such state) then the specific PIT effect should have been abolished in the presence of *all* the compounds.

In humans there is one study that assessed the effect of a non-reinforced stimulus on instrumental performance (Colagiuri &

Lovibond, 2015). However, in the experiments reported these authors only one instrumental response was trained so it is not possible to distinguish the general from the specific form of the PIT effect. Participants received pairings of a stimulus with food delivery and a second stimulus was presented without being reinforced (A->O<sub>1</sub>; B-). Then participants were trained to perform one response using food as a reinforcer (R->O<sub>1</sub>) and at test they had the opportunity to perform the response in the presence and absence of A and B. Unlike in the previous studies reported here participants continued receiving reinforcers during the test, and the results suggest that A increased responding over baseline levels while B reduced performance. But besides the fact that only one response was trained (and only one US was presented) the authors did not report evidence that B was a conditioned inhibitor so it is not entirely clear what the exact effect of this stimulus was.

In summary, the limited evidence in animal literature suggests that the CIs affect the specific PIT effect but, to my knowledge, there is no study of the effect of conditioned inhibition on specific PIT in humans. Thus, the experiments presented in this chapter aimed to use conditioned inhibition to discriminate between the different accounts of specific PIT. The goal of Experiment 1 was to establish a procedure that allows the formation of CIs and also produces the specific PIT effect. In this experiment two responses were reinforced with one of two outcomes (i.e. R<sub>1</sub>->O<sub>1</sub>; R<sub>2</sub>->O<sub>2</sub>) and in the Pavlovian phase different CS+s were paired with one of these outcomes (e.g. A->O<sub>1</sub>; B-

>O<sub>2</sub>). Also in this phase two CIs were trained to signal the absence of one of the outcomes (e.g. AX-; BY-). In the PIT test participants could perform both instrumental responses while the CS+s and the CIs were presented as single cues, and the rate of performance of each response was measured in the presence and absence of these stimuli. In Experiment 2 the CS+s were presented in the PIT test in compound with CIs that signalled the absence of the same outcomes predicted by the CS+s, to examine whether the CIs reduce the specific PIT effect produced by the CS+s. Finally Experiment 3 included the PIT tests of the first and second experiments in the same task, and also explored the extent to which the effect of the CIs on responding depended on their being presented with a CS that signalled the same outcome, as it had in the experiment reported by Laurent et al. (2015). To achieve this, in the PIT test the CS+s were also presented in compound with CIs trained with a different US. In all the experiments the inhibitory properties of the CIs were assessed independently by conducting a summation test before the PIT test.

## **2.8 Experiment 1**

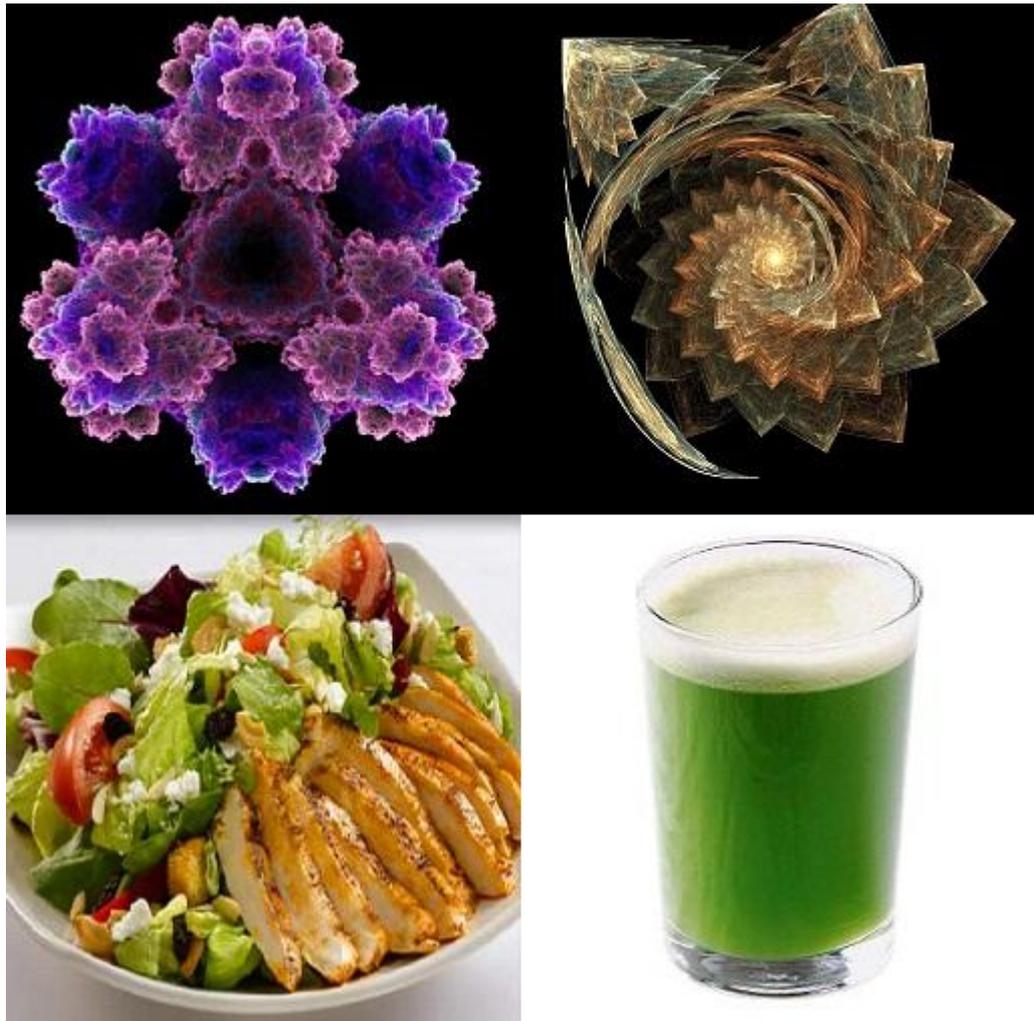
In this and the rest of the experiments in this thesis a computer task with a specific PIT design was used. This design consisted of an instrumental training phase and a Pavlovian conditioning phase, followed by a summation test. After this, an instrumental re-training phase was conducted and then a PIT test (see Table 1). The CSs

were neutral fractal images and the outcomes were pictures of food and drinks (see Figure 1).

*Table 1.* Design of Experiment 1.

Instrumental	Pavlovian phase		Summation test	PIT test
	Pre-training	Inhibition		
R <sub>1</sub> ->O <sub>1</sub>	A->O <sub>1</sub>	AD->O <sub>1</sub>	FX/GY <i>Inhibitory</i>	FX <i>Inhibitory</i>
R <sub>2</sub> ->O <sub>2</sub>	B->O <sub>2</sub>	AX <sup>-</sup>	FC/GH <i>Pre-exp</i>	GY <i>Inhibitory</i>
		X <sup>-</sup>	FN <sub>1</sub> /GN <sub>2</sub> <i>Novel</i>	FC <i>Pre-exp</i>
		BE->O <sub>2</sub>	FY/GX <i>Unrelated</i>	GH <i>Pre-exp</i>
		BY <sup>-</sup>	F/G <i>Excitatory</i>	FN <sub>1</sub> <i>Novel</i>
		Y <sup>-</sup>	X/Y <i>Inhibitory</i>	GN <sub>2</sub> <i>Novel</i>
		CH <sup>-</sup>	C/H <i>Pre-exp</i>	
		F->O <sub>1</sub>	N <sub>1</sub> /N <sub>2</sub> <i>Novel</i>	
		G->O <sub>2</sub>		

*Note:* A, B, C, D, E, F, G, H, X and Y: neutral fractal images; R<sub>1</sub> and R<sub>2</sub>: keyboard responses; O<sub>1</sub> and O<sub>2</sub>: food and drink images. - denotes no outcome.



*Figure 1.* Sample images of the pictures used as CSs and outcomes. Top panel: Fractal images used as CSs. Bottom panel: food and drink pictures used as outcomes.

In the *instrumental training* phase participants had to press the 'z' and 'm' keys ( $R_1$ ,  $R_2$ ) in order to obtain presentations of food and drinks pictures ( $O_1$ ,  $O_2$ ), each key being reinforced with one of these outcomes ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). This was followed by the *Pavlovian phase*, which was divided in two parts. In the *pre-training* participants received presentations of two CSs, each of them paired with one of the outcomes ( $A \rightarrow O_1$ ;  $B \rightarrow O_2$ ). Previous research suggests that this *pre-training* facilitates learning about the CIs in the following part of the

phase (He, Cassaday, Howard, Khalifa and Bonardi, 2012). Then in the *inhibitory training* A and B were presented in a non-reinforced compound with two stimuli (AX-, BY-) in order to establish X and Y as CIs. In addition A and B were also paired with the outcomes but in compound with two other cues (AD->O<sub>1</sub>; BE->O<sub>2</sub>), an arrangement made to maintain the excitatory strength of A and B and also to avoid participants learning that all compounds were not reinforced. Two more stimuli were presented in a non-reinforced compound (CH-) similarly to AX and BY, but because this compound did not predict the absence of an expected outcome (no CS+ in the compound), C and H should not become CIs, and so served as control stimuli for the summation test and PIT test.

Although this conditioned inhibition procedure should allow X and Y to acquire inhibitory properties, it has been suggested (Williams, Travis and Overmier, 1986) that it also allows the formation of within-compound associations between the CSs+ and the target inhibitors (A-X and B-Y in this experiment). Thus later at test, X may be able to activate a representation of A, which in turn could lead to the activation of O<sub>1</sub>, diminishing the inhibitory properties of X. For this reason, and in order to extinguish any within-compound association with the CSs+, single presentations of X and Y were included. Additionally two more stimuli were established as test excitors (F->O<sub>1</sub>; G->O<sub>2</sub>), which were used to measure the inhibitory strength of X and Y in the summation test and also as excitatory stimuli in the PIT test.

In the summation test F and G were presented with the inhibitory cues X and Y (FX, GY) and with the pre-exposed control stimuli C and H (FC, GH), and participants had to rate the likelihood of each of the outcomes ( $O_1$ ,  $O_2$ ) for each of these compounds. If X and Y had acquired inhibitory strength, then participants' expectation of the outcomes should be lower during FX and GY than during FC and GH. However, because the CH compound was presented non-reinforced in an excitatory context, it is possible that each of these cues acquired, to some degree, inhibitory properties due to *differential inhibition*. For this reason, F and G were also presented with two novel stimuli (FN<sub>1</sub> and GN<sub>2</sub>) to serve as an alternative control.

After the summation test participants received instrumental re-training, followed by the PIT test. In this test participants had the chance to perform both instrumental responses in the presence and absence of the CSs+ (F, G), CIs (X, Y) and control cues (C, H). It was expected that F would elicit more R<sub>1</sub> than R<sub>2</sub> responding, and the reverse for G, i.e. the specific PIT effect, because F and R<sub>1</sub> were paired with O<sub>1</sub>, while G and R<sub>2</sub> were reinforced with O<sub>2</sub>. But if the R-O associations contain information about the outcomes (e.g. Adams & Dickinson, 1981) and the CIs suppress the representation of these outcomes, then according to the S-O-R accounts the CIs should selectively reduce responding. In other words X should suppress activation of O<sub>1</sub>, reducing the performance of R<sub>1</sub> compared to R<sub>2</sub>, and the opposite pattern should be seen for Y. Finally, the pre-exposed control stimuli, C and H, did not predict the absence of any particular

outcome, so it was anticipated that these cues would have no effect on instrumental performance.

### **2.8.1 Method**

*Participants.* Thirty-five students from the University of Nottingham participated in this experiment (6 males and 29 females), aged between 18 and 35 years old. In this and the rest of the experiments the students from the School of Psychology were given course credit for their participation, while the remaining received an inconvenience allowance of £4.

*Apparatus and materials.* The experiment was programmed in PsychoPy (Peirce, 2007) and the task was conducted on a standard computer with a 20-inch screen. The CSs were 12 neutral fractal images (5.7 x 5.7 cm), and the outcomes were 4 pictures of food ( $O_1$ ) and 4 pictures of drinks ( $O_2$ ) of 8 x 8 cm in size. In the compound trials of the Pavlovian phase, two CSs appeared side by side in the centre of the screen, i.e. one on the left and the other on the right, and in the single trials only one CS was presented (either on the left or the right). All CSs appeared an equal number of times on each side of the screen. In each of the reinforced trials one of the four images (food or drink depending on the trial) was positioned at the centre of the screen. In the non-reinforced trials the screen remained unchanged for the same amount of time. In the instrumental phase and PIT test, a

fixation cross (10 x 10 mm) was positioned at the centre of the screen, which was replaced by a fixation dot (3 x 3 mm) in the Pavlovian phase. The instrumental responses were pressing the keys 'z' and 'm' and for half of the participants 'z' was reinforced with O<sub>1</sub> and 'm' with O<sub>2</sub>, while the opposite for the other half. Participants also had to use the mouse in the summation test; finally, in order to advance to each of the phases, participants had to press the 'space' bar. The stimuli X, Y, C, H, N<sub>1</sub> and N<sub>2</sub> were partially counterbalanced with each other, resulting in a sub-set of 12 experimental conditions. Then this sub-set was doubled before F and G were also counterbalanced, resulting in 24 counterbalancing conditions. The complete task instructions and the counterbalancing tables are presented in Appendix A and B, respectively (pp. A-1 to B-4).

### *Procedure*

Participants received an information sheet and the chance to ask any questions related to the task. Then they completed a consent form and were guided to a quiet room. A general description and specific instructions were presented on the screen before each of the phases.

*Instrumental phase.* Participants were instructed to discover the relationship between key pressing and the delivery of rewards (outcomes) and to try to obtain as many rewards as they could. Each response was reinforced according to a variable ratio (VR) 5 schedule, i.e. on average, every 5 responses on the same key produced an

outcome presentation. The fixation cross remained on the screen except when the outcomes were presented on the reinforced trials. Each outcome presentation lasted 0.8 s unless any key (R<sub>1</sub> or R<sub>2</sub>) was pressed, in which case the outcome image was replaced by the fixation cross. The phase ended when participants had received 100 outcome presentations in total.

*Pavlovian phase.* Participants were instructed to pay attention to the relationship between the neutral images and the rewards but told that no response was required. Each trial began with the fixation dot for 2s, followed by a CS (single or compound) for 2s. In the case of the reinforced trials, the CS was followed by an outcome that lasted 0.8s, but on non-reinforced trials the background screen remained unchanged for 0.8s. After this the fixation dot appeared again starting a new trial.

This phase was divided in two stages: *pre-training* and *inhibitory training*. The *pre-training* stage was divided in two blocks, each comprising four trials of A->O<sub>1</sub> and four of B->O<sub>2</sub>. The *inhibitory training* stage was divided in 4 blocks, and each of them consisted of 2 trials of AD->O<sub>1</sub>, BE->O<sub>2</sub>, F->O<sub>1</sub> and G->O<sub>2</sub>; and 3 trials of AX-, BY-, CH-, X- and Y-. The order of the different types of trial was semi-random within each block.

A *Pavlovian test* was conducted after the second and fourth block of the inhibitory training stage, in which participants had to answer a series of questions about the relationship between each CS

or CS compound and the outcomes. At the top of the screen the text "In a scale from 1 to 100, how likely is it that this image will be followed by" together with "FOOD" or "DRINK" was presented. A CS(s) appeared at the centre of the screen and a rating scale below it with the number '0' on the left and '100' on the right. Participants used the mouse to click at the point of the scale they considered appropriate. Each of these series contained two questions per stimulus/compound: one for  $O_1$  and one  $O_2$ . At the end of the second series of questions participants had the chance to take a break. The order of the questions was randomized. The purpose of these questions was to assess whether participants had paid attention during the phase, and also to serve as a tool to exclude those that did not learn the F- $O_1$  and G- $O_2$  associations, which were considered critical because these cues were used to assess inhibition in the summation test.

*Summation test.* This was conducted immediately after the break and it was identical to the Pavlovian test, except that the questions referred to different CS and CS compounds. F and G were presented with X and Y (FX, GY), C and H (FC, GH) and two novel stimuli (FN<sub>1</sub> and GN<sub>2</sub>) and each of these stimuli was also presented alone (F, G, X, Y, C, H, N<sub>1</sub> and N<sub>2</sub>). This test had one block with two questions per cue/compound, one about  $O_1$  and the other about  $O_2$ , and the order of the questions was randomized.

*Instrumental retraining.* The same as the instrumental phase, except that it ended when participants received 50 outcomes in total.

*PIT test.* Participants were instructed to press the 'z' and 'm' keys as much as they wanted to and that sometimes some of the neutral pictures would appear. Each trial consisted of the fixation cross for 2s (preCS period), followed by a 2-s presentation of the CS (CS period). The CSs presented were F, G, X, Y, C and H. This test was conducted in extinction (i.e. none of the responses was reinforced) and it was divided into 6 blocks, each of them with one trial per CS, and the order of the CS presentations was randomized in each block. The rate of R<sub>1</sub> and R<sub>2</sub> performance was recorded during the preCS and CS periods.

*Statistical analysis.* All the data was analysed using ANOVAs and significant main effects with pairwise comparisons post hoc tests using the Bonferroni correction. Significant three-way interactions were analysed with further two-way ANOVAs, and two-way interactions with simple main effects. As a measure of effect size partial eta-squared ( $\eta_p^2$ ) are reported.

The answers to the rating scales in the Pavlovian phase were used to determine an exclusion criterion. When the questions were about the F-O<sub>1</sub> and G-O<sub>2</sub> relationships, i.e. whether F would be followed by O<sub>1</sub> or G by O<sub>2</sub>, the ratings were expected to be close to 100, but close to 0 when the questions related to F-O<sub>2</sub> and G-O<sub>1</sub>. Thus a learning score was created by taking the mean of the rating scores to F-O<sub>1</sub> and G-O<sub>2</sub>, and subtracting the mean of the scores to F-O<sub>2</sub> and G-O<sub>1</sub>, so that the closer the value to 100 the better the learning. Based on this a learning score of 50 in the last block of questions was set as

a minimum, because lower values imply that the discrimination between F and G was not acquired.

The data from the PIT test were grouped by congruence: if the response ( $R_1$  or  $R_2$ ) made during the preCS or CS periods was previously reinforced with the same outcome ( $O_1$  or  $O_2$ ) as the CS being presented, then this response was defined as *congruent*, if not as *incongruent*. For example, both F and  $R_1$  were paired with  $O_1$  in training so  $R_1$  responses made before or during F were considered *congruent* responses, while  $R_2$  responses were *incongruent*. In the case of the inhibitory cues, X was trained as an inhibitor for  $O_1$  and Y for  $O_2$  so the same logic was applied as for F and G; but because the pre-exposed control cues did not have any relationship with the outcomes they were arbitrarily assigned to one of them; thus for half of the participants  $R_1$  responses were considered congruent for C and  $R_2$  responses were congruent for H, and the opposite for the other half. After grouping the responses, and if no differences were found during the preCS period, *PIT scores* were computed by subtracting the responses made during the preCS period from the responses made during the CS period and converting to responses per minute (rpm). Thus more congruent than incongruent PIT scores indicated the *specific PIT effect*.

## 2.8.2 Results

*Pavlovian phase.* Eleven participants failed to pass the exclusion criterion set for this experiment and they were replaced to complete the 24 counterbalancing conditions.

*Instrumental phase.* The mean number of responses and outcome presentations are presented in Table 2. No differences were found between type of response made ( $R_1$ ;  $R_2$ ) nor type of outcome earned ( $O_1$ ;  $O_2$ ) in instrumental training,  $F_s < 1$ .

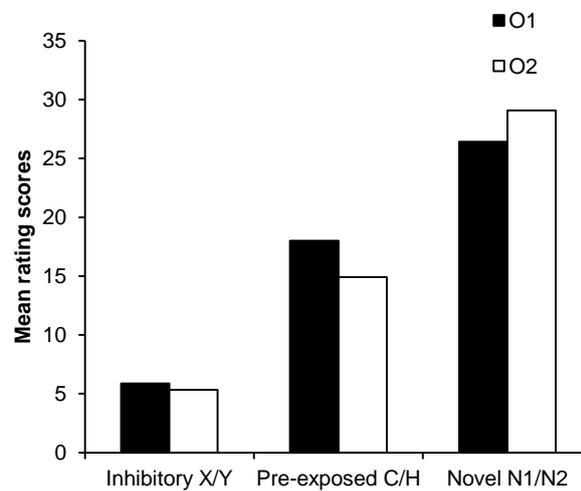
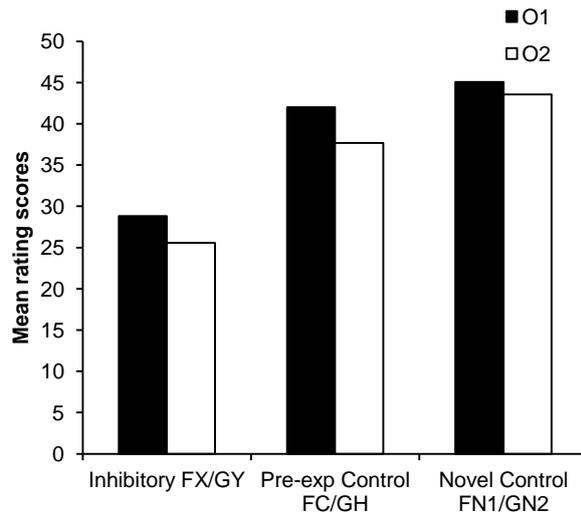
*Table 2.* Mean number of R1 and R2 responses and mean number of O1 and O2 deliveries in the instrumental phase of Experiments 1, 2 and 3.

	$R_1$	$R_2$	$O_1$	$O_2$
<b>Experiment 1</b>	281.92	255.79	54.46	50.83
<b>Experiment 2</b>	242.81	280.7	56.3	49.9
<b>Experiment 3</b>	247.38	284.13	50.79	55.17

*Summation test.* In this test participants rated the likelihood that different cues and compounds were going to be followed by  $O_1$  or  $O_2$ . The critical questions referred to the likelihood of  $O_1$  in the presence of FX, FC and  $FN_1$ , and of  $O_2$  in the presence of GY, GH and  $GN_1$ . The mean rating scores for these questions are presented in the top panel

of Figure 2, which suggests that participants rated the outcome as less likely to occur during the inhibitory than during the control compounds; however, this was not fully supported by the statistical analysis. An ANOVA with Outcome ( $O_1$ ,  $O_2$ ) and CS (inhibitory, pre-exposed, novel) as factors revealed a significant main effect of CS,  $F(1, 23) = 4.63$ ,  $p = .015$ ,  $MSe = .082$ ,  $\eta_p^2 = .17$ , and nothing else was significant, largest  $F(1, 23) = 1.09$ ,  $p = .306$ ,  $MSe = .03$ . Analysis of the significant main effect showed a significant difference between FX/GY and FN<sub>1</sub>/GN<sub>2</sub> ( $p < .01$ ), but not between FX/GY and FC/GH. However, no difference was found between FC/GH and FN<sub>1</sub>/GN<sub>2</sub>.

The ratings to the questions that referred to the opposite outcome of training were also analysed. The ratings to the likelihood of  $O_2$  for FX, FC and FN<sub>1</sub> were 7.3, 9.5 and 13.4 respectively, and to the likelihood of  $O_1$  for GY, GH and GN<sub>2</sub> were 6, 12.1 and 11.7 respectively. The ANOVA showed no significant differences or interactions, largest  $F(2, 46) = 2.9$ ,  $p = .06$ ,  $MSe = .015$ .



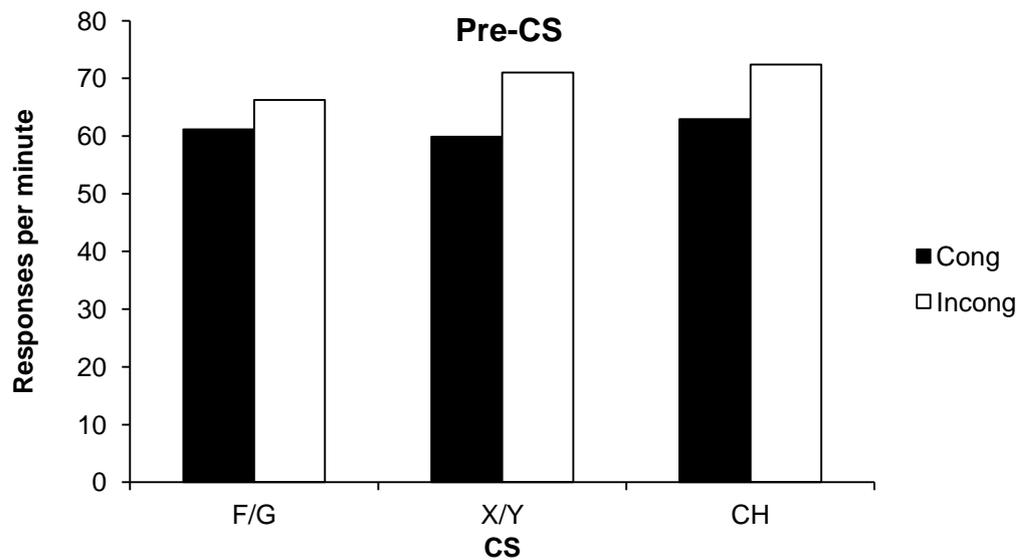
*Figure 2.* Summation test in Experiment 1. Top panel: Mean ratings of the likelihood of O1 occurrence during FX, FC and FN1, and O2 occurrence during GY, GH and GN2. Bottom panel: Mean ratings of the likelihood of O1 occurrence during X, C and N1, and O2 during Y, H and N2.

The mean ratings to the single cues (X/Y, C/H and N<sub>1</sub>/N<sub>2</sub>) were also analysed and they are plotted in the bottom panel of Figure 2. As

the control cues had no relationship with the outcomes they were arbitrarily assigned to one of them ( $O_1$  for C and  $N_1$ , and  $O_2$  for H and  $N_2$ ). Neither the CIs nor the control stimuli were ever presented with the outcomes, so ratings close to zero were expected. However, the ratings to the control stimuli seem to be higher than those to the inhibitory cues. The ANOVA showed a significant main effect of CS,  $F(2, 46) = 10.34, p < .001, MSe = 284, \eta_p^2 = .31$ , and nothing else was significant,  $F_s < 1$ . Post-hoc tests showed a significant difference between inhibitory and novel cues,  $p = .001$ , but not between inhibitory and pre-exposed,  $p = .057$ , or between pre-exposed and novel cues,  $p = .093$ . As in the case of the compounds, the novel cues were rated significantly higher than the inhibitory cues, but no different from the pre-exposed control stimuli. It is possible that the excitatory properties of the CS+ transferred to the rest of the cues through stimulus generalisation. However, because the CIs (X and Y) and pre-exposed cues (C and H) were presented during training, any generalised associative strength should have been extinguish at the moment of the test, which would explain why the novel cues produced higher ratings than these stimuli.

*PIT test.* The responses from the pre-CS period were collapsed across blocks and presented in Figure 3. The graph shows unexpected differences between congruent and incongruent responses, which was confirmed by the statistical analysis. An ANOVA with block (6), congruence (congruent, incongruent) and CS (F/G, X/Y, C/H) revealed a significant main effect of congruence,  $F(1, 23) = 8.94, p = .007, MSe$

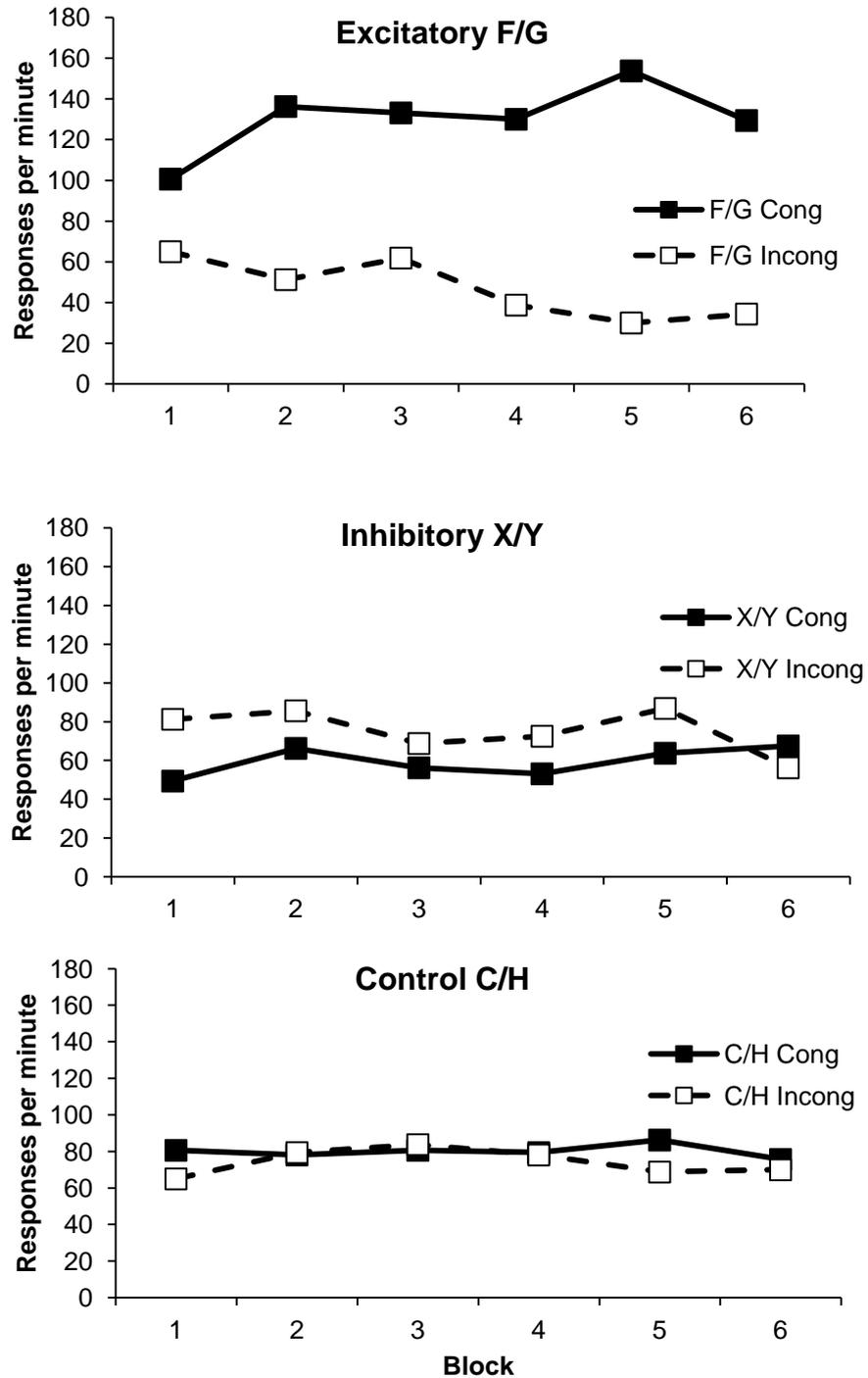
= 1.97,  $\eta_p^2 = .28$ . Nothing else was significant, largest  $F(5, 115) = 1.13$ ,  $p = .35$ ,  $MSe = 3.22$ . Although it is possible that participants continued responding after CS presentation, affecting the pre-CS period for the next CS, it is not clear why the elevation in the pre-CS responses was only found in the congruent trials.



*Figure 3.* Response rate during the pre-CS period of the PIT test in Experiment 1. F/G: excitatory CS; X/Y: inhibitory stimuli; C/H: pre-exposed control stimuli.

These differences might affect and compromise the interpretation of the PIT scores: If the responses of the preCS are subtracted from those made during the CS period, this would increase the difference between congruent and incongruent responses, resulting in an apparently larger specific PIT effect. For this reason the CS responses were analysed without subtracting the preCS

responding. Although the CS period does not allow one to infer an increase in performance produced by the CS presentations, it still allows observation of the specific PIT effect, i.e. more congruent than incongruent responses. The responses performed in the CS period during F/G, X/Y and C/H are plotted in Figure 4, which suggests that the specific PIT effect was found in the presence of F/G but not of X/Y nor C/H. Moreover, the opposite pattern was found for X/Y, i.e. more incongruent than congruent responses. The ANOVA revealed a significant main effect of congruence,  $F(1, 23) = 17.61, p < .001, MSe = 8.24, \eta_p^2 = .43$ ; CS,  $F(2, 46) = 5.12, p = .01, MSe = 7.18, \eta_p^2 = .18$ ; a significant Congruence x CS interaction,  $F(2, 46) = 26.73, p < .001, MSe = 8.18, \eta_p^2 = .54$ , and a significant Block x Congruence x CS interaction,  $F(10, 230) = 2.1, p = .025, MSe = 2.19, \eta_p^2 = .08$ . To analyze the triple interaction, two-way ANOVAs were conducted for each of the CSs. For F/G the analysis showed a significant main effect of congruence,  $F(1, 23) = 34.29, p < .001, MSe = 16.32, \eta_p^2 = .60$ ; and a significant Block x Congruence interaction,  $F(5, 115) = 4.02, p = .002, MSe = 2.82, \eta_p^2 = .15$ . Simple main effects on the interaction revealed an effect of congruence on blocks 2 - 6 inclusive, smallest  $F(1, 23) = 13.47, p = .001$  on block 3. This confirms that F and G produced the *specific transfer effect* after the first block of the test. The ANOVA conducted on data for X/Y showed no significant main effect or interaction, largest  $F(1, 23) = 3.55, p = .07, MSe = 5.65$  for the main effect of congruence, and neither did the analysis of the data for C/H, largest  $F(1, 23) = 1.07, p = .312, MSe = 2.64$ .



*Figure 4.* Mean rate of congruent and incongruent responding for each type of stimulus in the CS period of the PIT test of Experiment 1. Top panel: responses during the excitatory cues F/G. Middle panel: responses during the inhibitory stimuli X/Y. Bottom panel: responses during the pre-exposed control stimuli C/H.

### 2.8.3 Discussion

These results confirm that this procedure is effective in producing the specific PIT effect. Presentations of CSs+ resulted in performance of more congruent than incongruent responses, i.e.  $R_1 > R_2$  for F and  $R_2 > R_1$  for G. Although numerically the CIs produced less congruent than incongruent responses, this was not significant. Thus their effect at test was no different from that of the control stimuli (C/H).

The results of the summation test are not entirely clear about the inhibitory properties of the CIs. While the CIs reduced the expectancy of the outcomes more than the novel cues ( $N_1/N_2$ ), this effect was not strong enough to differentiate the CIs from the pre-exposed stimuli C and H, which presumably are a more conservative control. However, the ratings of C/H were no different to those of the novel control cues. One possibility is that the CIs did not acquire enough inhibitory strength during training, but a second is that C and H also acquired inhibitory properties to some degree, perhaps due to differential inhibition.

Some of the limitations of this study are noteworthy. One of them is that in the Pavlovian phase participants had to passively watch the screen while the stimuli were presented, which could have provoked a loss in the attention to the task. This is relevant considering that more than 30% of the participants were replaced because they could not discriminate between the associations of F

and G with the outcomes. A second issue is that the pre-exposed control cues (C, H) did not receive the same training as the CIs (X, Y). C and H were only presented as a compound (CH<sup>-</sup>), while X and Y were also presented alone (AX<sup>-</sup>, BY<sup>-</sup>, X<sup>-</sup>, Y<sup>-</sup>). These problems were addressed in Experiment 2.

## 2.9 Experiment 2

A more direct evaluation of the effect of inhibition on PIT is to present the CIs in compound with CS+s. If activation of the outcome representation at test is crucial to produce the specific PIT effect and the CIs suppress this representation then the CS should produce a reduced effect. But if activation of the outcome representation at test is irrelevant then the CIs should not diminish the CSs' ability to increase responding. Thus one of the main changes in this experiment (see Table 3) was replacing the single cues presented during the PIT test with the CS+s in compound with the CIs (FX, GY). To confirm that any reduction of the PIT effect was not caused by presenting the CSs with any non-reinforced stimulus, the CSs were also presented with the pre-exposed cues (FC, GH) and novel stimuli (FN<sub>1</sub>, GN<sub>2</sub>), which served as control compounds.

Table 3. Design of Experiment 2.

Instrumental	Pavlovian phase		Summation test	PIT test
	Pre-training	Inhibition		
R <sub>1</sub> ->O <sub>1</sub>	A->O <sub>1</sub>	AD->O <sub>1</sub>	FX/GY <i>Inhibitory</i>	FX <i>Inhibitory</i>
R <sub>2</sub> ->O <sub>2</sub>	B->O <sub>2</sub>	AX <sup>-</sup>	FC/GH <i>Pre-exp</i>	GY <i>Inhibitory</i>
		X <sup>-</sup>	FN <sub>1</sub> /GN <sub>2</sub> <i>Novel</i>	FC <i>Pre-exp</i>
		BE->O <sub>2</sub>	FY/GX <i>Unrelated</i>	GH <i>Pre-exp</i>
		BY <sup>-</sup>		FN <sub>1</sub> <i>Novel</i>
		Y <sup>-</sup>		GN <sub>2</sub> <i>Novel</i>
		CH <sup>-</sup>		
		C <sup>-</sup>		
		H <sup>-</sup>		
		F->O <sub>1</sub>		
		G->O <sub>2</sub>		

Note: A, B, C, D, E, F, G, H, X and Y: neutral fractal images; R<sub>1</sub> and R<sub>2</sub>: keyboard responses; O<sub>1</sub> and O<sub>2</sub>: food and drink images. - denotes no outcome.

A second goal of this experiment was to assess the outcome-specificity of conditioned inhibition - whether a CI that predicts the absence of one outcome is capable of reducing participants' expectation of a different outcome. In the summation test two *unrelated* compounds were included: FY and GX. For example, F predicted O<sub>1</sub> deliveries but Y signalled the absence of O<sub>2</sub>. If inhibition is outcome-specific, Y should only suppress the activation of O<sub>2</sub>, and have no effect on the participants' expectations of O<sub>1</sub> elicited by F, and so ratings of FX should be lower than those of FY. If inhibition is not outcome specific, participants' ratings of FY and FX should not differ.

In order to address some of the limitations of the previous experiment, a second main change was the modification of the Pavlovian phase. In this experiment participants had to predict the outcome that followed each of the CSs by pressing a key, and they

received feedback after each trial. This replaced the rating questions and it should increase participants' attention to the Pavlovian relationships and also their motivation to pay attention while performing the task. Additionally, providing the correct and incorrect choices of the participants trial by trial was thought to be a better assessment of learning.

Finally, to address the fact that the CIs (X and Y) and the control stimuli (C and H) received different treatment in the previous experiment, in this study C and H were also presented alone. One unintended benefit of this new arrangement is that if it is true that C and H acquired inhibitory properties due to differential inhibition, then the inclusion of more presentations of these cues should increase their negative strength, which should be evident in the summation test.

In summary, this experiment differed from Experiment 1 in that: a) compounds instead of single cues were presented in the PIT test; b) *unrelated* compounds were included in the summation test to assess outcome-specificity of conditioned inhibition; c) the rating scales were removed from the Pavlovian phase and participants had to actively predict the outcomes; and d) single trials of C and H were included.

### **2.9.1 Method**

*Participants.* Twenty-seven students participated in this Experiment (4 males and 23 females) aged between 18 and 28 years old.

*Procedure.* Everything was the same as in Experiment 1, unless otherwise stated.

*Instrumental phase.*

*Pavlovian phase.* After the fixation dot (2s), the question "Which reward will appear now?" was presented at the top of the screen, a CS(s) below it and the text "1) Food 5) Drink 9) Nothing" at the bottom. This screen remained until participants pressed one of these numbers on the keyboard. Then the question, text and CSs were removed and the corresponding outcome (or background screen) appeared for 2s with a feedback message: when the response was correct, the text "Correct!" was positioned at the top of the screen in green letters, but if it was an incorrect answer, the message "Oops! That was wrong" was presented at the bottom of the screen in red letters. Then the fixation dot reappeared, marking the beginning of a new trial.

The *inhibitory training* now included single presentations of C and H. It was divided into 2 blocks, each of which consisted of 4 presentations of AD->O<sub>1</sub> and BE->O<sub>2</sub>, 6 presentations of F->O<sub>1</sub>, G->O<sub>2</sub>, AX-, BY- and CH-, and 3 presentations of X-, Y-, C- and H-. The number of presentations of the non-reinforced single cues was reduced in order to maintain the same proportion of reinforced to nonreinforced trials as in the previous experiment. Participants had the chance to take a break between these blocks.

*Summation test.* This test was identical to that of Experiment 1, except that it included trials of FY and GX and the single cue presentations

were removed (F, G, X, Y, C, H, N<sub>1</sub> and N<sub>2</sub>). The test comprised two blocks, each of them containing two questions per compound. Unlike in the previous experiment, all the questions referred to the relevant outcome: that is, for compounds containing F the questions referred to O<sub>1</sub> and for those containing G they referred to O<sub>2</sub>.

*Instrumental re-training.*

*PIT test.* Participants received presentations of the FX, GY, FC, GH, FN<sub>1</sub> and GN<sub>2</sub>. The CS images were presented side by side in the centre of the screen, and the position of the images was counterbalanced, that is, all the images were presented the same number of times in the left and right position. The test was divided into 3 blocks, and each of them consisted of 2 presentations of each compound.

## **2.9.2 Results**

*Pavlovian phase.* Participants had to predict the outcome that followed each CS, and in each of the trials a correct response was recorded as 1 and an incorrect choice as 0. The mean scores from inhibitory training were grouped by block and are presented in Figure 5. The graph suggests that participants improved their ability to predict the outcomes for all the compounds except AD/BE. An ANOVA with Block (1, 2) and CS (AD/BE, AX/BY, CH, F/G, X/Y and C/H) showed a

significant main effect of block,  $F(1, 26) = 67.31, p < .001, MSe = .03, \eta_p^2 = .72$ , CS,  $F(5, 130) = 12.23, p < .001, MSe = .043, \eta_p^2 = .32$ , and a significant Block x CS interaction,  $F(5, 130) = 5.69, p < .001, MSe = .024, \eta_p^2 = .18$ . Simple main effects revealed an effect of block for all the compounds except for AD/BE,  $F < 1$ , smallest  $F(1, 26) = 11.61, p = .002$  (for CH). It is possible that participants had trouble predicting the outcomes for AD/BE either because they received fewer presentations of these compounds than the rest (8 trials of each compound vs 12 trials F, G, AX, BY and CH), or because A and B were also presented in non-reinforced compounds with the CIs (AX-, BY-), which resulted in lower accuracy for these compounds.

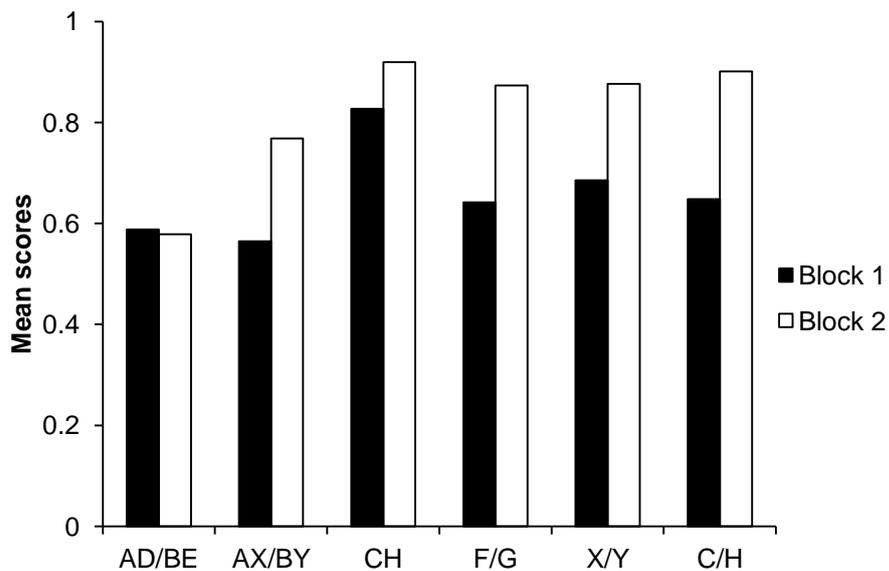


Figure 5. Mean scores grouped by CS and block for responses made in the Pavlovian phase of Experiment 2.

*Instrumental phase.* The mean number of responses and outcome deliveries are presented in Table 2. No differences were found

between type of response made ( $R_1$ ;  $R_2$ ) nor type of outcome received ( $O_1$ ;  $O_2$ ), largest  $F(1, 26) = 1.31$ ,  $p = .262$ ,  $MSe = 14747.36$ .

*Summation test.* The mean rating scores for the summation test are presented in Figure 6, which suggests that participants rated  $FN_1/GN_2$  higher than all the other compounds, which did not differ. An ANOVA with outcome ( $O_1$ ,  $O_2$ ) and compound (FX/GY, FC/GH,  $FN_1/GN_2$ , FY/GX) as factors revealed a significant main effect of compound,  $F(3, 78) = 11.18$ ,  $p < .001$ ,  $MSe = .059$ ,  $\eta_p^2 = .30$ , and nothing else was significant, largest  $F(3, 78) = 1.639$ ,  $p = .19$ ,  $MSe = .018$ . Analysis of the significant main effect confirmed differences between  $FN_1/GN_2$  and FX/GY,  $p = .007$ , FC/GH,  $p = .004$ , and FY/GX,  $p = .003$ . No other difference was significant ( $p = 1$ ). As in Experiment 1, these results showed that the effect of the CIs on the outcome expectancy was no different to that of the pre-exposed control cues. Moreover, in this experiment C and H reduced the expectancy of the outcomes more than  $N_1$  and  $N_2$ , which might have been caused by the inclusion of single presentations of C and H. The fact that more training with these cues increased their ability to reduce participants' expectation of the outcomes supports the idea that C and H acquired some degree of inhibition. The second important results relates to FY and GX. These *unrelated* compounds were rated as low as the inhibitory and pre-exposed controls, which suggest that at least in this procedure the effect of the inhibitors is not specific to the outcome of training. However, the fact that the CIs were not rated different that the pre-

exposed control cues needs to be consider before making this conclusion.

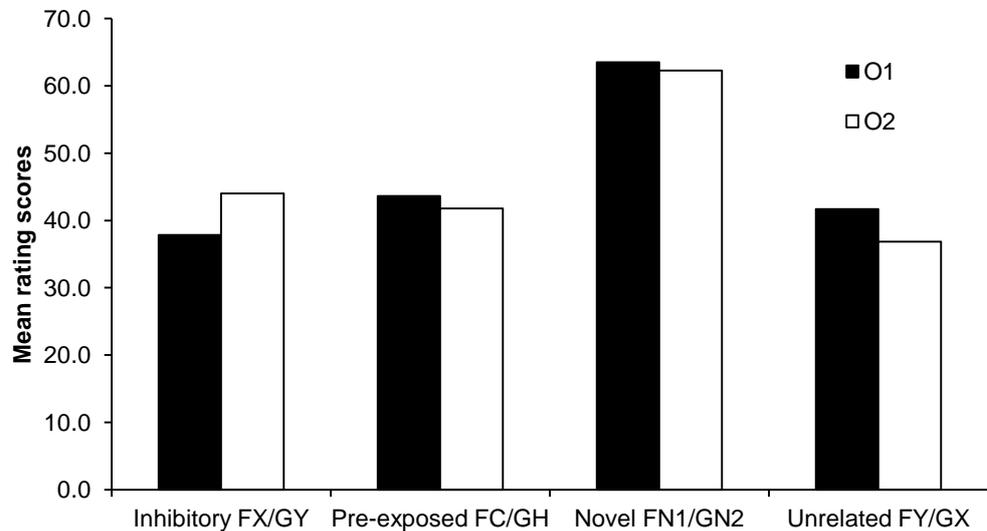


Figure 6. Mean ratings of the likelihood of O1 occurrence during FX, FC, FN1 and FY; and O2 occurrence during GY, GH, GN2 and GX in the summation test of Experiment 2.

*PIT test.* No significant differences were found in the preCS period (see below for analysis), so the PIT scores were calculated and presented in the top panel of Figure 7. The graph suggests that the specific PIT effect (i.e. more congruent than incongruent responses) was present on both FC/GH and FN<sub>1</sub>/GN<sub>2</sub> trials, but this effect was greatly diminished on the FX/GY trials. An ANOVA with block (1-3), congruence and compound (FC/GH, FX/GY, FN<sub>1</sub>/GN<sub>2</sub>) as factors showed a significant main effect of congruence,  $F(1, 26) = 9.74, p = .004, MSe = 49.82, \eta_p^2 = .27$ , and a significant Congruence x Compound interaction,  $F(2, 52) = 3.3, p = .045, MSe = 15.88, \eta_p^2 = .11$ . Nothing else was significant, largest  $F(2, 52) = 2.8, p = .07, MSe$

= 9.55 for the Block x Congruence interaction. The analysis of the significant interaction showed an effect of congruence on FC/GH,  $F(1, 26) = 11.54, p = .002$ , and FN<sub>1</sub>/GN<sub>2</sub>,  $F(1, 26) = 11.44, p = .002$ , but not on FX/GY trials,  $F < 1$ . These analyses confirmed that the CS+s produced the specific PIT effect when they were in compound with pre-exposed and novel control stimuli, but this effect disappeared when they were accompanied by the CIs.

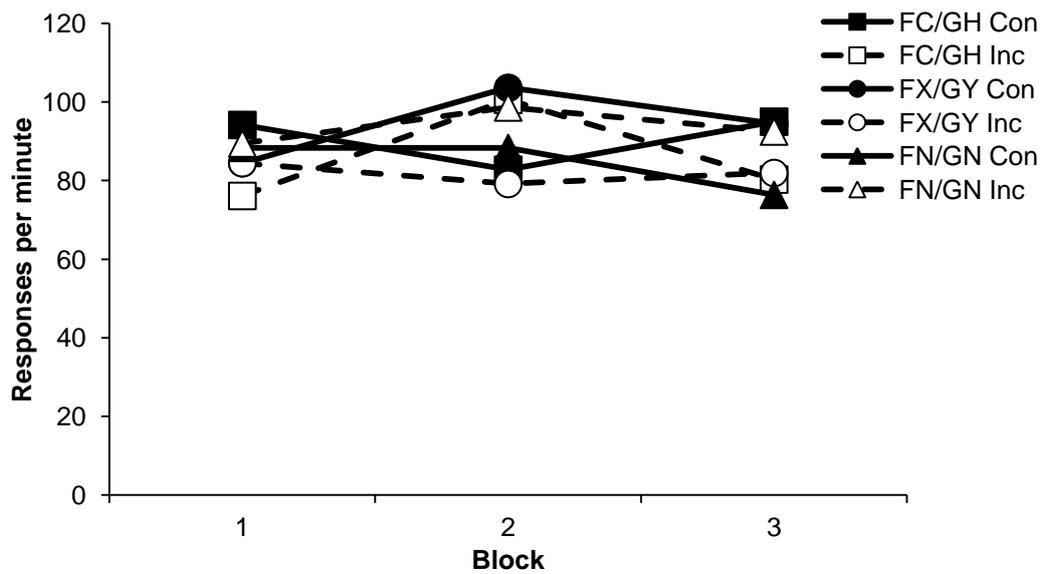
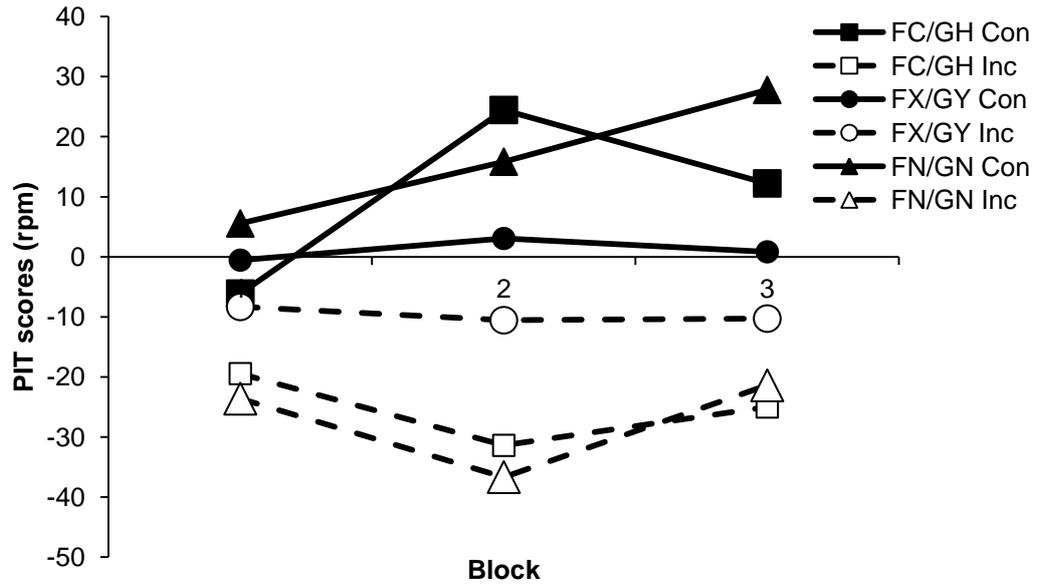


Figure 7. Mean responses grouped by congruence for each of the compounds in the PIT test of Experiment 2. Top panel: PIT scores. Bottom panel: responses during the preCS period.

The data from the pre-CS period are presented in the bottom panel of Figure 7, which suggests no clear differences in performance. The ANOVA on these data showed a significant triple interaction,  $F(4,$

104) = 16.83,  $p = .037$ ,  $MSe = 6.35$ ,  $\eta_p^2 = .39$ . Nothing else was significant, largest  $F(2, 52) = 1.25$ ,  $p = .29$ ,  $MSe = 17.13$ . Two-way ANOVAs with block and congruence as factors were conducted for each type of compound. For FC/GH the analysis revealed a significant Block x Congruence interaction,  $F(2, 52) = 3.6$ ,  $p = .034$ ,  $MSe = 6.53$ ,  $\eta_p^2 = .12$ . Although the Figure suggests more responding during the incongruent trials in the second and third block, simple main effects on the interaction showed no significant effect of congruency at any of the blocks, largest  $F(1, 26) = 3.27$ ,  $p = .082$ . The analyses of FX/GY and FN<sub>1</sub>/GN<sub>2</sub> showed no significant main effect or interaction, largest  $F(2, 52) = 1.72$ ,  $p = .19$ ,  $MSe = 5.21$ .

### **2.9.3 Discussion**

The results of PIT test confirmed that the CS+s produced the specific PIT effect when presented with neutral stimuli, but this effect was greatly reduced when they were in compound with a CI. These results support the S-O-R accounts of the PIT effect, according to which the CS activates a representation of the outcome that is responsible for the specific PIT effect. Then if the CI suppresses the outcome representation the specific PIT effect should be reduced, as was found in this experiment. On the other hand, if the PIT effect is produced by the CS eliciting responses through an S-R association,

then presenting CSs with the CIs should not influence the specific PIT effect, at least not more than presenting them with the neutral stimuli.

The results of the summation test showed that the CIs reduced participants' expectations of the outcomes more than the novel stimuli, replicating the findings of Experiment 1. Furthermore, the ratings of FY/GX were no different to those of FX/GY and lower than those for FN<sub>1</sub>/GN<sub>2</sub>, supporting the idea that the inhibition is not outcome-specific. But the effect of X and Y did not differ from that of C and H. This experiment included single trials of C and H and unlike the previous experiment participants rated these compounds (FC, GH) lower than the novel stimuli (FN<sub>1</sub>, GN<sub>2</sub>). Although this supports the idea that C and H acquired inhibitory strength by differential inhibition, these stimuli had no effect in the PIT test. Thus it is possible that the summation test was not sensitive enough to distinguish between both types of cues.

### **2.10 Experiment 3**

The main goal of this experiment was to replicate the results of Experiments 1 and 2 in the same study, and to accomplish this two PIT tests were conducted (see Table 4). *Test A* was exactly as in Experiment 2 and *Test B* included presentations of X, Y, F and G, as in Experiment 1, and also FY and GX. These compounds were

included to assess if the non-specific effect of inhibition found in the summation test of Experiment 2 can also be found in the PIT test.

In this experiment the Pavlovian phase was conducted before instrumental training, an arrangement that was made to reduce the chances of C and H acquiring inhibitory properties. If differential inhibition occurred in the previous experiment, then this might have been caused by the excitatory strength of the context. In the instrumental phase a great number of unsignalled outcomes were delivered, which might have provided the context with sufficient excitatory properties for C and H to acquire inhibitory strength (Baker, 1977). Thus this experiment began with the Pavlovian phase in order to conduct inhibitory training in a relatively neutral context. In addition to this, the questions and the scales of the summation test were also modified to improve its ability to detect differences between the stimuli.

The Pavlovian phase was also slightly modified: new reinforced compound (DE->O<sub>1</sub> and IJ->O<sub>2</sub>) were included in both the pre-training and inhibitory training stages. In the previous experiments participants' accuracy to AD and BE, which were paired with O<sub>1</sub> and O<sub>2</sub> respectively, was lower than to the rest of the stimuli. However, the cues A and B were also partially non-reinforced in the AX- and BY-compounds. If this was the cause for the lower scores the AD and BE, then participants' scores to the new compounds (DE and IJ) should not be lower than to other CSs because each of these stimuli was always reinforced.

Table 4. Design of Experiment 3.

Pavlovian phase	Summation test	Instrumental	PIT test A	PIT test B	
Pre-training	Inhibition				
A->O <sub>1</sub>	A->O <sub>1</sub>	FX/GY <i>Inhibitory</i>	R <sub>1</sub> ->O <sub>1</sub>	FX <i>Inhibitory</i>	FY <i>Unrelated</i>
B->O <sub>2</sub>	AX <sup>-</sup>	FC/GH <i>Pre-exp</i>	R <sub>2</sub> ->O <sub>2</sub>	GY <i>Inhibitory</i>	GX <i>Unrelated</i>
DE->O <sub>1</sub>	X <sup>-</sup>	FN <sub>1</sub> /GN <sub>2</sub> <i>Novel</i>		FC <i>Pre-exp</i>	F <i>Excitatory</i>
IJ->O <sub>2</sub>	B->O <sub>2</sub>	FY/GX <i>Unrelated</i>		GH <i>Pre-exp</i>	G <i>Excitatory</i>
	BY <sup>-</sup>			FN <sub>1</sub> <i>Novel</i>	X <i>Inhibitory</i>
	Y <sup>-</sup>			GN <sub>2</sub> <i>Novel</i>	Y <i>Inhibitory</i>
	CH <sup>-</sup>				
	C <sup>-</sup>				
	H				
	F->O <sub>1</sub>				
	G->O <sub>2</sub>				
	DE->O <sub>1</sub>				
	IJ->O <sub>2</sub>				

Note: A, B, C, D, E, F, G, H, I, J, X and Y: neutral fractal images; R<sub>1</sub> and R<sub>2</sub>: keyboard responses; O<sub>1</sub> and O<sub>2</sub>: food and drink images. - denotes no outcome.

### 2.10.1 Method

*Participants.* Twenty-four students participated in this Experiment (9 males and 15 females) aged between 18 and 36 years old.

*Procedure.* Everything was the same as in Experiment 2, unless otherwise stated.

*Pavlovian phase.* The *pre-training* was divided into 2 blocks, each with 4 trials of A->O<sub>1</sub>, B->O<sub>2</sub>, DE->O<sub>1</sub> and IJ->O<sub>2</sub>. The *inhibitory training* was divided into 3 blocks; the various types of trial A->O<sub>1</sub>, B->O<sub>2</sub>, DE->O<sub>1</sub>, IJ->O<sub>2</sub>, AX<sup>-</sup>, BY<sup>-</sup>, CH<sup>-</sup>, F->O<sub>1</sub>, G->O<sub>2</sub>, X<sup>-</sup>, Y<sup>-</sup>, C<sup>-</sup> and H<sup>-</sup> were

presented 3 times in each of the first 2 blocks, but only twice in the third block.

*Summation test.* The question "In a scale from 1 to 100, how likely is it that this image will be followed by FOOD/DRINK" was replaced by "How likely is that this image will be followed by FOOD/DRINK?" and at the extreme left of the scale the text was "very UNLIKELY" and on the extreme right "very LIKELY".

*Instrumental phase.*

*PIT tests.* The compounds FX, GY, FC, GH, FN<sub>1</sub> and GN<sub>2</sub> were presented in *Test A*, and F, G, X, Y, FY and GY in *Test B*. The tests were separated by the *instrumental re-training* and half of the participants received *Test A* before *Test B*, and the remainder the reverse.

### **2.10.2 Results**

*Pavlovian phase.* The mean scores of *inhibitory training* were grouped according to CS type and are presented in Figure 8. The graph suggests that participants improved their accuracy across the blocks. A/B and DE/IJ were also presented in pre-training which explains the higher scores for these trial types from the first block. An ANOVA with block (1-3) and CS (A/B, DE/IJ, AX/BY, CH, F/G, X/Y, C/H) as factors showed a significant main effect of block,  $F(2, 46) = 54.21, p < .001$ ,

$MSe = .04$ ,  $\eta_p^2 = .70$ , CS,  $F(6, 138) = 9.3$ ,  $p < .001$ ,  $MSe = .05$ ,  $\eta_p^2 = .29$ , and a significant CS x Block interaction,  $F(12, 276) = 4.53$ ,  $p < .001$ ,  $MSe = .03$ ,  $\eta_p^2 = .03$ . Simple main effects on the interaction revealed an effect of CS only on block 1,  $F(6, 18) = 15.3$ ,  $p < .001$ , largest  $F(6, 18) = 2.02$ ,  $p = .12$  in the third block.

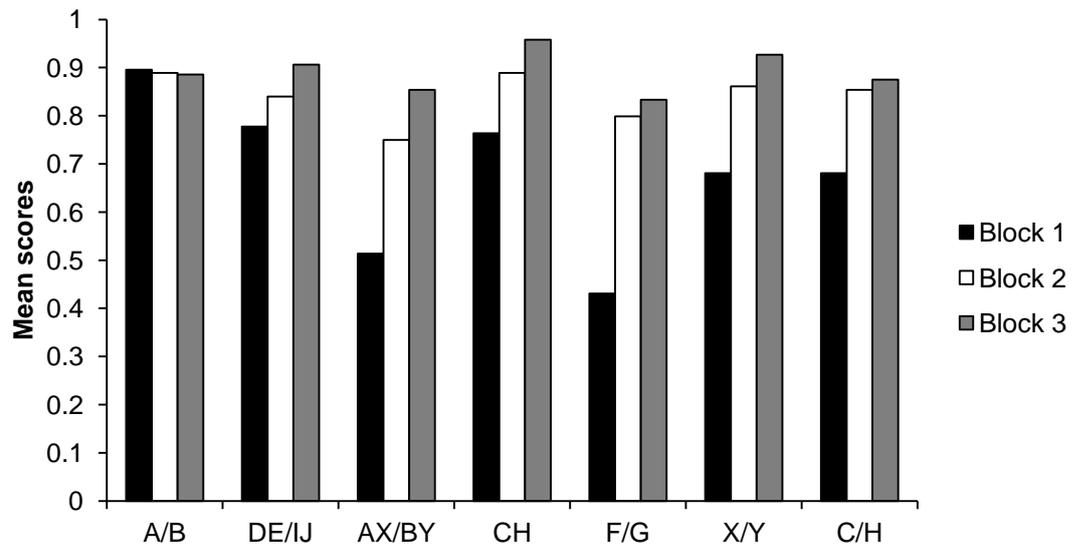
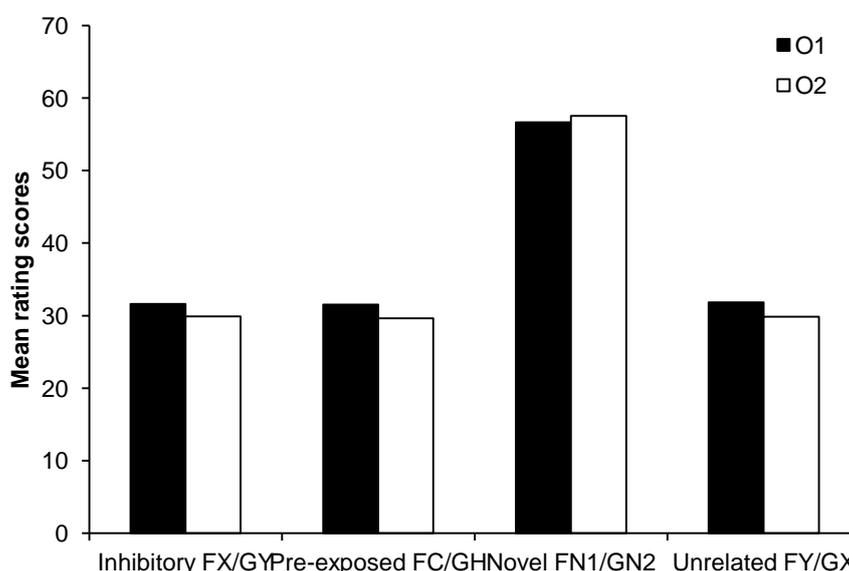


Figure 8. Mean scores grouped by CS and block in the Pavlovian phase of Experiment 3.

*Instrumental phase.* The mean numbers of responses made and outcomes delivered during the instrumental phase are presented in Table 2. The analysis showed a significant difference between  $R_1$  and  $R_2$ ,  $F(1, 23) = 4.29$ ,  $p = .05$ ,  $MSe = 3777$ ,  $\eta_p^2 = .16$ , but not between  $O_1$  and  $O_2$ ,  $F(1, 23) = 1.36$ ,  $p = .255$ ,  $MSe = 168.43$ . In order to rule out the possibility that this difference persisted in the PIT test, the number

of  $R_1$  and  $R_2$  responses was compared for each of the tests, but no difference was found, largest  $F(1, 23) = 1.79$ ,  $p = .19$ ,  $MSe = 1.65$ .

*Summation test.* The mean scores of the summation test were grouped and plotted in Figure 9, which suggests that participants gave higher scores to  $FN_1$  and  $GN_2$  than to the rest of the compounds, replicating the results of the previous experiment. An ANOVA with outcome ( $O_1$ ,  $O_2$ ) and compound (FX/FY, FC/GH,  $FN_1/GN_2$ , FY/GX) as factors showed a significant main effect of compound,  $F(3, 69) = 15.85$ ,  $p < .001$ ,  $MSe = .053$ ,  $\eta_p^2 = .41$ , and nothing else was significant,  $F_s < 1$ . The analysis of the significant main effect confirmed differences between  $FN_1/GN_2$  and FC/GH, FY/GX and FY/GX ( $p < .01$ ). Nothing else was significant.



*Figure 9.* Mean ratings of the likelihood of the occurrence of  $O_1$  during FX, FC,  $FN_1$  and FY, and of  $O_2$  during GY, GH,  $GN_2$  and GX in the summation test of Experiment 3.

*PIT tests.* The PIT scores of *Test A* are presented in the top panel of Figure 10. The graph shows the specific PIT effect for all the compounds, although it seems absent in the first block of the FX/GY trials. An ANOVA with block, congruence and compound (FC/GH, FX/GY, FN<sub>1</sub>/GN<sub>2</sub>) as factors showed a significant main effect of congruence,  $F(1, 23) = 5.83, p = .024, MSe = 108, \eta_p^2 = .20$ , but nothing else was significant, largest  $F(4, 92) = 2.44, p = .052, MSe = 3.65$ , for the Block x Compound interaction. To explore the possibility that the CIs had a transient effect, an ANOVA with congruence and compound as factors was conducted on the data of the first block only. This revealed a significant main effect of congruence,  $F(1, 23) = 7.95, p = .01, MSe = 44.2, \eta_p^2 = .26$ , and compound,  $F(2, 46) = 3.62, p = .035, MSe = 5.45, \eta_p^2 = .14$ , and a significant interaction,  $F(2, 46) = 3.25, p = .048, MSe = 17.1, \eta_p^2 = .12$ . Simple main effects on the interaction showed an effect of congruence on FC/GH,  $F(1, 23) = 14.03, p = .001$ , and FN<sub>1</sub>/GN<sub>2</sub>,  $F(1, 23) = 8.82, p = .007$ , but not on FX/GY,  $F < 1$ . This confirms that the CIs reduced the specific PIT, although this effect was not strong enough to continue across the test.

The PIT scores of *Test B* are presented in the bottom panel of Figure 10. The graph shows a clear specific PIT effect in the F/G trials, which seems smaller for FY/GX. In addition X and Y seem to produce more incongruent than congruent responses but only in the second block. An ANOVA with block, congruence and CS (F/G, FY/GX, X/Y) revealed a significant main effect of congruence,  $F(1, 23) = 5.8, p = .024, MSe = 50, \eta_p^2 = .20$ , and CS,  $F(2, 46) = 6.89, p = .002, MSe =$

10.04,  $\eta_p^2 = .23$ , a significant Block x CS interaction,  $F(4, 92) = 2.48$ ,  $p = .05$ ,  $MSe = 3.28$ ,  $\eta_p^2 = .10$ , and a significant Congruence x CS interaction,  $F(2, 46) = 3.84$ ,  $p = .029$ ,  $MSe = 32.58$ ,  $\eta_p^2 = .14$ . Nothing else was significant, largest  $F(4, 92) = 2.11$ ,  $p = .086$ ,  $MSe = 9.52$ . Simple main effects on the critical Congruence x CS interaction showed an effect of congruence on F/G trials,  $F(1, 23) = 5.53$ ,  $p = .028$ , but not on FY/GX,  $F(1, 23) = 3.72$ ,  $p = .07$ , or X/Y trials,  $F < 1$ . This confirms that the specific PIT effect produced by F/G was reduced by the CIs, even though these CIs were trained to predict the absence of an outcome different to that predicted by F and G.

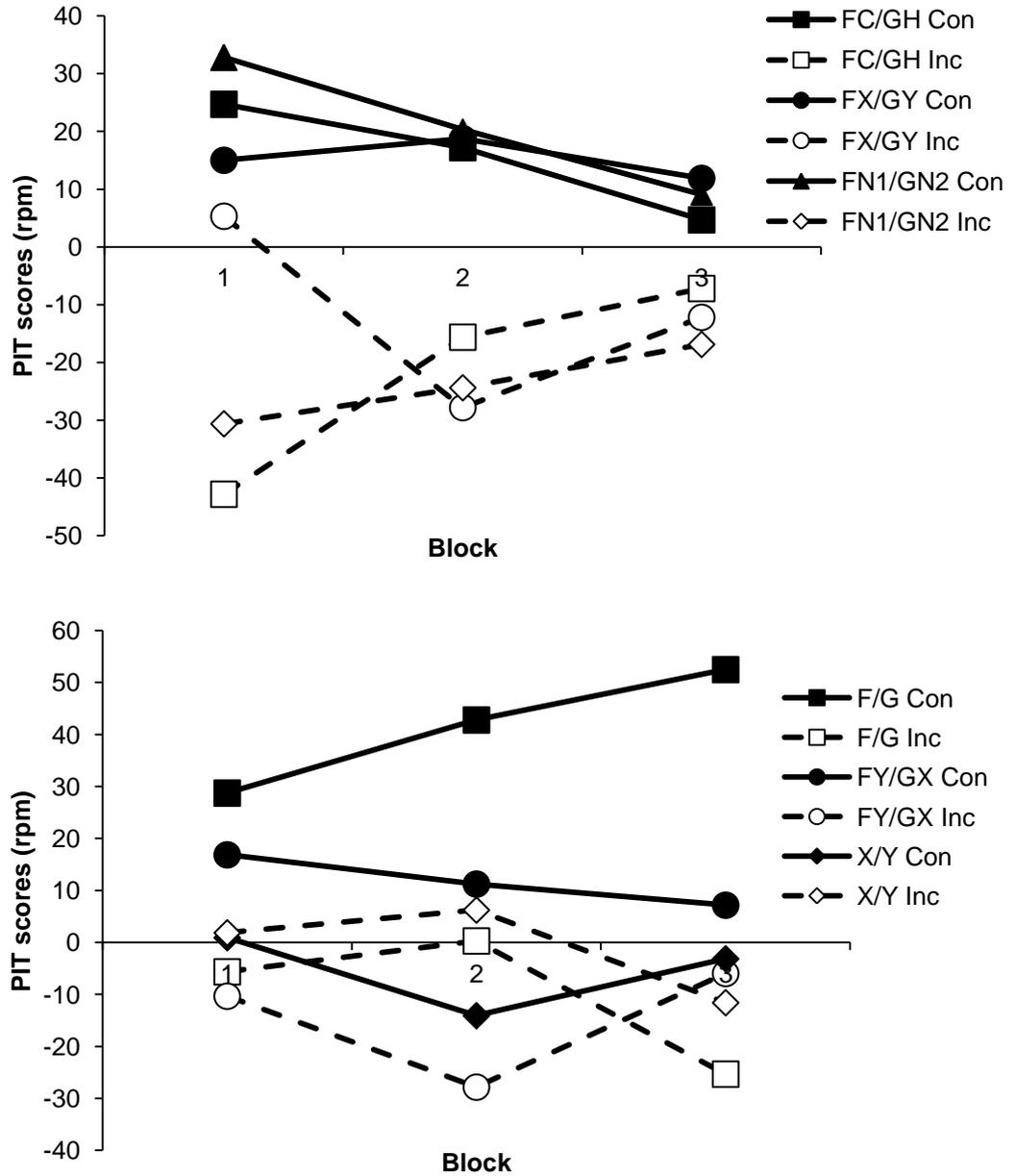
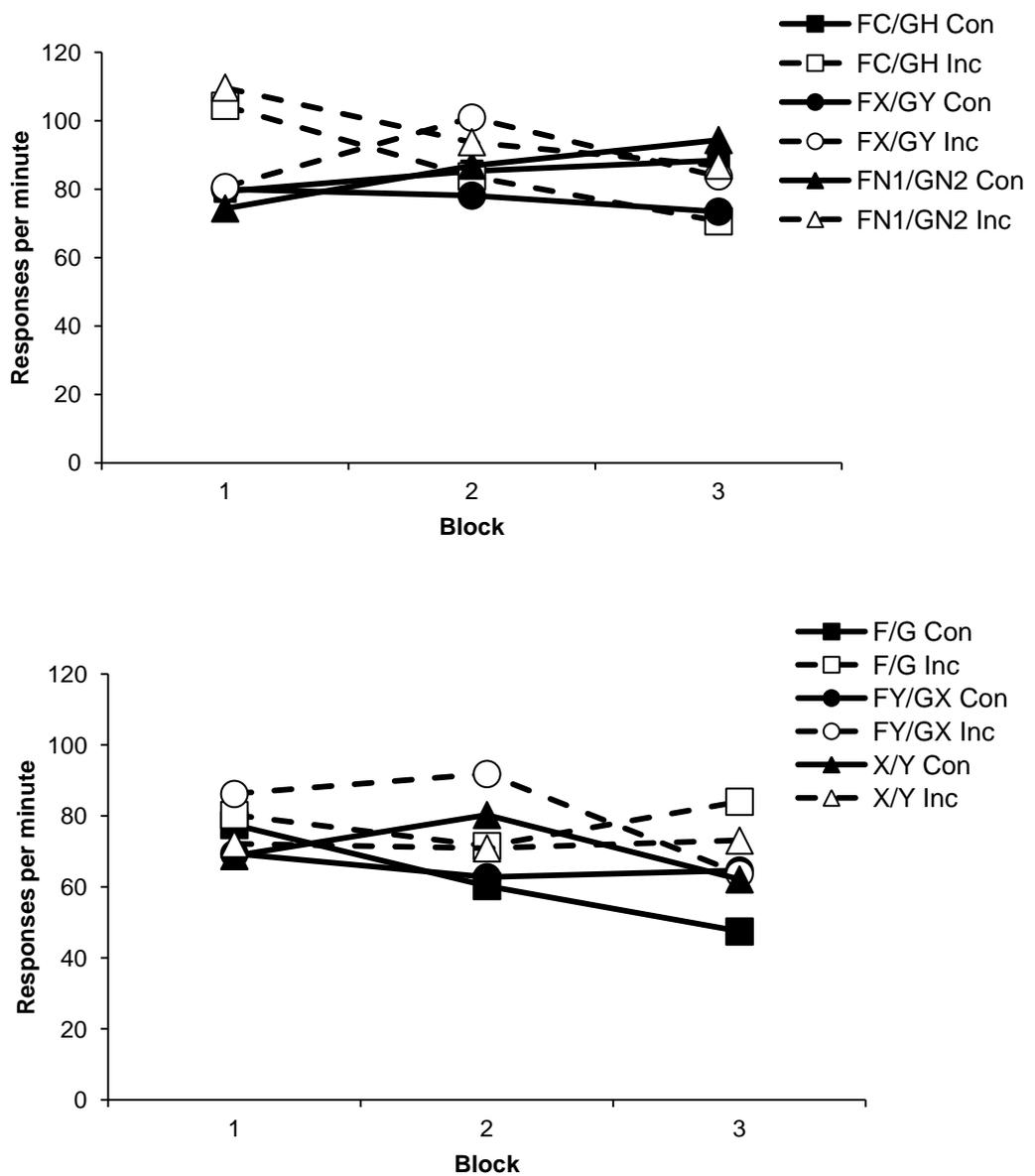


Figure 10. PIT scores in the Test A and B of Experiment 3. Top panel: FC/GH, FX/GY and FN1/GN2 in Test A. Bottom panel: F/G, FY/GX and X/Y in Test B.

The responses of the pre-CS of *Test A* and *B* were also analysed and presented in Figure 11. The statistical analyses showed no significant differences in any of the tests, largest  $F(1, 23) = 3.86$ ,  $p = .062$ ,  $MSe = 15.55$  for the main effect of congruence in *Test B*.



*Figure 11.* Mean rates of congruent and incongruent responding for each of the compounds in the pre-CS period of Test A and B of Experiment 3. Top panel: responses during the pre-CS of Test A. Bottom panel: responses during the pre-CS of Test B.

### 2.10.3 Discussion

The goal of this experiment was to assess the effect of the CIs in the PIT test when presented alone and in compound with F and G in the same task. As in Experiment 1, when X and Y were presented as single cues in the PIT test they appeared to produce the opposite to specific PIT, i.e. more incongruent than congruent responses, but this was not statistically significant (Test B). But when X and Y were presented with the CS+s that were trained with the same outcome, i.e. FX and GY, the specific PIT effect was smaller than that produced by the CS+s in compound with the control stimuli, i.e. FC, FN<sub>1</sub>, GH and GN<sub>2</sub>, although only in the first test block (Test A). The CIs were also presented with the CS+s trained with a different outcome, i.e. FY and GX, and these compounds produced a smaller PIT effect than the CS+s presented as single cues (Test B). Considering this test in isolation is possible to argue that FY and GX produced less specific PIT than F and G merely because in the FY and GX compounds F and G were accompanied by a non-reinforced stimulus, and not because X and Y were exerting any inhibitory properties. However, F and G were capable of producing the specific PIT effect when they were presented with the control stimuli (FC, GH, FN<sub>1</sub> and GN<sub>2</sub>) in Test A of this experiment and in the PIT test of Experiment 2, which suggests that the reduction in the specific PIT effect produced by FY/GX can be attributed to the inhibitory properties of the CIs. Nevertheless, it is not clear how a CI that signalled the absence of a particular outcome, e.g.

X- no  $O_1$ , could affect the specific PIT effect produced by a CS predictor of a different outcome, e.g.  $F \rightarrow O_2$ .

In the summation test participants' expectations of the outcomes were reduced by X and Y more than by the novel stimuli  $N_1$  and  $N_2$ , confirming the inhibitory properties of the CIs. However, the effect of X and Y did not differ from that of C and H. The idea that C and H might have acquired some inhibitory strength due to differential inhibition was considered and in this experiments the Pavlovian phase was conducted before instrumental training. This arrangement aimed to hinder the development of differential inhibition, but it did not affect the ability of C and H to reduce participants' expectations of the outcomes. This, together with the fact that these cues had no effect in the PIT test relative to the CIs, suggests that differential inhibition was not the cause of the results of the summation test.

## **2.11 General discussion**

These experiments aimed to establish a PIT task in which CIs could be tested to discriminate between the different accounts of the specific PIT effect. The results of these experiments confirmed that this task is effective in producing the specific PIT effect: CS+ presentations increased responses previously reinforced with the same outcome as the CS+, compared to responses reinforced with a different outcome. This specific PIT effect persisted when the CS+s

were presented in compound with either pre-exposed control stimuli or novel cues. However, the PIT effect was reduced when the CS+s were presented with the CIs, even if the CIs signalled the absence of a different outcome to that predicted by the CS+s.

The inhibitory properties of the CIs were independently assessed in the summation test. In all the experiments the participants' expectations of the outcomes elicited by the CS+s were reduced more by presenting the CIs than by presenting novel stimuli, confirming the inhibitory strength of the CIs. Nevertheless, the effect of the CIs was no different to that produced by cues that were simply pre-exposed during training, which was a more conservative control. One of the explanations considered was that the pre-exposed stimuli also acquired inhibitory strength during training due to differential inhibition. However, this idea was not supported by the fact that the pre-exposed cues did not affect the specific PIT effect as the CIs did. Moreover, Experiment 3 was arranged in a way that the excitatory strength of the context was reduced, which should have decreased the ease with which the control cues could become inhibitors. But even then the pre-exposed and the CIs produced a similar effect in the summation test. One possible interpretation is that the CIs acquired inhibitory strength but not enough to produce an effect distinguishable from that of the pre-exposed stimuli in the summation tests, perhaps due to a low sensitivity of this test.

The outcome-specificity of conditioned inhibition was also assessed in the summation test of Experiments 2 and 3. CIs that

predicted the absence of an outcome were presented with CS+s trained with a different outcome, and these CIs reduced participants' expectations of the outcomes more than the novel stimuli. These results suggest that in these experiments conditioned inhibition was not outcome-specific. However, this interpretation must be taken cautiously because, as mentioned above, the summation test might not be sensitive enough to detect small differences in the inhibitory properties of the stimuli. Although the CIs also reduced specific PIT when they were presented with CS+s that predicted a different outcome (Experiment 3), which suggests that the CIs were not outcome-specific, these results alone are not sufficient to conclude definitively that conditioned inhibition is not outcome-specific, and more research was conducted to address this issue (see Chapter 3).

Overall, the results provide evidence supporting some of the two-process accounts, which assume that both S-O and R-O associations are critical for the specific PIT effect. When a CS+ is presented at test, e.g. A that signals  $O_1$ , it activates an outcome representation, e.g.  $A \rightarrow [O_1]$ , which increases the response trained with the same outcome, e.g.  $A \rightarrow [O_1] \rightarrow R_1$ . Then if A is presented with a CI that suppresses activation of  $O_1$  (Holland & Lamarre, 1984; Rescorla & Holland, 1977), the specific PIT effect should be reduced, which was found in Experiment 2 and partially in Experiment 3.

However, the present experiments do not support the predictions of all two-process theory accounts. According to the expectancy version (Trapold and Overmier, 1972) the expectancy of

the outcome, triggered by the CS, elicits responding in a forward manner (O-R). This O-R association is formed during training, in which the response is usually preceded and followed by the same outcome. But in all these experiments the instrumental training was conducted concurrently: both responses were preceded equally often by  $O_1$  and  $O_2$ , i.e.  $O_1/O_2-R_1$  and  $O_1/O_2-R_2$ , but followed only by one of the outcomes, i.e.  $R_1-O_1$  and  $R_2-O_2$ . If the specific PIT effect is determined by O-R associations then each of the outcomes should elicit  $R_1$  and  $R_2$  equally, resulting in no specific PIT effect (or a general PIT effect). Although this does not mean that O-R associations never contribute to the PIT effect, it does imply that in the present task the effect must be determined by a bidirectional R-O associative structure (Rescorla & Colwill, 1989; Rescorla, 1992a; Rescorla, 1994b). Finally, these results cannot easily be explained by the S-R account. According to this account an association is formed between the CSs and the responses during training; thus the specific PIT effect is caused by the CSs directly eliciting the responses at test. Then presenting a CS+ together with a CI should not affect its ability to produce the specific PIT effect, even if this CI suppresses the activation of the outcome representation.

## Chapter III

### **Outcome-specificity of conditioned inhibition in the specific PIT effect.**

#### **3.1 Overview**

The results of Experiment 3 suggest that conditioned inhibition is not outcome-specific. The CIs reduced the specific PIT effect, and this occurred even when they were presented with a CS+ trained with a different outcome. Furthermore, the results of the summation tests also suggest that the CIs were not outcome-specific in these experiments - although the reliability of this conclusion is tempered by the fact that the inhibitory properties of the CIs did not differ from those of the pre-exposed control stimuli in this test. This series of experiments used a simplified version of the task and aimed to replicate the effect of the CIs on the specific PIT effect, and also to further assess the specificity of conditioned inhibition. It was expected that this new task might be more sensitive in detecting differences both in the summation and PIT tests. In Experiments 4, 5 and 6, a CI trained to signal the absence of one outcome was presented in the PIT test together with a CS+ predicting the same outcome and with a CS+ trained with a different outcome. In Experiment 4 and 5 the Pavlovian phase was conducted prior to instrumental training, while in Experiment 6 the order of the phases was reversed. Experiment 7 and 8 assessed alternative explanations of the effect of the CI in the PIT test: Experiment 7 aimed to test the possibility that the CI acted by

eliciting competing responses and Experiment 8 that the CI activated a competing neutral outcome representation.

### 3.2 Introduction

The experiments presented in the previous chapter support the idea that the specific PIT effect is mediated by the activation of an outcome representation at test and that this effect can be reduced by a CI. But the results of Experiment 3 showed that a CI signalling the absence of a particular outcome (e.g. X->no  $O_1$ ) reduced the specific PIT effect produced by a CS+ predicting a different outcome (e.g. F-> $O_1$ ). These results suggest that a CI can influence the specific PIT effect regardless of the US with which it was trained, which is problematic for the S-O-R accounts. According to these accounts a CS+ activates an outcome representation that elevates performance of those responses reinforced with the same outcome. But this representation must encode specific information about the sensory properties of the outcomes. For instance, in an experiment in which two responses are trained ( $R_1$ -> $O_1$ ;  $R_2$ -> $O_2$ ) and two CSs are paired each with one outcome (F-> $O_1$ ; G-> $O_2$ ), F will elevate performance of  $R_1$  and not  $R_2$  only if it activates an outcome representation with the aspects of  $O_1$  that are different from those of  $O_2$ . If F activates a representation of the common elements of both  $O_1$  and  $O_2$  it would increase performance of both  $R_1$  and  $R_2$ , i.e. general PIT. For this

reason it is not clear how a CI that is not specific to the features that discriminate the outcomes, but instead acts on their common elements, can reduce the specific PIT effect. Moreover, these results are inconsistent with those reported by Laurent and colleagues (2014) described in the previous chapter. The authors found that a CI predicting the absence of an outcome ( $X \rightarrow \text{no } O_1$ ) abolished the specific PIT effect produced by a CS+ that signalled the same outcome ( $A \rightarrow O_1$ ), but had no effect when presented with a CS+ predicting a different outcome ( $B \rightarrow O_2$ ).

It is possible to explain these results by making two assumptions. One is that in the experiments reported in Chapter 2 the CIs did not suppress a sensory representation of the US as proposed by Konorski (1947, 1967), but rather suppressed the motivational state elicited by the outcomes. The outcomes used in these experiments were pictures of food and drinks, which have different sensory properties but may have a similar motivational value. Then the CIs might become associated with the motivational state elicited by the absence of the outcomes. A parallel can be made in research regarding the blocking effect (Kamin, 1969). In this effect, one group of subjects initially receives CS-US pairings, i.e.  $CS_1 \rightarrow O_1$ , and in a second stage  $CS_1$  is presented in compound with a second CS ( $CS_2$ ), compound that is also paired with the US, i.e.  $CS_1CS_2 \rightarrow O_1$ . Evidence indicates that learning about  $CS_2$  is impaired, that is,  $CS_2$  produces less CRs, in a subsequent test, relative to a group that did not receive the initial  $CS_1$ -US pairings. One of the explanation of this effect is that

in the second stage the US is already predicted by CS<sub>1</sub>, which results in less learning about the CS<sub>2</sub>-US relation (Kamin, 1969; Rescorla & Wagner, 1972). Consistent with this interpretation, it has been found that changing the magnitude of the US in the second stage results in an unblocking effect, that is, if the US presented in the second stage is not fully predicted by CS<sub>1</sub>, then CS<sub>2</sub> acquires excitatory strength, being capable of eliciting CRs at test (Kamin, 1969). However, it has also been observed that changing the US in the second stage, e.g. CS<sub>1</sub>->Food, CS<sub>1</sub>CS<sub>2</sub>->Water, does not affect the blocking effect (Ganesan & Pearce, 1988). These results seem to be inconsistent with the idea that the blocking effect is due to the lack of predictive value of CS<sub>2</sub>, it can be interpreted by Konorski's conceptualization of Pavlovian conditioning (Konorski, 1948, 1967). Although both food and water have distinct sensory properties, they share a similar motivational value. Therefore, it is possible that in the second stage CS<sub>1</sub> activates a motivational US representation of the US, making CS<sub>2</sub> redundant and thus explaining the blocking effect found by Ganesan and Pearce (1988). An analogous mechanism can be considered in the case of inhibition. If a CI can suppress a motivational representation of the US, then it can explain the results of the PIT tests.

The second assumption is that the specific PIT effect is mainly determined by the sensory outcome representation, but it also needs a certain motivational state in order to be observed. For example, the representation of an outcome provides information about the response that was reinforced with that outcome, but the response may not be

performed if the outcome is no longer desired. This is because the R-O association encodes not only sensory information about the outcome but also depends on the motivational value of the reinforcer (Adams & Dickinson, 1981; Balleine & Dickinson, 1998). Thus, if in the PIT test the CIs elicit a motivational state antagonistic to that required for production of the instrumental responses then they might reduce the specific PIT effect otherwise produced by the CS+s.

The main challenge to this interpretation is the evidence from the studies that have used different techniques to reduce the motivational value of the outcomes. As was described in Chapter 1, it has been found that devaluing the outcomes before the test does not affect the specific PIT effect (e.g. Holland, 2004). Particularly, the specific PIT effect has been found even after inducing satiety before the test (e.g. Corbit et al., 2007; Watson et al. 2014), which should have an impact on subjects' motivational state. However, the evidence also indicates that outcome devaluation is outcome-specific. For instance, if two outcomes are used to train two responses ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ) and then one of the outcomes is devalued (e.g.,  $O_1$ ), this will only reduce performance of the response trained with the devalued outcome (i.e.  $R_1$ ), when instrumental behaviour is measured in the absence of Pavlovian stimuli (i.e. extinction test). Importantly, this occurs even if both outcomes have the same motivational valence, which implies that the outcome devaluation affects performance via the sensory components of the outcome representation. One possibility is that the effect of outcome devaluation is different from

that of the CIs trained in the experiments presented here. For this reason it is important to confirm that the effect of conditioned inhibition on the specific PIT effect is not outcome-specific. In the experiments described in this chapter a CI that signalled the absence of an outcome, e.g.  $O_1$ , was presented in the PIT test with a CS+ predictor of the same outcome and also with a CS+ that signalled deliveries of a different outcome, e.g.  $O_2$ . This procedure aimed to compare the outcome-specificity of conditioned inhibition in the same test.

The results of the summation test in the previous experiments are also ambiguous. Although the results of Experiments 2 and 3 showed that the CIs reduced participants' expectations of the outcomes even if they were paired with CS+s that predicted different outcomes, suggesting a form of conditioned inhibition that was not outcome-specific, the effect of these CIs was not different that of control stimuli that were pre-exposed during training. This raises the possibility that the summation test was not sensitive enough to discriminate between the inhibitory properties of the CIs, their US-specificity and the effect of non-reinforced stimuli. For this reason, the following series of experiments focused on establish a CI strong enough to reduce responding in the summation test compared to pre-exposed control stimuli. To achieve this, the task was simplified in such a way that only one CI was established to predict the absence of one outcome ( $X \rightarrow \text{no } O_1$ ), and its effect was assessed by presenting it with a CS+ trained with the same outcome ( $F \rightarrow O_1$ ) and a CS+ trained with a different outcome ( $G \rightarrow O_2$ ). Thus the aim of this chapter was to

provide further evidence on the effect of conditioned inhibition in the specific PIT effect and to confirm whether inhibition is outcome-specific or not in both summation and PIT tests. In Experiment 4 and 5, the Pavlovian phase was followed by instrumental training while in Experiment 6, the order of the phases was reversed. In these three experiments the CIs were presented in compound with CS+s in the PIT test. Experiments 7 and 8 were conducted in order to reject alternative explanations of the effect of the CIs in the PIT test.

### **3.3 Experiment 4**

This experiment used a simplified version of the task used in Experiments 2 and 3 (see Table 5). In the Pavlovian phase, participants initially received pairings of A and AB with  $O_1$  ( $A \rightarrow O_1$ ,  $AB \rightarrow O_1$ ), followed by non-reinforced presentations of  $AX^-$  and  $CH^-$ . This arrangement should allow X to develop inhibitory properties while keeping C and H as neutral cues. After this, F was paired with  $O_1$  and G with the novel  $O_2$  ( $F \rightarrow O_1$ ,  $G \rightarrow O_2$ ). The purpose of training  $G \rightarrow O_2$  was to use G in the summation test that followed to assess the outcome-specificity of conditioned inhibition. In this test participants rated the likelihood of  $O_1$  in the presence of FX and FC, and  $O_2$  in the presence of GX and GH. If X became a CI it was expected to lower the ratings of FX compared to those of FC. Moreover, as X was trained to predict the absence of  $O_1$ , if inhibition is outcome-specific then no difference

should be found between the ratings to GH and GX; otherwise ratings to GX were expected to be lower than to GH.

In the instrumental training that followed, two responses were reinforced, each of them with one of the outcomes ( $R_1 \rightarrow O_1$ ,  $R_2 \rightarrow O_2$ ), after which the PIT test was conducted. In this test, participants had the chance to perform both  $R_1$  and  $R_2$  as much as they wanted in the presence and absence of FX, FC, GX and GH. It was expected that the specific PIT effect would be found in the presence of FC but not FX. In addition, if conditioned inhibition is outcome-specific, then both GH and GX should produce the specific PIT effect; but if the CIs act on the common motivational elements of  $O_1$  and  $O_2$ , then X should reduce the specific PIT produced by F as effectively as that produced by G.

Table 5. Design of Experiment 4.

	Pavlovian phase		Summation	Instrumental	PIT test
Pre-training	Inhibition	Exciters			
A- $\rightarrow$ O <sub>1</sub>	A- $\rightarrow$ O <sub>1</sub>	F- $\rightarrow$ O <sub>1</sub>	FX <i>Inhibitory</i>	R <sub>1</sub> - $\rightarrow$ O <sub>1</sub>	FX <i>Inhibitory</i>
AB- $\rightarrow$ O <sub>1</sub>	AB- $\rightarrow$ O <sub>1</sub>	G- $\rightarrow$ O <sub>2</sub>	FC Pre- <i>exp</i>	R <sub>2</sub> - $\rightarrow$ O <sub>2</sub>	FC Pre- <i>exp</i>
	AX <sup>-</sup>		GX <i>Unrelated</i>		GX <i>Unrelated</i>
	CH <sup>-</sup>		GH Pre- <i>exp</i>		GH Pre- <i>exp</i>

Note: A, B, C, F, G, H and X: neutral fractal images; R<sub>1</sub> and R<sub>2</sub>:

keyboard responses; O<sub>1</sub> and O<sub>2</sub>: food and drink images. - denotes no outcome.

### 3.3.1 Method

*Participants.* Twenty-four students from the University of Nottingham and CUNY-Brooklyn College (USA) participated in this experiment (5 males, 19 females) aged between 18 and 24 years old.

*Apparatus and materials.* Seven fractal images were used as CSs. In the non-reinforced trials a white square was presented for the same period as the outcomes. The CSs, the white square and the outcomes had the same size (8 x 8 cm). The CSs appeared either on the left, right or both sides of the screen, leaving an 8 cm space at the centre of the screen, in which the outcome/white square was positioned. For half of the participants  $O_1$  was food images and  $O_2$  drink images, while the opposite was true for the other half. For all the participants the 'z' key was reinforced with drink images and 'm' with food images, thus for half of the participants 'z' was  $R_1$  and 'm'  $R_2$ , and the reverse for the other half. A and B were counterbalanced with each other, as were X and C and also H, F and G, resulting in 12 counterbalancing conditions.

*Procedure.* Everything was the same as in Experiment 3 unless otherwise stated.

*Pavlovian phase.* This stage was divided in 3 parts, each of them comprising one block. In the *pre-training* stage participants received 4 trials of A- $O_1$  and AB- $O_1$ , while in the *inhibitory training* stage that

followed they received the same trials as in the pre-training plus 8 additional trials of A-O<sub>1</sub> and AB-O<sub>1</sub> and 12 trials of AX and CH. After this, in the *excitor training* stage 8 trials of F-O<sub>1</sub> and G-O<sub>2</sub> were presented.

Each trial began with the fixation dot followed by a CS with the question "*Which reward will appear now?*" above it and a text with the possible answers below it (food, drink or nothing). In the *pre-training* and *inhibitory training* stages the text was "1) Food 5) Nothing" when O<sub>1</sub> was food images and "5) Nothing 9) Drink" when O<sub>1</sub> was drink images. In the *test excitor training* stage O<sub>2</sub> trials were introduced, so the text was "1) Food 5) Nothing 9) Drink".

After participants pressed a number on the keyboard (1, 5 or 9), the text was removed but the CS(s) remained on the screen while the corresponding outcome (or white square) appeared with a feedback message. If the answer was correct the text "Correct!" was presented at the top in green letter, otherwise the text "Oops! That was wrong" was positioned at the bottom in red letters. After 2 seconds the CS(s), outcome and feedback were replaced by the fixation dot starting a new trial.

*Summation test.* The test was divided in two blocks, each of them with 2 trials of FC, FX, GH and GX.

*Instrumental training.* The texts "*Drink = 0*" in orange and "*Food = 0*" in blue were positioned at the top left and right of the screen, respectively. Each time an outcome was delivered, the corresponding

number was increased by 1. The phase ended when participants received at least 50 deliveries of each outcome.

*PIT test.* The fixation cross was present throughout the entire test. The pre-CS was 2 seconds in duration, followed by the CSs presented on the left and right of the cross for 2 more seconds. The test was divided into 4 blocks, each of them with 2 trials of FC, FX, GH and GX.

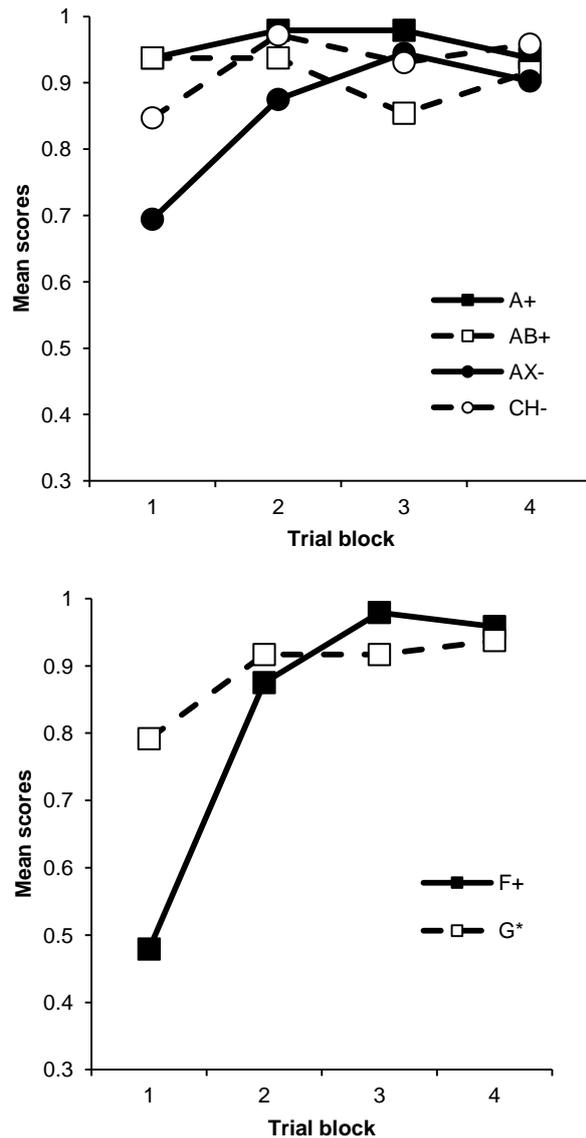
*Statistical Analysis.* Because the results of the previous experiments suggested that the CI has only a transient effect on the PIT test, only two blocks were analysed.

### **3.3.2 Results**

*Pavlovian phase.* The mean scores of the *inhibitory training* stage were grouped in 4-trial blocks comprising 2 A, 2 AB, 3 AX and 3 CH trials each, and they are presented in the top panel of Figure 12. The Figure suggests that participants rapidly learned to predict the outcome for all the CSs, although learning seems to be slower for AX. An ANOVA with trial type (A, AB, AX, CH) and Block (1-4) showed a significant main effect of trial type,  $F(3, 69) = 3.68, p = .016, MSe = .05, \eta_p^2 = .14$ , block,  $F(3, 69) = 6.41, p = .001, MSe = .023, \eta_p^2 = .22$ , and a significant interaction,  $F(9, 207) = 4.38, p < .001, MSe = .020, \eta_p^2 = .16$ . Simple main effects on the interaction revealed a significant difference between the CSs only on block 1,  $F(3, 21) = 7.54, p = .001,$

in which the scores of AX were significantly lower than those of A, AB and CH ( $p < .01$ ). This might be caused by A being constantly reinforced in the absence of X in the *pre-training* stage, producing difficulties for the participants at the beginning of the *inhibitory training* stage.

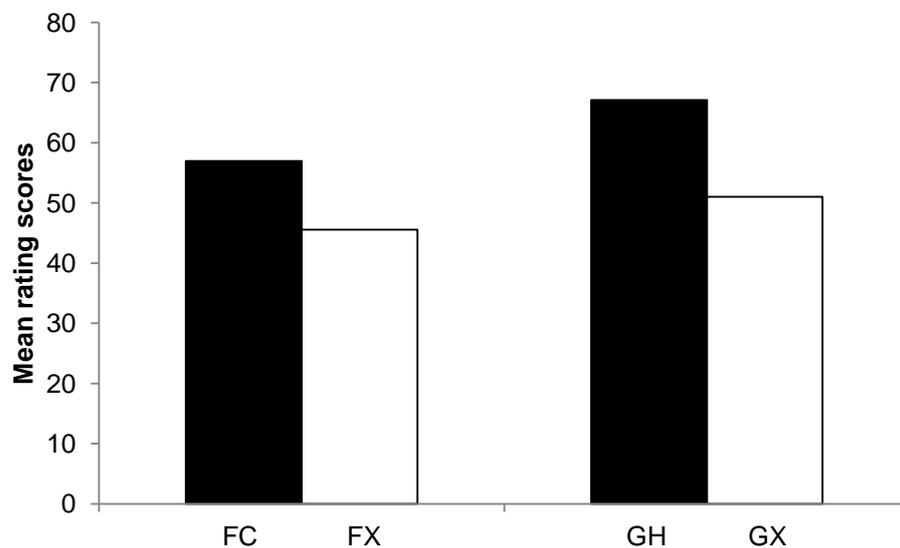
A similar analysis was conducted on the *excitor training* stage. The mean scores for F and G were grouped in 4 blocks of 2 trials each and they are plotted in the bottom panel of Figure 12. The graph shows an initially better performance to G than to F, but this difference disappeared by the second block. An ANOVA with trial type (F, G) and block (1-4) showed a significant main effect of trial type,  $F(1, 23) = 4.77, p = .039, MSe = .046, \eta_p^2 = .17$ , block,  $F(3, 69) = 24.74, p < .001, MSe = .043, \eta_p^2 = .52$ , and a significant interaction,  $F(3, 69) = 10.16, p < .001, MSe = .034, \eta_p^2 = .16, \eta_p^2 = .31$ . The analysis of the interaction confirmed a significant difference between F and G only on block 1,  $F(1, 23) = 13.8, p = .001$ . This difference in the scores to F and G might be caused by a pre-exposure effect to  $O_1$ , that is, that participants had previous experience to  $O_1$  but not to  $O_2$ , which was introduced at this stage. It has been found that a CS paired with a pre-exposed outcome requires more trials to elicit CRs relative to a CS paired with a novel outcome (Randich & LoLordo, 1979), which is analogous to F- $O_1$  and G- $O_2$ .



*Figure 12.* Mean scores grouped by CS and block in the Pavlovian phase of Experiment 4. Top panel: the mean scores to the CSs in the inhibitory training. Bottom panel: the mean scores to the CSs in the test excitator training.

*Summation test.* The mean scores of the summation test are presented in Figure 13. The graph suggests that participants' expectancies of the outcomes were reduced when X was presented with F and G, compared to when they were presented with C or H. An

ANOVA with CS (F, G) and trial type (C/H, X) showed a significant main effect of CS,  $F(1, 23) = 4.94$ ,  $p = .036$ ,  $MSe = .03$ ,  $\eta_p^2 = .18$ , and a significant main effect of trial type,  $F(1, 23) = 9.05$ ,  $p = .006$ ,  $MSe = .05$ ,  $\eta_p^2 = .28$ . The interaction was not significant,  $F < 1$ . These results confirmed X as a CI and also support the non-specificity of its conditioned inhibition: X was trained to predict the absence of O<sub>1</sub> but it reduced the expectancy of O<sub>1</sub> and O<sub>2</sub>, which is consistent with the results of the previous experiments.



*Figure 13.* Mean ratings of the likelihood of O<sub>1</sub> occurrence during FC and FX, and O<sub>2</sub> occurrence during GH and GX in the summation test of Experiment 4.

*Instrumental training.* All participants completed the instrumental phase successfully. The mean number of responses and outcome presentations are presented in Table 6. No differences were found

between type of response made ( $R_1$ ;  $R_2$ ) nor type of outcome received ( $O_1$ ;  $O_2$ ) in the instrumental training,  $F_s < 1$ .

*Table 6.* Mean number of R1 and R2 responses and mean number of O1 and O2 deliveries in the instrumental phases of Experiments 4, 5, 6 and 8.

	$R_1$	$R_2$	$O_1$	$O_2$
<b>Experiment 4</b>	261.5	268.3	52.6	52.5
<b>Experiment 5</b>	268.1	264.2	51.9	53.6
<b>Experiment 6</b>	252	254.2	50.6	51.1
<b>Experiment 8</b>	261.2	261.7	53.2	53.2

*PIT test.* The PIT scores are presented in Figure 14. The scores to FC and FX are plotted in the top panel, which shows a clear specific PIT for FC but the opposite for FX, in which more incongruent responses were performed in the first block. The left panel of Figure shows the PIT scores to GH and GX, suggesting the presence of the specific PIT effect, i.e. more congruent than incongruent responses, but it seems to be reduced in the case of GX. An ANOVA with block (1, 2), congruence, CS (F, G) and trial type (C, X) revealed a significant main effect of congruence,  $F(1, 23) = 5.97$ ,  $p = .023$ ,  $MSe = 27.95$ ,  $\eta_p^2 = .21$ , a significant Block x Congruence interaction,  $F(1, 23) = 6.05$ ,  $p = .022$ ,  $MSe = 12.59$ ,  $\eta_p^2 = .21$ , and a Congruence x Trial type interaction,  $F(1, 23) = 6.1$ ,  $p = .021$ ,  $MSe = 16.91$ ,  $\eta_p^2 = .21$ . The

analysis of the critical Congruence x Trial type interaction showed a significant effect of congruence on C/H trials,  $F(1, 23) = 9.12, p = .006$ , but not on X trials,  $F < 1$ . This confirms the specific PIT effect in the case of FC and GH but not of FX and GX.

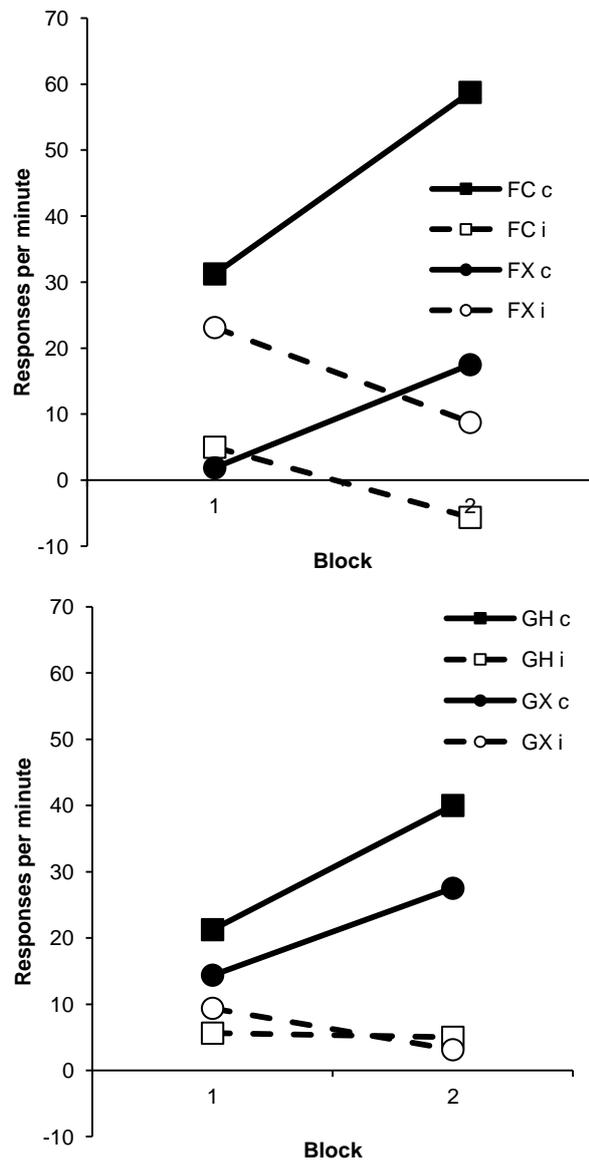


Figure 14. Mean rate of congruent and incongruent responses for each type of CS in the PIT Test of Experiment 4. Top panel: PIT scores for FC and FX. Bottom panel: PIT scores for GH and GX.

The responses during the pre-CS period were also analysed to ensure that no differences in the baseline were present (Table 7). A similar ANOVA showed no significant main effect or interaction, largest  $F(1, 23) = 3.97, p = .058, MSe = 1.64, \eta_p^2 = .15$  for the main effect of congruence.

*Table 7.* Mean preCS response rates in each block of the PIT test of Experiment 4.

	Congruent				Incongruent			
	FC	FX	GH	GX	FC	FX	GH	GX
<b>1</b>	49	49	48	49	42	36	47	44
<b>2</b>	45	37	48	49	51	35	38.8	49.4

### 3.3.3 Discussion

The results confirm the suggestion that a CI can reduce the specific PIT effect otherwise produced by a CS+. F selectively elevated performance when presented with the control cue C but not with X. Moreover, in the presence of FX the specific PIT effect was numerically reversed in the first block of test. This is consistent with the S-O-R accounts that attribute the specific PIT effect to the ability of the CS to activate a representation of the outcome, which in turns elicits the response previously reinforced with that outcome. If the CI

can suppress such a representation, then the specific PIT effect should be reduced.

The results of the PIT test also suggest that the effect of the CI might not be outcome-specific. The PIT effect produced by G, which predicted  $O_2$ , was also reduced by presentations of X, which signalled the absence of  $O_1$ . Because the specific PIT effect must be determined by the sensory properties of the outcomes, it is not clear how a CI can affect it in a general manner, as these results suggest.

The inhibitory properties of X were independently assessed in the summation test. Unlike in the previous experiments, this time X reduced the expectancies of  $O_1$  produced by F more than C. Furthermore, X also reduced the expectancies of  $O_2$  produced by G, even though X was only trained to predict the absence of  $O_1$ . This evidence supports the idea that conditioned inhibition is not outcome-specific in these experiments.

### **3.4 Experiment 5**

The aim of this experiment was to replicate the results of Experiment 4. The design was identical to Experiment 4 except that single presentations of X were included in the *inhibitory training* stage in order to increase the inhibitory strength of the CI. Single presentations of C were also included in that stage to match the

number of X trials. Because in this experiment C but not H had similar training to X, in the summation and PIT test H was replaced by C.

### **3.4.1 Method**

*Participants.* 32 students from the University of Nottingham participated in this experiment (12 males and 20 females) aged between 17 and 39 years old. One participant was excluded because he performed only one response in the instrumental phase.

*Procedure.* Everything was the same as in Experiment 4 unless otherwise stated. In this experiment F and G were counterbalanced with each other, as were A and B, and C and X, resulting in 8 counterbalancing conditions.

*Pavlovian phase.* The *inhibitory training* was divided in two blocks, each comprising 4 trials of A-O<sub>1</sub> and AB-O<sub>1</sub>, 6 trials of AX<sup>-</sup> and CH<sup>-</sup> and 3 trials of X<sup>-</sup> and C<sup>-</sup>.

*Summation test.* Both F and G were presented with X and with C (FC, FX, GC, GX).

*Instrumental training.*

*PIT test.* The compounds presented in this test were FC, FX, GC and GX. After the CS presentation an additional 1-s ITI period was included. This ITI period was exactly the same as the pre-CS period.

However, the responses performed during the ITI were not considered in the analysis. This was implemented in order to reduce the interference of post-CS responding on the baseline level, i.e. pre-CS responding.

### 3.4.2 Results

*Pavlovian phase.* The mean scores of *inhibitory training* are presented on the top panel of Figure 15 and as in Experiment 4 the scores for AX seemed to be lower than the rest at the beginning of training. An ANOVA with trial type (A, AB, AX, CH) and block (1-4) as factors revealed a significant main effect of trial type,  $F(3, 90) = 23.11, p < .001, MSe = .019, \eta_p^2 = .44$ , block,  $F(3, 90) = 24.64, p < .001, MSe = .13, \eta_p^2 = .45$ , and a significant interaction,  $F(9, 270) = 16.43, p < .001, MSe = .012, \eta_p^2 = .35$ . The analysis of the interaction showed a significant effect of trial type only at block 1,  $F(3, 28) = 28.73, p < .001$ , in which the scores for AX were significantly lower than those for A, AB and CH,  $p < .001$ . Nothing else was significant, largest  $F(2, 29) = 2.5, p = .1$ .

The mean scores of *excitor training* are plotted on the bottom panel of Figure 15. Again the scores to F were lower than to G at the beginning of training. An ANOVA with trial type (F, G) and block (1-4) showed a significant main effect of trial type,  $F(1, 30) = 37.92, p < .001, MSe = .024, \eta_p^2 = .56$ , block,  $F(3, 90) = 65.26, p < .001, MSe =$

.026,  $\eta_p^2 = .69$ , and a significant interaction,  $F(3, 90) = 27.66$ ,  $p < .001$ ,  $MSe = .022$ ,  $\eta_p^2 = .48$ . Simple main effects on the interaction showed a significant effect of trial type only at block 1,  $F(1, 30) = 46.09$ ,  $p < .001$ . Nothing else was significant.

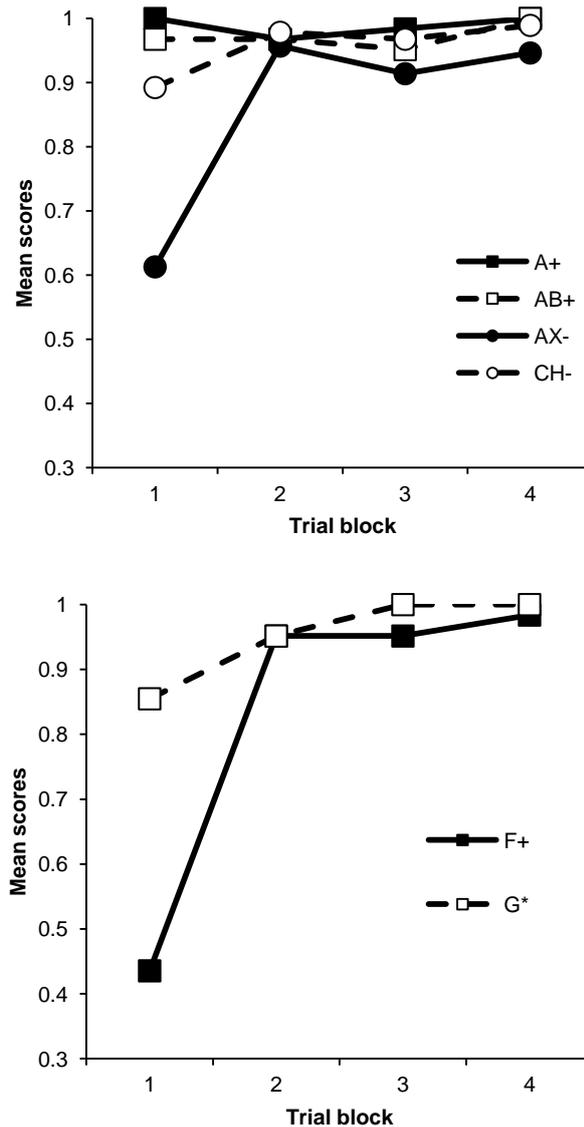
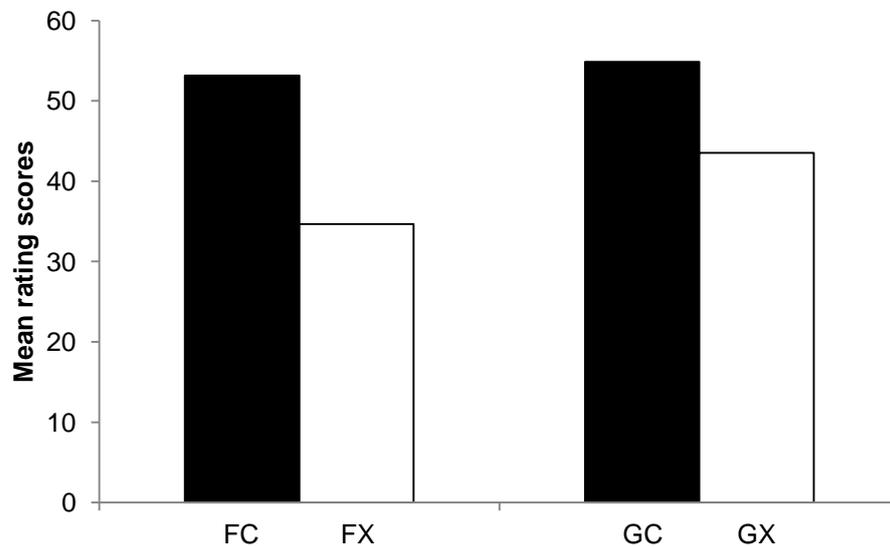


Figure 15. Mean scores grouped by CS and block in the Pavlovian phase of Experiment 5. Top panel: the mean scores to the CSs in the inhibitory training. Bottom panel: the mean scores to the CSs in the test excitator training.

*Summation test.* The mean scores of the test are plotted in Figure 16, which suggests lower rating scores to FX and GX than to FC and GC. An ANOVA with CS (F, G) and trial type (C, X) revealed a significant main effect of trial type,  $F(1, 30) = 10.06$ ,  $p = 0.003$ ,  $MSe = 0.07$ ,  $\eta_p^2 = .25$ . Nothing else was significant, largest  $F(1, 30) = 2.78$ ,  $p = 0.11$ ,  $MSe = 0.03$ . These results replicate the findings of Experiment 4, confirming the inhibitory properties of X and supporting the non-specificity of conditioned inhibition in this procedure.



*Figure 16.* Mean ratings of the likelihood of O1 occurrence during FC and FX, and O2 occurrence during GC and GX in the summation test of Experiment 5.

*Instrumental training.* The participants completed the phase successfully (Table 6), and no significant difference between the type of responses made ( $R_1$ ;  $R_2$ ) or outcomes delivered ( $O_1$ ;  $O_2$ ) was found,  $F_s < 1$ .

*PIT test.* The PIT scores are presented in Figure 17, which suggests that the specific PIT effect was present and steady for FC and GC, but was abolished in the first block for FX and GX. Importantly, it seems that X reduced the specific PIT effect for F and G equally, regardless of the fact that G predicted a different outcome. An ANOVA with block, congruence, CS (F, G) and trial type (C, X) revealed a significant main effect of congruence,  $F(1, 30) = 13.89, p = .001, MSe = 43.73, \eta_p^2 = .32$ , a significant Block x Congruence interaction,  $F(1, 30) = 5.59, p = .025, MSe = 19.59, \eta_p^2 = .16$ , and a significant Block x Congruence x Trial type interaction,  $F(1, 30) = 8.77, p = .006, MSe = 15.42, \eta_p^2 = .23$ . To analyse the triple interaction additional two-way ANOVAs for C and for X were conducted. The analysis of C revealed a significant main effect of congruence,  $F(1, 30) = 17.51, p < .001, MSe = 17.67, \eta_p^2 = .37$ . Nothing else was significant, largest  $F(1, 30) = 1.68, p = .204, MSe = 10.75$ . This confirms the specific PIT effect on the FC and GC trials. In contrast, the analysis of X showed a significant main effect of congruence,  $F(1, 30) = 8.76, p = .006, MSe = 34.06$ , which interacted with block,  $F(1, 30) = 12.15, p = .002, MSe = 20.09$ . Simple main effects on the interaction showed a significant congruence effect on block 2,  $F(1, 30) = 22.53, p < .001$ , but not on block 1,  $F < 1$ , confirming the absence of the specific PIT effect at the beginning of the test.

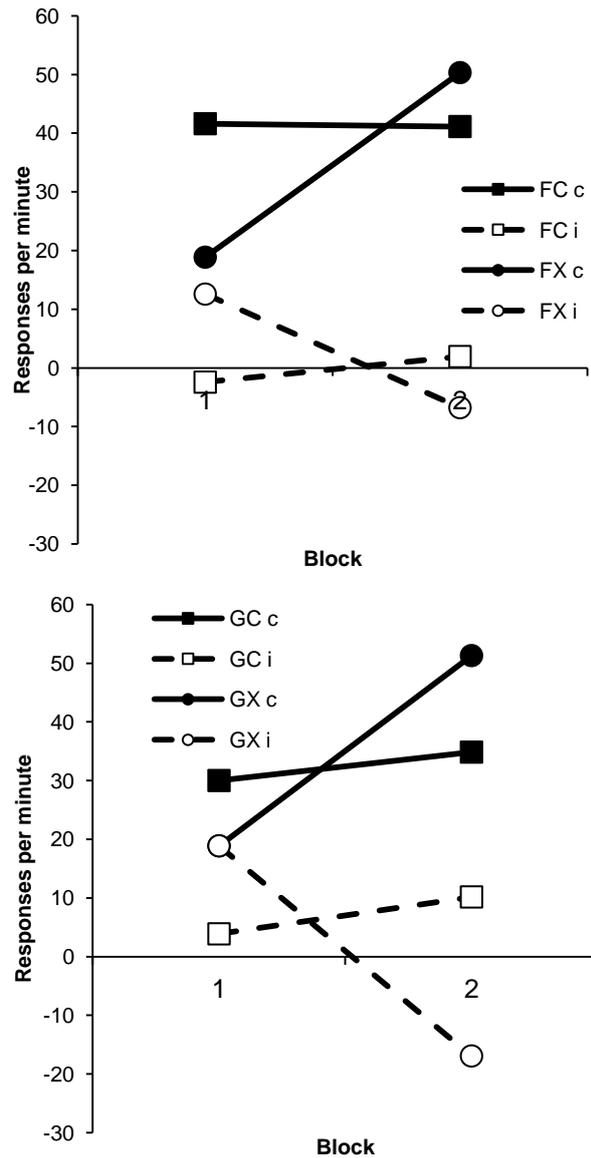


Figure 17. Mean rate of congruent and incongruent responses for each type of CS in the PIT Test of Experiment 5. Top panel: PIT scores for FC and FX. Bottom panel: PIT scores for GC and GX.

The mean responses during the pre-CS period are presented in Table 8. An identical analysis showed no significant effects or interactions, largest  $F(1, 30) = 1.02, p = .32, MSe = 13.31$ .

*Table 8.* Mean preCS response rates in each block of the PIT test of Experiments 5, 6 and 8.

	Congruent				Incongruent			
	FC	FX	GC	GX	FC	FX	GC	GX
<b>Exp 5 Block 1</b>	67	76	75	72	67	68	67	56
<b>Exp 5 Block 2</b>	74	74	81	58	79	67	67	80
<b>Exp 6 Block 1</b>	58	68	24	62	77	59	88	42
<b>Exp 6 Block 2</b>	55	62	74	74	76	62	56	58
<b>Exp 8 Block 1</b>	46	49	44	58	41	64	40	44
<b>Exp 8 Block 2</b>	58	56	64	62	63	56	56	53

### 3.4.3 Discussion

The results of this experiment successfully replicated those of Experiment 4, in which the specific PIT effect was reduced by the CI. Although the effect of X was transient, being present only at the first block of the test, these results support the S-O-R accounts. Importantly, in this experiment X had the same effect on both F and G, even though X signalled the absence of O<sub>1</sub> and G the deliveries of O<sub>2</sub>. These results are consistent with the findings of the summation test, in which the effect of X was not outcome-specific.

It is not clear why the effect of X disappeared on the second block of the test. One possibility is that participants rapidly learned to ignore the less informative CS of the compound (X and C), focusing on the CSs that predicted the outcomes. Another possible explanation is that both S-O-R and S-R mechanisms are competing mechanisms part of the specific PIT effect. The procedure used in these experiments might have favoured the S-O-R mechanisms and that is

why X initially disrupted the specific PIT effect by suppressing the activation of the outcome representation. However, this disruption might have allowed the S-R mechanism to take control of the specific PIT effect, which explains why this effect reappears. This possibility was explored in Experiment 6.

### 3.5 Experiment 6

If both S-O-R and S-R mechanisms work in parallel it is possible that the procedures used in the previous experiments favoured an S-O-R mechanism of PIT over the formation of S-R associations. Applying the S-R perspective to Experiments 4 and 5, in the Pavlovian training an association was formed between F and  $O_1$  (and between G and  $O_2$ ). Then in instrumental training when  $R_1$  was reinforced with  $O_1$ , this outcome activated a representation of F, which should allow the formation of an F- $R_1$  link. This association would allow F to directly elicit  $R_1$  in the PIT test. However, in these experiments both responses were trained in the same session, which resulted in  $R_1$  always being followed by  $O_1$  but preceded by *both*  $O_1$  and  $O_2$ . Thus if a representation of F is activated by  $O_1$  this should produce two types of associations: an  $R_1$ -F and an F- $R_1/R_2$  association. Although only the  $R_1$ -F relationship is always consistent, this arrangement should be detrimental to the ability of F to produce the specific PIT effect. A second issue is that because the instrumental

training phase was conducted after the Pavlovian phase, each time an outcome followed a response, e.g.  $R_1 \rightarrow O_1$ , the outcome activated a representation of the CS, e.g.  $R_1 \rightarrow O_1 \rightarrow F$ , forming a backward association between the CS and the response (R-S association). However, it is arguable that the ability of S to elicit responding at test will be greater if the CS precedes the activation of a response representation during training, rather than vice versa. Thus, the procedure used in the previous experiments was not ideal to produce strong S-R associations. A third problematic aspect of the task is that the summation test was conducted between training of F/G and the instrumental stage. Non-reinforced presentations of these stimuli might have reduced their ability to activate the outcomes during instrumental phase, which would also hinder any S-R contribution to the specific PIT effect.

For these reasons, Experiment 6 aimed to foster the S-R associations and to assess if a CI can reduce the specific PIT effect under such conditions. In this experiment the order of the phases was reversed. The study began with the instrumental phase followed *immediately* by training of  $F \rightarrow O_1$  and  $G \rightarrow O_2$  (Table 9). Although this arrangement does not completely eliminate the contribution of the S-O-R mechanism to the PIT effect, it should facilitate the formation of S-R associations for two reasons. First, this procedure should allow the formation of forward S-R associations instead of a backward R-S links. For instance, in each trial that F is followed by  $O_1$ , the outcome activates a representation of  $R_1$ , forming an  $F-R_1$  association. Second,

training of the rest of the stimuli and the summation test were conducted after instrumental training and F-O<sub>1</sub> and G-O<sub>2</sub> pairings. This should eliminate any interference or extinction effect of the summation test from the formation of the S-R associations. If this arrangement produces stronger S-R associations then the ability of the CSs to elicit responding might be greater. If this is correct, and the S-R mechanism is part of the specific PIT effect, then the effect of the CI on the PIT test should be harder to found.

*Table 9.* Design of Experiment 6.

Instrumental	Pavlovian		Summation	PIT test
	Excitators	Pre-training Inhibition		
R <sub>1</sub> ->O <sub>1</sub>	F->O <sub>1</sub>	A->O <sub>1</sub>	FX	FX
R <sub>2</sub> ->O <sub>2</sub>	G->O <sub>2</sub>	AB->O <sub>1</sub>	FC	FC
		AX <sup>-</sup>	GX	GX
		X <sup>-</sup>	GC	GC
		CH <sup>-</sup>		
		C <sup>-</sup>		

*Note:* A, B, C, F, G, H and X: neutral fractal images; R<sub>1</sub> and R<sub>2</sub>:

keyboard responses; O<sub>1</sub> and O<sub>2</sub>: food and drink images. - denotes no outcome.

### 3.5.1 Method

*Participants.* 32 students from the University of Nottingham participated in this experiments (12 males and 20 females) aged between 18 and 32 years old.

*Procedure.* Everything was the same as in Experiment 5 unless otherwise stated.

*Instrumental training.* In this experiment instrumental training was conducted first.

*Pavlovian phase.* The *excitor training* was conducted immediately after the instrumental phase, and was followed by *pre-training* and then *inhibitory training*.

*Summation test.*

*PIT test.*

### **3.5.2 Results**

*Instrumental training.* All participants completed this phase successfully (Table 6), and no significant differences were found between the type of response performed ( $R_1$ ;  $R_2$ ) nor the type of outcome received ( $O_1$ ;  $O_2$ ) by the participants,  $F_s < 1$ .

*Pavlovian phase.* The mean scores from *excitor training* are plotted on the top panel of Figure 18. The Figure suggests that participants rapidly learned to predict the outcomes for both CSs. An ANOVA with CS (F, G) and block (1-4) as factors showed a significant main effect of block,  $F(3, 90) = 25.15$ ,  $p < 0.001$ ,  $MSe = 0.046$ ,  $\eta_p^2 = .46$ , and nothing else was significant, largest  $F(3, 90) = 1.48$ ,  $p = 0.224$ ,  $MSe = 0.046$ . The analysis of the significant main effects revealed differences only between block 1 and the rest of the blocks ( $p < 0.001$ ). In this experiment participants did not receive previous exposure to  $O_1$ , and

unlike the previous experiments there were no differences between F and G. This confirms the idea that the lower scores to F found in the previous experiments were caused by a pre-exposure effect of O<sub>1</sub>.

The mean scores of *inhibitory training* are presented on the bottom panel of Figure 18. The graph suggests that, as in the previous experiments, participants learned about AX slightly slower than the rest of compounds. An ANOVA with CS (A, AB, AX, CH) and block (1-4) as factors showed a significant main effect of CS,  $F(3, 90) = 7.04$ ,  $p < .001$ ,  $MSe = .039$ ,  $\eta_p^2 = .19$ , a significant main effect of block,  $F(3, 90) = 18.71$ ,  $p < .001$ ,  $MSe = .029$ ,  $\eta_p^2 = .38$ , and a significant interaction,  $F(9, 270) = 5.82$ ,  $p < .001$ ,  $MSe = .031$ ,  $\eta_p^2 = .16$ . The analysis of the interaction revealed an effect of CS in blocks 1, 2 and 3 ( $p < 0.05$ ) but not in block 4,  $F < 1$ . In block 1 AX scores were lower than those of A, AB and CH ( $p < .05$ ), while in block 3 AB scores were lower than those of A and CH ( $p < .05$ ). No other differences were significant, smaller  $p = .18$ .

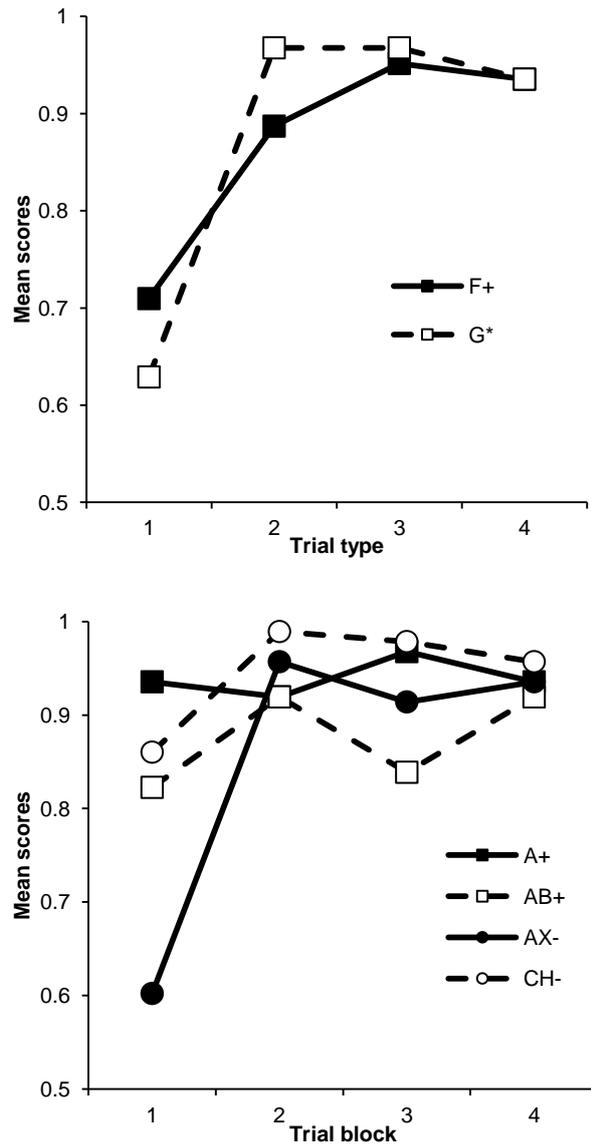
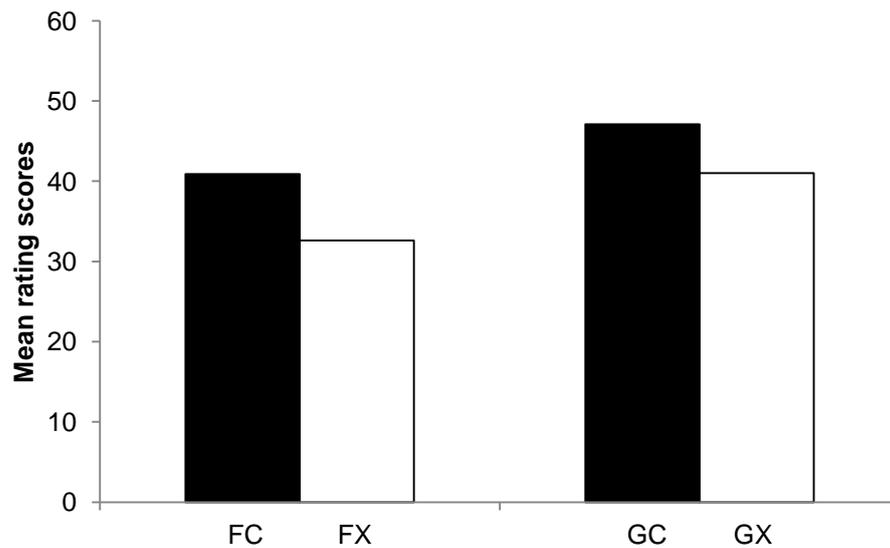


Figure 18. Mean scores grouped by CS and block in the Pavlovian phase of Experiment 6. Top panel: the mean scores to the CSs in the test excitator training. Bottom panel: the mean scores to the CSs in the inhibitory training.

*Summation test.* The mean rating scores of the summation are presented in Figure 19. The graph shows lower ratings for FX and GX than for FC and GC, which was confirmed by the statistical analysis. An ANOVA with CS (F, G) and trial type (C, X) as factors revealed a

significant main effect of trial type,  $F(1, 30) = 4.83$ ,  $p = 0.036$ ,  $MSe = 0.033$ ,  $\eta_p^2 = .14$ . Nothing else was significant, largest  $F(1, 30) = 3.15$ ,  $p = 0.086$ ,  $MSe = 0.053$  for the main effect of CS. These results are consistent with those of the previous experiments, in which X reduced the expectancies of both  $O_1$  and  $O_2$ , even though it was only trained to signal the absence of  $O_1$ .



*Figure 19.* Mean ratings of the likelihood of  $O_1$  occurrence during FC and FX, and  $O_2$  occurrence during GC and GX in the summation test of Experiment 6.

*PIT test.* The PIT scores are presented in Figure 20. The top panel of the Figure suggests that the specific PIT effect was produced by FC, but this effect seems to be reduced in the FX trials. Similarly, the bottom panel suggests that the specific PIT produced by GX was smaller than that produced by GC, although it seems a transient effect. An ANOVA with block (1-3), congruence, CS (F, G) and trial type (C, X) as factors revealed a significant main effect of congruence,  $F(1, 29)$

= 10.27,  $p = .003$ ,  $MSe = 84.18$ ,  $\eta_p^2 = .26$ , which interacted with trial type,  $F(1, 29) = 5.48$ ,  $p = .026$ ,  $MSe = 19.74$ ,  $\eta_p^2 = .16$ . Nothing else was significant, largest  $F(2, 58) = 2.9$ ,  $p = .06$ ,  $MSe = 18.37$  for the Block x Congruence x Trial type interaction. The analysis of the significant interaction revealed an effect of congruence at C,  $F(1, 29) = 13.19$ ,  $p = .001$ , but not at X,  $F(1, 29) = 4.12$ ,  $p = .052$ . This analysis confirms that the specific PIT effect produced by F and G was reduced by X presentations.

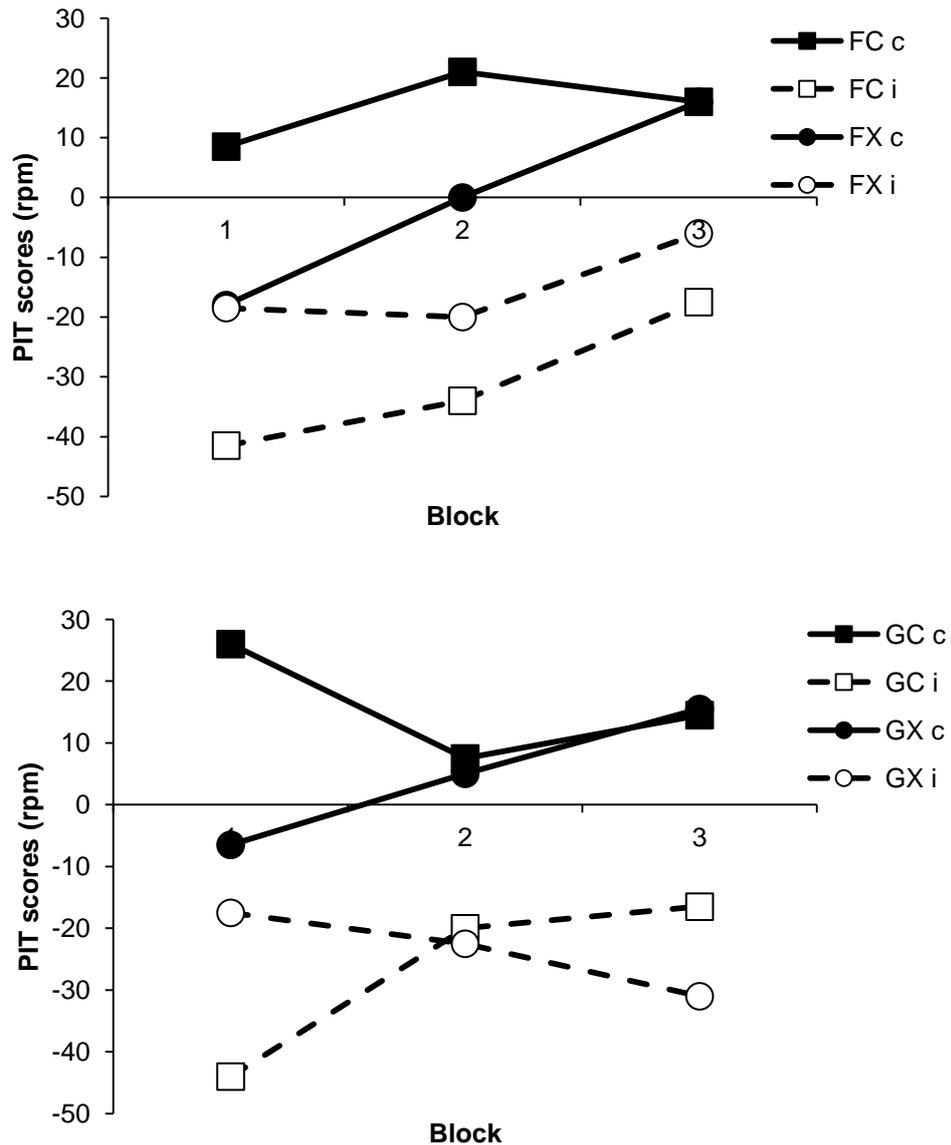


Figure 20. Mean rate of congruent and incongruent responses for each type of CS in the PIT Test of Experiment 6. Top panel: PIT scores for FC and FX. Bottom panel: PIT scores for GC and GX.

The mean rates of responding during the pre-CS period are plotted in Figure 21. Although the Figure does not show a clear pattern of differences, it suggests that in the first block participants performed more incongruent than congruent responses in the presence of GC. The corresponding ANOVA revealed a significant Block x CS

interaction,  $F(2, 58) = 3.29$ ,  $p = .044$ ,  $MSe = 13.42$ ,  $\eta_p^2 = .10$ , and a significant Block x Congruence x Trial type interaction,  $F(2, 58) = 3.52$ ,  $p = .036$ ,  $MSe = 25.8$ ,  $\eta_p^2 = .11$ . Nothing else was significant, largest  $F(1, 29) = 4.11$ ,  $p = .052$ ,  $MSe = 19.16$  for the Congruence x Trial type interaction. Two-way ANOVAs were conducted with congruence and trial type as factors for each of blocks, which showed a showed an interaction in block 1,  $F(1, 29) = 7.97$ ,  $p = .009$ ,  $MSe = 29.62$ ,  $\eta_p^2 = .22$ , but not in block 2 or 3, largest  $F(1, 29) = 1.03$ ,  $p = 0.32$ ,  $MSe = 11.79$ . This analysis confirmed a higher rate of responding during the incongruent trials for FC and GH in the first block. However, this difference was not present in the rest of the blocks, suggesting that this transient effect could have not affected the results of the PIT scores, in which lower responding to X compared to C was present across the three blocks of the test.

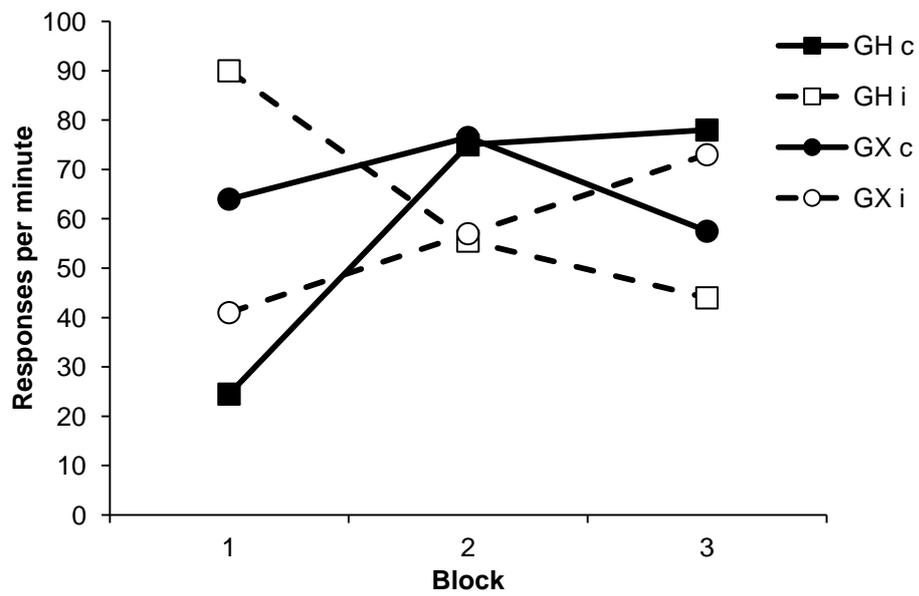
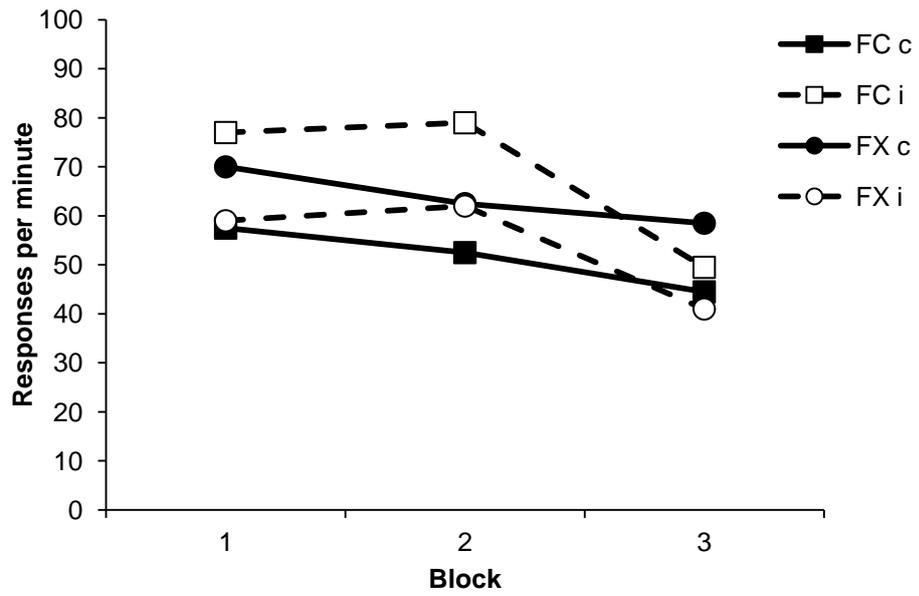


Figure 21. Mean rate of congruent and incongruent responses for each type of CS in the pre-CS period of the PIT Test of Experiment 6. Top panel: responses during FC and FX. Bottom panel: responses for GC and GX.

### **3.5.3 Discussion**

These results replicated those found in the previous experiments. The specific PIT effect was reduced in the FX and GX trials compared to the FC and GC, which is consistent with the results of the summation test. In this test, participants' expectancies of  $O_1$  and also  $O_2$  were reduced by the presence of X, which confirmed X as a CI and also provides more evidence supporting the proposal that in these experiments conditioned inhibition is not outcome-specific.

In this experiment instrumental training was conducted at the beginning, followed immediately by training of F and G. Although this arrangement was thought to facilitate the formation of S-R associations, no evidence was found to support the suggestion that these association were formed nor that they contribute to the specific PIT effect. However, it is still possible that the S-R associations formed during training were not strong enough to counteract the effect of the CI on suppressing the activation of the outcome representation.

### **3.6 Experiment 7**

It is important to note that the results of these experiments have been interpreted as supporting the S-O-R mechanisms by assuming that the CI suppressed the outcome representation. Although there is a strong basis in animal work for this assumption (e.g. Rescorla &

Holland, 1977), it is possible that the effect of the CI on the specific PIT effect is due to different reasons. For instance, an S-R account of inhibition could explain the results presented here by assuming that X became associated with responses opposite to those elicited by F and G. Then at test X elicited these competing responses, interfering with performance of both  $R_1$  and  $R_2$ . This would explain the non-specific effect of conditioned inhibition in these experiments because X would have similar effect when presented with F and G. Although it is not clear why X would elicit more competing responses more than C, which underwent similar training, the goal of this study was to test this possibility.

The design was similar to that of Experiment 5 (Table 10) with two critical differences. First, the instrumental responses were trained without outcome deliveries. Instead, participants had to perform 'z' and 'm' responses until they reached the number of responses required. Then X and C were presented in the PIT test. If X elicits competing responses then it should reduce performance of both responses even if these responses were not reinforced with either of the outcomes. But if the effect of X was due to a suppression of the outcome presentation, it should have no effect on performance in this experiment.

Table 10. Design of Experiment 7.

Pavlovian phase			Summation	Instrumental	PIT test
Pre-training	Inhibition	Exciters			
A->O <sub>1</sub>	A->O <sub>1</sub>	F->O <sub>1</sub>	FX	Left	X
AB->O <sub>1</sub>	AB->O <sub>1</sub>	G->O <sub>2</sub>	FC	Right	C
	AX <sup>-</sup>		GX		
	X <sup>-</sup>		GC		
	CH <sup>-</sup>				
	C <sup>-</sup>				

Note: A, B, C, F, G, H and X: neutral fractal images; Left and Right:

keyboard responses; O<sub>1</sub> and O<sub>2</sub>: food and drink images. - denotes no outcome.

### 3.6.2 Method

*Participants.* 32 students from the University of Nottingham participated in this experiments (8 males and 24 females) aged between 18 and 25 years old.

*Procedure.* Everything was the same as in Experiment 5 unless otherwise stated.

*Pavlovian phase.*

*Summation test.*

*Instrumental training.* The only difference in this phase was that the outcomes were not presented. Instead, participants received the instructions: "Now in this part you have to press the keys 'Z' and 'M'. You will see a "+" on the screen and you will have a counter for each key, so your goal is to press the buttons until you obtain 50 of each.

*You will not increase the counters all the times, so you have to keep trying- you can press as many times or as quickly as you see fit!"*

Although no outcome was delivered, each counter increased its value by 1 according to the same VR5 schedule as before.

*PIT test.* The only cues that were presented were X and C.

## Results

*Pavlovian phase.* The mean scores of *inhibitory training* are presented in the top panel of Figure 22. As in the previous experiments, scores to AX seems to be initially lower than to the rest of the cues, but this effect had dissipated by the end of the training. An ANOVA with CS (A, AB, AX, CH) and block (1-4) as factors showed a significant main effect of CS,  $F(3, 93) = 4.13, p = .008, MSe = .041, \eta_p^2 = .12$ , and block,  $F(3, 93) = 3.74, p = .014, MSe = .027, \eta_p^2 = .11$ . Nothing else was significant,  $F(9, 279) = 1.25, p = .263, MSe = .037$ . The analysis of the main effect of CS revealed significant differences only between AX and CH ( $p = .002$ ) but no significant differences were found between the blocks.

The mean scores of *excitor training* are plotted on the bottom panel of Figure 22, and as in experiments 4 and 5 F scores were lower than G scores, difference that persisted across training. The results of the ANOVA showed a significant main effect of CS,  $F(1, 31) = 15.27, p < .001, MSe = .061, \eta_p^2 = .33$ , and nothing else was significant, largest  $F(3, 93) = 1.09, p = .36, MSe = .051$ .

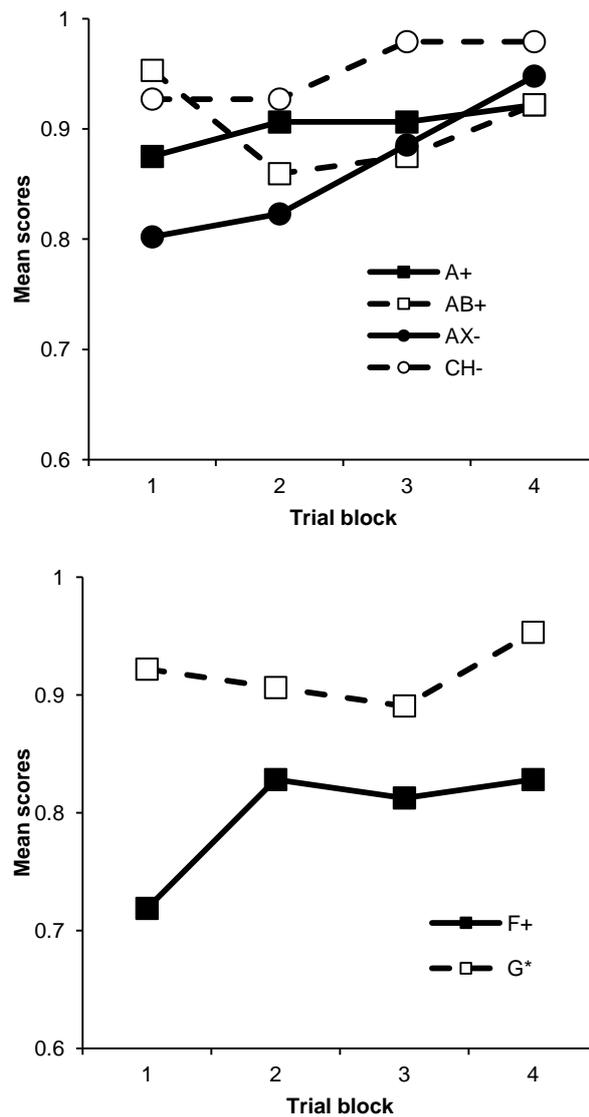
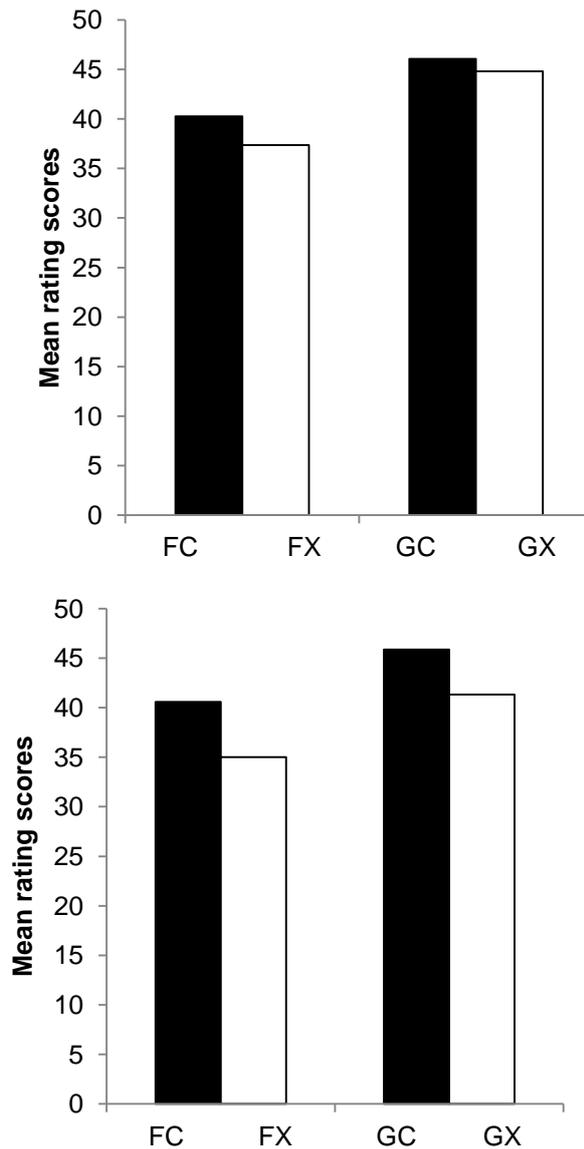


Figure 22. Mean scores grouped by CS and block in the Pavlovian phase of Experiment 7. Top panel: the mean scores to the CSs in the inhibitory training. Bottom panel: the mean scores to the CSs in the test excitator training.

*Summation test.* Unlike in the previous experiments, the results of the summation test were not significant. The mean scores are presented on the top panel of Figure 23, and an ANOVA with CS (F, G) and trial type (C, X) as factors showed a significant main effect of CS,  $F(1, 31)$

= 7.3,  $p = .011$ ,  $MSe = .019$ ,  $\eta_p^2 = .19$ , and nothing else was significant,  $F_s < 1$ . The reasons for this are unclear, but because the purpose of this experiment was to assess the effect of a CI on instrumental performance it was critical to obtain evidence of inhibition. For this reason 6 participants that rated the outcome as more likely to occur in the FX and GX trials than in the FC and GC trials were excluded from this analysis and from the PIT test, leaving 26 subjects. The new scores of the summation test are presented on the right panel of Figure 23 and it can be observed that there were higher rating scores to the control than to the inhibitory compounds. The ANOVA showed a significant main effect of CS,  $F(1, 25) = 6.9$ ,  $p = .014$ ,  $MSe = .013$ ,  $\eta_p^2 = .22$ , and a significant main effect of trial type,  $F(1, 25) = 5.22$ ,  $p = .031$ ,  $MSe = .013$ ,  $\eta_p^2 = .17$ . The interaction was not significant,  $F < 1$ . These results confirm X as a CI, and makes possible a comparison between X and C during the PIT test.



*Figure 23.* Mean ratings of the likelihood of O1 occurrence during FC and FX, and O2 occurrence during GC and GX in the summation test of Experiment 7. Top panel: ratings before participants' exclusion. Bottom panel: ratings after participants' exclusion.

*Instrumental training.* All participants performed both responses. A mean of 256.9 'z' responses and 255.3 'm' responses was performed. The mean number of times the left counter was increased was 51.97,

and the right counter was 51.22. Neither of these differences was significant,  $F_s < 1$ .

*PIT test.* In this experiment there were no congruent or incongruent responses because neither of the outcomes was presented at instrumental training. Thus, the number of z and m responses was averaged and presented in Figure 24. The graph suggests that although X and C produced a slight elevation of performance at the beginning of the test, there seems to be no difference between these two cues. An ANOVA with block (1, 2) and trial type (C, X) as factors showed a significant main effect of block,  $F(1, 25) = 5.35$ ,  $p = .029$ ,  $MSe = 11.36$ ,  $\eta_p^2 = .18$ . Nothing else was significant,  $F_s < 1$ .

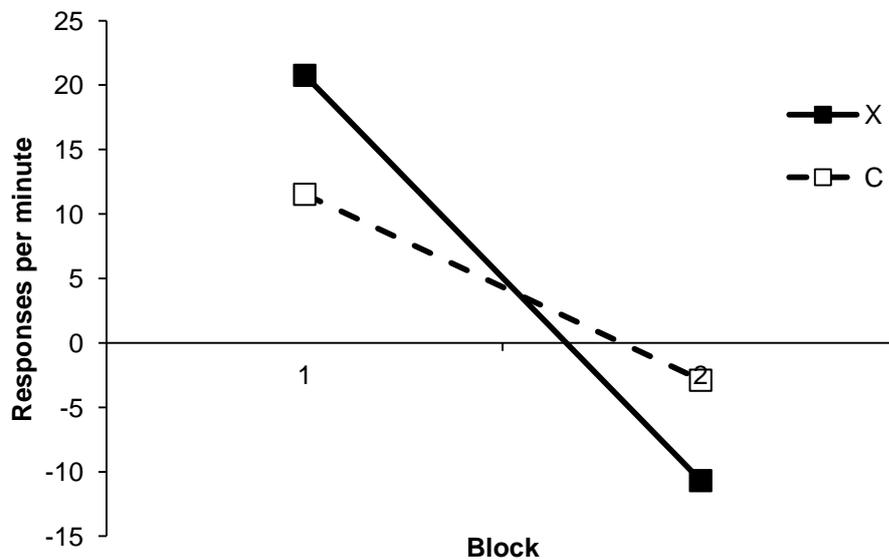


Figure 24. Mean response rate for X and C in the PIT test of Experiment 7.

The mean pre-CS responses per minute for X and C respectively were 134 and 141 in block 1, and 155 and 159 in block 2.

The statistical analysis showed no significant main effect or interaction, largest  $F(1, 25) = 2.7, p = .11, MSe = 16.7$ .

### **3.6.3 Discussion**

These results provide no evidence that X elicited a competing response that might interfere with instrumental performance. Both X and C had the same effect during the PIT test, even though the results of the summation test confirmed X as a CI. Thus, it may be concluded that this mechanism was not responsible for the results of the previous experiments.

### **3.7 Experiment 8**

Another alternative explanation for the effect of the CI in the PIT test is that X did not suppress the outcome representation but instead activated the representation of something else, which interfered with instrumental performance. For instance, during training X was followed by nothing, which was represented by a white square; thus if an association was formed between X and this outcome then X could activate such a representation later at test. Because instrumental performance depends on the R-O associations formed during training, this performance must be determined, at least partially, by a representation of the reinforcer. Thus, X activating an alternative

representation might have interfered with the retrieval of the outcome representation encoded in the R-O associations. If this is correct it is not clear why X had a differential effect compared to C, considering that both cues had the same contingency with nothing/square. Nevertheless this experiment was conducted to evaluate this possibility.

This experiment was the same as Experiment 5 except that in the Pavlovian phase X was presented with the novel cue D, replacing AX (Table 11). DX and X were both followed by a white square with the word '*Nothing*' on it, while CH and C were paired only with the white square. This arrangement should facilitate the formation of an association between X and the white square with the word nothing on it. Then if the CI reduced the PIT effect due to the activation of this different outcome, a larger reduction of the PIT effect in the presence of X than C is expected. Furthermore, this reduction should be larger than in the previous experiments because X was explicitly trained with an alternative outcome. In this study X was not trained as a conditioned inhibitor, and for this reason the summation test was not included in this experiment.

Table 11. Design of Experiment 8.

Pavlovian phase			Instrumental	PIT test
Pre-training	'Inhibition'	Exciters		
A->O <sub>1</sub>	A->O <sub>1</sub>	F->O <sub>1</sub>	R <sub>1</sub> ->O <sub>1</sub>	FX <i>Inhibitory</i>
AB->O <sub>1</sub>	AB->O <sub>1</sub>	G->O <sub>2</sub>	R <sub>2</sub> ->O <sub>2</sub>	FC <i>Pre-exp</i>
	DX->'Nothing'			GX <i>Unrelated</i>
	X->'Nothing'			GC <i>Pre-exp</i>
	CH			
	C			

Note: A, B, C, D, F, G, H and X: neutral fractal images; R<sub>1</sub> and R<sub>2</sub>:

keyboard responses; O<sub>1</sub> and O<sub>2</sub>: food and drink images. - denotes no outcome.

### 3.7.1 Method

*Participants.* 32 students from the University of Nottingham participated in this experiment (8 males and 24 females) aged between 18 and 29 years old.

*Procedure.* Everything was the same as in Experiment 5 unless otherwise stated.

*Pavlovian phase.* AX trials were replaced by DX. The X and DX presentations were followed by a white square with the word 'Nothing' in the middle, occupying approximately 20% of the square.

*Instrumental training.*

*PIT test.*

### 3.7.2 Results

*Pavlovian phase.* The mean scores for A, AB, DX and CH are presented on the top panel of Figure 25. The graph suggests that participants learned to predict the outcome for each of the compounds, although they were less accurate in the case of AB. An ANOVA with CS and block (1-4) as factors showed a significant main effect of CS,  $F(3, 93) = 3.34, p = .023, MSe = .025, \eta_p^2 = .10$ , and nothing else was significant,  $F_s < 1$ . The analysis of the significant main effect showed differences between AB and DX ( $p = .038$ ) and between AB and CH ( $p = .032$ ). In the previous experiments the scores to AX were lower than the rest, but in this study participants were accurate in predicting the outcome of DX. This confirms the previous suggestion that participants were less accurate in the AX trials due to the fact that A was also reinforced in the absence of X.

The mean scores of F and G are plotted on the bottom panel of Figure 25. As in the previous experiment, the scores to G were higher than to F. Although it seems that both scores were lower at the end of the training, this was not confirmed by the statistical analysis. ANOVA with CS and block (1-4) as factors showed only a significant main effect of CS,  $F(1, 31) = 11.12, p = 0.002, MSe = 0.059, \eta_p^2 = .26$ . Nothing else was significant, largest  $F(3, 93) = 1.65, p = 0.184, MSe = 0.083$ .

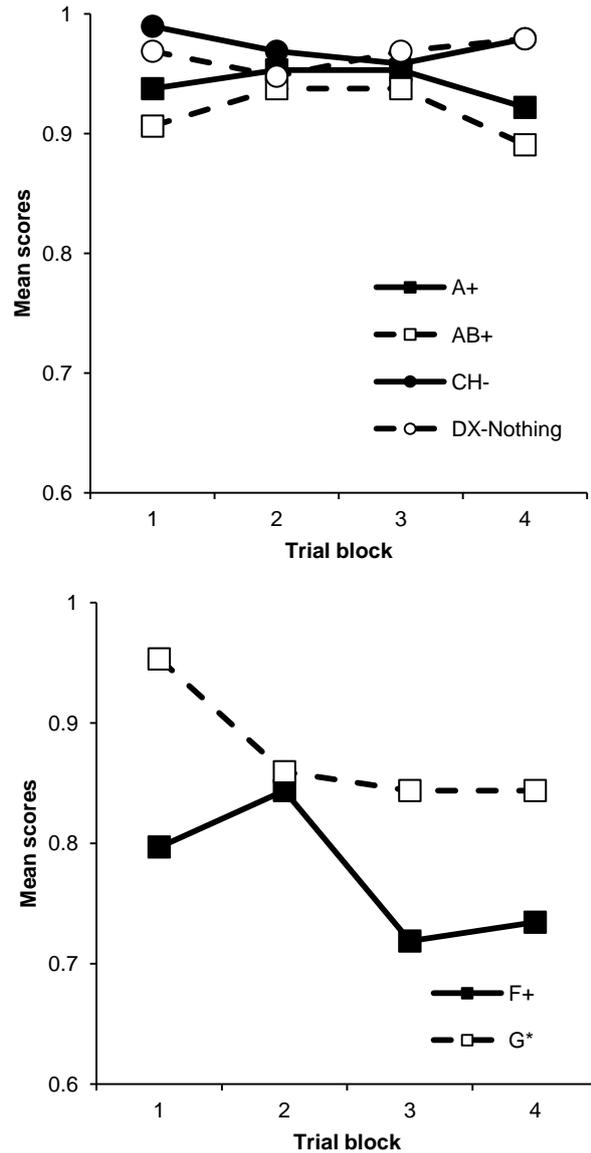


Figure 25. Mean scores grouped by CS and block in the Pavlovian phase of Experiment 8. Top panel: the mean scores to A, AB, CH and DX. Bottom panel: the mean scores to F and G.

*Instrumental phase.* All participants completed the training successfully (see Table 6). No significant differences were found between the type of response performed nor between the types of outcome delivered,  $F_s < 1$ .

*PIT test.* The PIT scores for FC and FX are presented on the top panel Figure 26, and the PIT scores for GC and GX on the bottom. The graphs suggest that a specific PIT effect of similar size was found in the presence of all the compounds. An ANOVA with block (1, 2), congruence, CS (F, G) and trial type (C, X) as factors revealed a significant main effect of congruence,  $F(1, 31) = 14.66$ ,  $p = .001$ ,  $MSe = 72.74$ ,  $\eta_p^2 = .32$ . Nothing else was significant, largest  $F(1, 31) = 3.15$ ,  $p = .086$ ,  $MSe = 13.38$ .

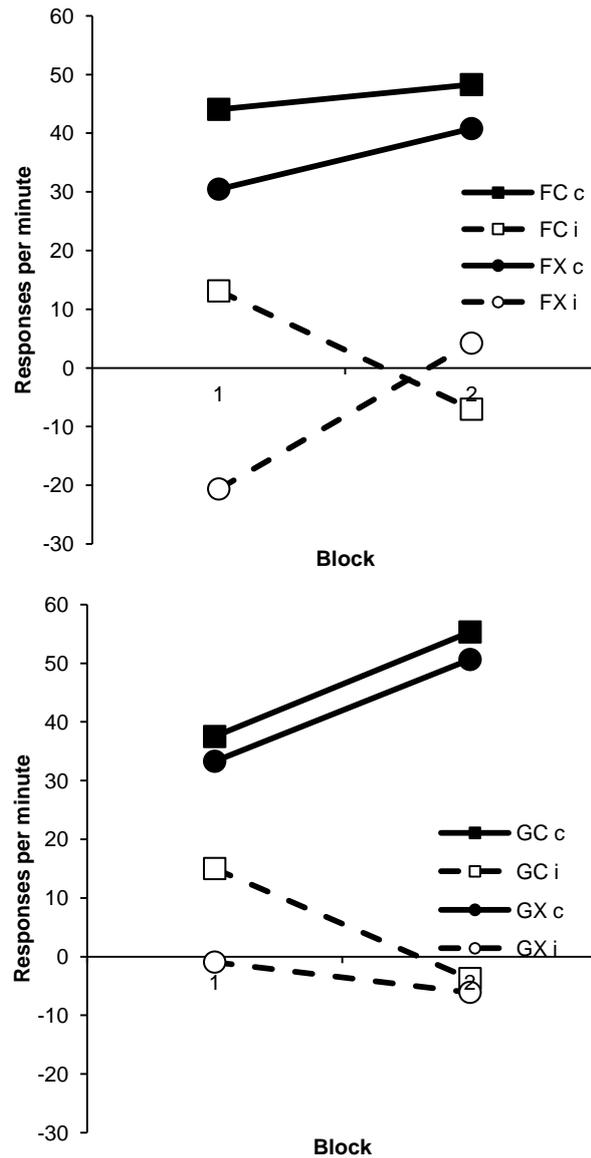


Figure 26. Mean rate of congruent and incongruent responses for each type of CS in the PIT Test of Experiment 8. Top panel: PIT scores for FC and FX. Bottom panel: PIT scores for GC and GX.

The mean rates of responding during the pre-CS period are presented in Table 8. The ANOVA showed no significant main effect or interaction, largest  $F(1, 31) = 3.29, p = .08, MSe = 6.93$ .

### 3.7.3 Discussion

The results of this experiment do not support the idea that X reduced the specific PIT effect due to the activation of a different outcome. In this experiment X was explicitly paired with the outcome 'Nothing' but it had no effect in the PIT test. These results strongly suggest that this was not the mechanism behind the results of the previous experiments.

### 3.8 General Discussion

This series of experiments aimed to provide evidence on the outcome-specificity of conditioned inhibition, and to further assess the effect of conditioned inhibition in the specific PIT effect. In Experiments 4, 5 and 6, two CS+s were paired with one of two outcomes ( $F \rightarrow O_1$ ;  $G \rightarrow O_2$ ) and a CI was trained to signal the absence of one of these outcomes ( $X \rightarrow \text{no } O_1$ ). Additionally, two responses were trained, each of them with one of the outcomes ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). In the PIT test, F and G were presented either with X or with a pre-exposed stimulus. In all these experiments the specific PIT effect was found when F and G were presented with the control stimulus, but this effect was greatly reduced by the presence of X. Moreover, X reduced the specific PIT effect even when it was presented with G, which was a predictor of a different outcome ( $O_2$ ). These findings are consistent with those of the

summation tests. In these tests participants rated the likelihood of the outcomes in the presence of F and G, either in compound with X or with the pre-exposed stimulus. Participants' expectations of  $O_1$  and  $O_2$  were reduced in the presence of X compared to the control stimulus, confirming the inhibitory properties of X. The fact that X reduced the expectations of  $O_2$  even when it was trained to signal the absence of  $O_1$ , strongly suggest that conditioned inhibition was not outcome-specific in these experiments. Although evidence supporting this non-specific effect of conditioned inhibition has been reported before (LoLordo and Fairless, 1985; Nieto, 1984; Pearce, Montgomery and Dickinson, 1981) to my knowledge there is no evidence on inhibitors producing this *general* effect in the specific PIT test.

The fact that a CI reduced the specific PIT effect regardless of the outcome of training is not easy to explain. According to the S-O-R accounts, the specific PIT effect must be determined by the CS+ retrieving specific information about the outcome of training, which in turn produces the response previously reinforced with the same outcome. Thus if a CI does not suppress a specific representation of this outcome, then it is not clear how it could affect the specific PIT effect. This raised the possibility that the results of Experiments 4, 5 and 6 were not caused by the inhibitory properties of X but to something else. One alternative was that X did not suppress the outcome representation but instead elicited competing responses at test, reducing the specific PIT effect for both F and G. This idea was tested in Experiment 7, in which the critical aspect was that the

instrumental responses were trained in the absence of  $O_1$  and  $O_2$ . Thus, if the effect of X, which was trained to signal the absence of  $O_1$ , was caused by eliciting competing responses, then X should reduce performance even if the responses were reinforced with valueless outcomes. Contrary to this idea, the effect of X in the PIT test was not different to that of a pre-exposed stimulus. These results rule out the alternative that the reduction in the specific PIT effect produced by X in the previous experiments must be caused by X somehow interfering with the outcome representation encoded in the R-O associations. However, this does not necessarily mean that X suppressed a representation of the outcome. If during training X became associated not with the absence of  $O_1$  or  $O_2$ , but with a different outcome, then at test X might have activated a representation of this outcome. This could have interfered with the retrieval of the outcome representation by the instrumental responses, reducing the specific effect for both F and G. This idea was tested in Experiment 8, in which X was not trained as a CI but instead was explicitly paired with a different outcome and then presented with F and G in the PIT test. Again, X had no effect on the specific PIT effect produced by these CSs+ compared to a pre-exposed control stimulus. These results confirm that X must be trained in a conditioned inhibition procedure to reduce the specific PIT effect.

Overall these results are consistent with the idea that the role of X in the reduction of the specific PIT effect was caused by its inhibitory properties. But the question of an inhibitor can reduce the specific PIT

effect without suppressing a specific sensory representation of the outcome still remains. Applying Konorski's conceptualization (1967) of conditioned inhibition to these experiments, F activated an internal representation of  $O_1$  at test that encoded sensory information about  $O_1$ , i.e. the  $O_1^s$  centre, and another of the motivational state elicited by this outcome, i.e. the  $O_1^m$  centre. Similarly, G activated  $O_2^s$  and  $O_2^m$  centres. As a result of its inhibitory training, X activated a no- $O_1^s$  and a no- $O_1^m$  that antagonised the centres activated by F, but left the  $O_2^s$  and  $O_2^m$  centres unchanged. In this sense, X should have reduced the specific PIT effect produced by F but not by G. However, in these experiments the outcomes were pictures of food and drinks, so presumably they had lower motivational value than real outcomes. Moreover, it is not unlikely that both food and drink pictures elicited a similar motivational state in the participants. Assuming that this is correct, then F and G would have activated different sensory representations of the outcomes, but the same motivational centre,  $O_{12}^m$ . Thus, X might have suppressed this common centre, either directly (Konorski, 1948) or by activating a no- $O_{12}^m$  (Konorski, 1967), reducing the specific PIT effect produced by F and G. This may be a valid interpretation of the results presented here only given the assumptions outlined above - that (i) the specific PIT effect is mainly determined by the sensory aspects of the outcome, but that (ii) the motivational state elicited by these outcomes is also necessary to observe the effect.

This interpretation seems to be contrary to the evidence of those experiments using outcome devaluation procedures. As was mentioned before, it has been consistently found that reducing the value of the outcome before the test, either by pairing it with an aversive consequence (e.g. Holland, 2004) or by shifting the motivational state of the subjects (e.g. Corbit, Janak & Balleine, 2007), has no effect on the specific PIT effect. However, it is possible that a CI exerts a more pronounced effect on behaviour than simply devaluing the outcome. It is noteworthy that the effect of the CI in these experiments seems to be more marked in the first block of the PIT tests, disappearing on the second block (e.g. Experiment 5). This is consistent with the idea that the CS+s retrieve both aspects of the outcome, sensory and motivational, but the sensory aspects are dominant in the specific PIT effect. For instance, if F activated an  $O_1^{sm}$  representation but X suppressed the motivational component of it, this might affect the specific PIT initially, but the sensory representation of the outcome could still elicit responding across the test.

But there is an alternative interpretation of these results. Similarly to Konorski (1967), Dearing and Dickinson (1979) proposed that a CI forms an *excitatory* association with a certain motivational state during training. But while Konorski (1967) assumed that this motivational state was specific to the outcome of training, i.e. no-US centre, Dickinson and Dearing (1979) suggested that a CI becomes associated with a motivational state opposite to that elicited by the outcomes (appetitive vs. aversive). One of the implications of this idea

is that the effect of a CI that predicts the absence of an appetitive outcome will be similar to that produced by a CS+ predictor of an aversive US. Thus, in the experiments presented here, X might have elicited a state that was incompatible with that elicited by the pictures of food and drinks. But because the outcomes used in the experiments reported here had a low motivational value, it is possible that the effect produced by X on behaviour was also weak, which would explain the transient effect of X in these experiments. Importantly, it is possible that reducing the motivational value an outcome has less impact on behaviour than a CI eliciting an antagonist motivational state. This is because even if an outcome is no longer desirable, this should not affect subject's ability to perform instrumental responses. However, if a CI elicits an opposite motivational state in the subjects, e.g. an aversive state, this should hinder the chances of responding being observed (when this responses have been reinforced with appetitive outcomes). Furthermore, even if the motivational state of the subjects during training is modified before the test, e.g. from hungry to satiated (e.g. Corbit, Janak & Balleine, 2007), this is not the same as a CI eliciting the *opposite* motivational state. For instance, Overmier, Bull and Pack (1971) trained dogs with presentations of a CS+ followed either by food, shock or nothing, and then assessed the effect of these CSs+ on an instrumental response previously reinforced with shock. The authors found that the CS+ paired with food reduced avoidance responding, while the CS+ paired with shock facilitated performance. Thus, if a CI produces an effect similar to a CS+ paired with an

outcome of the opposite motivational value, then this conceptualization of conditioned inhibition can explain the reduction of the specific PIT effect produced by the CI in the present experiments.

It is important to highlight the results of Experiment 8, in which the CI had no effect on the performance of instrumental responses reinforced with valueless outcomes. Although it may be argued that a CI that exerts a *general* effect on subjects' behaviour should have reduced performance, the critical point of this interpretation is that the CI might affect the specific PIT effect either by suppressing the common elements of  $O_1$  and  $O_2$  or by eliciting a motivational state opposite to that produced by the outcomes of training. Because in that experiment the outcomes had no motivational value, the effect produced by the CI did not antagonise performance. Nevertheless, one of the problems with this interpretation are the results of Experiment 1 reported in the Chapter 1. In that experiment the CI presented by itself at test did not reduce performance of instrumental responding. If the CI elicits an antagonist motivational state to that produced by the outcomes, then it should have interfere with performance.

In summary, the results presented in this chapter are consistent with the idea that conditioned inhibition, at least in this task, is not outcome-specific. They also confirm that a CI is capable of reducing the specific PIT effect produced by CS+s, even when these CS+ were trained with a different outcome than the CI. Different interpretations of these results were provided but it is not clear which one, if any, of

these interpretations is correct. If the CI reduced the specific PIT effect by suppressing the activation of the common elements of both outcomes, e.g. motivational representation, then the evidence presented here supports the S-O-R accounts, according to which the activation of an outcome representation mediates PIT. However, if the CI reduced the specific PIT effect by eliciting a motivational state antagonist to the performance of the instrumental responses, then these results cannot only be explained by the S-O-R accounts but also by an S-R account. This is because performance should be reduced if subjects are tested under a motivational state opposite to that of training (elicited by the CI) even if responding is elicited by the activation of an outcome representation (S-O-R account) or by the CSs directly (S-R account). For this reason further research was conducted and reported in the next chapter, in which a different approach was taken. The Experiments reported in Chapter 4 aimed to assess if a backward conditioning procedure results in a CS capable of producing the specific PIT effect. As it was described in Chapter 1, contradictory evidence in animal research has been found. Delamater et al. (2003) and Laurent et al. (2015) found that presenting the outcome before the CS results in a CS that produce the reverse specific PIT effect, i.e. elevation of responses that were trained with a different outcome than the CS, which supports the S-O-R accounts of the specific PIT effect. However, Cohen-Hatton et al. (2013) compared the effect of a CS trained in a backward relation with an outcome (e.g. O1->CS1) to that produced by a CS trained in a forward relation (e.g.

CS2->O2). Cohen-Hatton and colleagues (2013) found that the CS preceded by the outcome in training produced a larger specific PIT effect than the CS followed by the outcome, which is inconsistent with the S-O-R mechanism but supports the S-R account. The next experiments aimed to resolve this conflicting evidence.

## Chapter IV

### Backward conditioning in the specific PIT effect

#### 4.1 Overview

The results found in the previous chapter indicate that a CI trained in a conditioned inhibition procedure exerts a general effect on behaviour, reducing the specific PIT effect produced by CS+s even if these CSs were trained with different outcomes. But these results can be explained by both S-O-R and S-R accounts; thus a different strategy was adopted in this chapter. The aim of this chapter was to explore the possibility of generating CIs by training a CS in a backward relation with the outcome (O->CS), and to assess if this CI can reduce the specific PIT effect. In Experiments 9, 10 and 11, different CSs were paired with an outcome in either a backward or a forward relation and then presented in a PIT test. A similar procedure was used in Experiments 12 and 13, except that in the Pavlovian phase participants' reaction time to the outcome presentations was measured during Pavlovian conditioning in order to increase participants' attention to the direction of the associations (forward and backward). In Experiment 14 the CSs were preceded and followed by either the same or different outcomes during Pavlovian training. Thus, the effect of forward and backward associations on the specific PIT effect was compared directly.

## **4.2 Introduction**

### **4.3 Backward conditioning**

In the *backward conditioning* procedure the US is presented before rather than after the CS. Although some authors have described this procedure as producing excitatory conditioning (e.g. Spetch, Wilkie & Pinel, 1981), the evidence is not entirely consistent. For instance Pavlov (1927) initially suggested that presenting the CS after the US does not result in any form of association, an idea that was shared by other authors (Mackintosh, 1974; Terrace, 1973). However, Pavlov (1928) later proposed that backward procedures might result in inhibition (Pavlov, 1928). This idea that a CS becomes a CI makes sense considering the negative relationship between the CS and US during backward conditioning (Rescorla, 1969b). As in conditioned inhibition, the US is presented during training but never after the CS; thus the CS becomes a predictor of the absence of the US just like a CI. Other authors have proposed that at the beginning of training the CS acquires excitatory properties due to its temporal relation with the outcome but during the training subjects learn that the CS signals the absence of the outcome, so it becomes inhibitory (Cole & Miller, 1999; Heth, 1976).

### **4.4 Inhibitory associations in backward conditioning**

There is evidence supporting the idea that a CS trained in a backward relation with an outcome results in CS with inhibitory

properties (Hall, 1984; Siegel & Domjan, 1971, 1974; Maier, Rapaport & Wheatley 1976; Moscovitch & LoLordo, 1968). For instance, Siegel and Domjan (1971) reported two experiments in which they directly compared backward and forward procedures using a retardation test. In the first experiment they trained rats to press a bar for food and then the subjects were divided into five groups. In the backward group, animals received the US (shock) followed by the CS (tone/light), while the rest of groups either received a) random presentations of the CS and the US; b) only presentations of the CS; c) only presentations of the US; d) neither presentations of the CS nor US. In the retardation test, rats received presentations of the CS followed by the US while they performed the instrumental response, and the acquisition of suppression to the CS was measured. The group that received backward pairings showed the slowest rate of learning, while the fastest group was that without experience with the CS or US. Between both groups were those that were pre-exposed to the CS, US or both. These results were replicated in a second experiment using rabbits as subjects and eyelid conditioning. This was taken as evidence of the CS acquiring inhibitory properties due to backward training. A few years later, Siegel and Domjan (1974) provided additional evidence on this issue by assessing the effect of different numbers of US-CS trials in both rats and rabbits. They found again that the CS retarded the acquisition of a new CS-US association, and that the more US-CS pairings, the slower the acquisition in the retardation test.

#### **4.5 Excitatory associations in backward conditioning**

However, the evidence supporting the idea that backward conditioning results in excitatory learning is also compelling (Ayres, Haddad & Albert, 1987; Burkhardt, 1980; Mahoney & Ayres, 1976; Shurtleff & Ayres, 1981). For example Burkhardt (1980) trained thirsty rats to lick water from a tube and then the subjects were divided into 5 groups. Four experimental groups received a shock delivery (US) immediately followed by the presentation of a white noise (CS), while a fifth group only received an US delivery. Each of the four experimental groups received a shock of a different intensity. Then all the subjects had access to the drinking tube and the latency of the licking response was measured during one CS presentation. The results showed that the CS increased the latency of the response in the experimental groups significantly more than in the control group, and that this increment was directly proportional to the intensity of the shock. Similarly, Ayres and colleagues (1987) used the same task except that in the Pavlovian phase, rats received one trial of a CS (A) followed by the US (shock) and a different CS (B) explicitly unpaired with the US. Then at test rats had the chance to perform the response initially trained while A and B were presented. The results showed that A not only suppressed licking more than B, but also elicited more fear responses (freezing). This sort of evidence suggest that backward conditioning can produce a CS with excitatory properties.

#### **4.6 Excitatory and inhibitory associations in backward conditioning**

The fact that these experiments have found opposite results may be explained by procedural differences. As Tait and Saladin (1986) pointed out, the experiments that found evidence of inhibition usually used a retardation test to measure the effect of the CSs (e.g. Siegel & Domjan, 1971, 1974), while those that found excitatory conditioning tested the effect of the CSs directly on performance (e.g. Ayres, Haddad & Albert, 1987). Tait and Saladin (1986) conducted an experiment to test this idea by using rabbits and eyelid conditioning. They initially trained five groups of thirsty rabbits to lick a tube for water, followed by Pavlovian conditioning using a tone (CS) and a shock (US). Each of the groups received similar training to those reported by Siegel and Domjan (1971) described above - that is a) US-CS; b) CS-US; c) CS alone, d) US alone; and e) no stimuli. In the first test all subjects received CS presentations while performing the licking response and the latency of the response was measured. Subsequently, a retardation test was conducted, in which the CS was paired with a shock and the acquisition of CR was measured (eyeblink). In the first test both CS-US and US-CS groups showed greater latency than the control groups, which suggests that the CS acquired equivalent excitatory strength regardless if it was trained in a forward or backward relation (as in Ayres, Haddad & Albert, 1987). However, in the second test the CS-US group showed more CRs than

the rest, while the US-CS showed the lowest, suggesting that the CS had acquired inhibitory properties (as in Siegel & Domjan, 1971).

It has been noted that in this experiment the order of the tests was not counterbalanced (Williams, Dyck & Tait, 1986). In the first test subjects received non-reinforced presentations of the CS, which might have resulted in extinction of the excitatory properties of this stimulus, thus unveiling any inhibitory strength of the CS at the second test. McNish and colleagues (1997) aimed to replicate the results reported by Tait and Saladin (1986) to eliminate any confounding variables. Using rabbits as subjects and an eyelid conditioning procedure, subjects received trials of a light that ended with a brief shock delivery (US). For the experimental group, a noise (CS) was presented immediately after the US, while for the control group the CS was delivered 6 min after the US. After this training, all the animals received two tests. In one of them, subjects received presentations of the light alone and also with the CS, while eyeblink responses were measured. In the second test, the light was replaced by an air puff and startle responses were measured. The order of the tests was counterbalanced and the results showed that in the experimental group presenting the CS with the light reduced the CRs elicited by the light (eyeblink) but the CS increased the responses elicited by the air puff (startle). These results are consistent with those reported by Tait and Saladin (1986), in which a CS trained in a backward conditioning procedure showed both excitatory and inhibitory properties depending on the type of measurement.

McNish and colleagues (1997) suggested that the CS acquired an excitatory and inhibitory association, but each of them with different aspects of the US. In the experiment described above, the CS increased fear responses (startle) when presented with the air puff, which might reflect an excitatory association between the CS and the motivational properties of the US (shock). But the CS reduced CRs elicited by a light previously paired with the shock, which according to the author might reflect an inhibitory association between the CS and the sensory aspects of the US. To explain this interpretation, McNish and colleagues (1997) appealed to the AESOP model (Wagner & Brandon, 1989). According to this model, the processing of the US representation has a primary state of activity, A1, and then it decays to a secondary state, A2, before becoming inactive. If the CS is processed while the US is in the A1 state an excitatory association is formed, but if it occurs during the A2 state the association is inhibitory. A further assumption that the model makes is that the motivational aspects of the US require more time to be processed; thus these remain in the A1 state longer than the sensory features of the US. Therefore when the CS is presented after the US an excitatory association is formed between the CS and the motivational properties of the US that are responsible for fear responses, such as startle. But the sensory information of the US might be already in the A2 state, so the CS forms an inhibitory association with this US representation, and this specific representation of the US might be the responsible for the CRs produced by the CS+s used in these experiments.

#### 4.7 The specific PIT effect and backward conditioning

If backward pairings result in an inhibitory association between the CS and the sensory aspects of the US, then according to the S-O-R accounts a CS trained in this procedure should not produce the specific PIT effect. Instead it should selectively reduce the performance of the responses trained with the same outcome as the CS, by suppressing the sensory US representation that mediates the PIT effect. Delamater, LoLordo and Sosa (2003) tested this by training CSs and outcomes in a backward relation and then presenting them in a PIT test. In one of the two experiments reported, rats were initially trained to perform two responses, each of them reinforced by either a sucrose solution or a food pellet ( $R_1 \rightarrow O_1$ ,  $R_2 \rightarrow O_2$ ). Then subjects received backward trials in which each of the outcomes was followed by a different CS ( $O_1 \rightarrow A$ ;  $O_2 \rightarrow B$ ) and in the PIT test the rate of  $R_1$  and  $R_2$  performance was measured in the presence and absence of these CSs. If an excitatory association was formed between the CSs and the US, then the standard specific PIT should have been found, that is more  $R_1$  than  $R_2$  responses during A, and the reverse for B. However, Delamater and colleagues (2003) found the opposite pattern of results: CS presentations selectively reduced those responses that were trained with the same outcome as the CS. This suggests that an inhibitory CS-US association was formed during training. Furthermore, this association must have encoded specific sensory information of the US, otherwise the reduction of instrumental responding could not have been specific, supporting the conclusions of McNish et al. (1997) and

the predictions of the AESOP model (Wagner & Brandon, 1989). These results were later replicated by Laurent, Wong and Balleine (2014), who conducted an almost identical experiment. The main difference was that in the Pavlovian phase two groups of mice were trained: one group was trained with backward ( $O_1 \rightarrow A$ ;  $O_2 \rightarrow B$ ) and the other with forward ( $A \rightarrow O_1$ ;  $B \rightarrow O_2$ ) CS-US pairings. In the PIT test both A and B were presented and the results showed the specific PIT effect in the forward group, but the opposite pattern in the backward group, exactly paralleling the results reported by Delamater and colleagues (2003).

However, Cohen-Hatton et al. (2013) found the opposite pattern of results in a series of experiments similar to those described above (Delamater et al., 2003; Laurent et al., 2014). In one of their experiments, and after instrumental training ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ), rats received a Pavlovian conditioning phase in which one CS was followed and another preceded by an outcome ( $A \rightarrow O_1$ ;  $O_2 \rightarrow B$ ). In the PIT test the rate of  $R_1$  and  $R_2$  performance was measured in the presence of the CSs, and the results showed that B, not A, produced the specific PIT effect. If these results are interpreted under the same logic as Delamater et al. (2003) and Laurent et al. (2014) they suggest that an excitatory association was formed between B and a sensory representation of  $O_2$ , even though there was a backward relation between B and  $O_2$ .

A critical difference between the procedures used by Cohen-Hatton et al. (2013) and those used by Delamater et al. (2003) and

Laurent et al. (2014) was the length of the interval between the outcome delivery and the CS presentation. While both Delamater et al. (2003) and Laurent et al. (2014) presented the CS 10s after the outcome delivery, in Cohen-Hatton and colleagues' (2013) experiments the CS was presented only 1s after subjects accessed to the outcome. This might allow the processing of the outcome and the CS in a simultaneous rather than backward relationship. In terms of the AESOP model (Wagner & Brandon, 1989), if the outcome representation was still in the A1 state when the CS was presented, this would result in an excitatory association.

The main goal of this chapter was to explore if a CS trained in a backward relation with an outcome reduces the specific PIT effect. The possibility that the different results described above (Cohen-Hatton et al., 2013; Delamater et al., 2003; Laurent et al., 2014) were caused by the length of the interval between the outcome delivery and the CS presentation was also explored. According to the S-O-R accounts if backward pairings result in an inhibitory association between the CS and a sensory representation of the outcome then it is expected that this CS will selectively reduce performance of a response that has been trained with the same outcome. But if the CS acquires excitatory strength it should produce the standard specific PIT effect, i.e. selectively increase those responses that have been reinforced with the same outcome. In Experiments 9, 10 and 11, participants were trained to perform two responses, each of them reinforced by a different outcome ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). In the Pavlovian

phase participants observed the relationship between different CSs being either followed or preceded by one of these outcomes (A->O<sub>1</sub>; B->O<sub>2</sub>; O<sub>1</sub>->C; O<sub>2</sub>->D). Then in the PIT test the rate of R<sub>1</sub> and R<sub>2</sub> performance was assessed in the presence and absence of each of these CSs. In Experiment 9 the CS and the outcomes were presented immediately after each other, while in Experiment 10 and 11 an interval of 1s (Experiment 11) or 2s (Experiment 12) was inserted between the CS and the outcomes in both type of trials (forward and backward). A similar procedure was used in Experiments 12 and 13, except that participants had to perform a response during each of the outcome presentations in the Pavlovian phase, to make the direction of the association (either forward or backward) more salient. Finally in Experiment 14 CSs were trained in a forward and a backward relation, either with the same or a different outcome, in order to compare the contributions of backward and forward associations to the specific PIT effect directly. In the Pavlovian phase of this experiment each of the CSs was preceded and followed by an outcome; for one group of participants the outcomes were the same (e.g. O<sub>1</sub>->A->O<sub>1</sub>) and for the other group the outcomes were different (e.g. O<sub>1</sub>->A->O<sub>2</sub>). In all the experiments the degree to which participants learned the different Pavlovian relationships was also assessed.

## 4.8 Experiment 9

This experiment began with the instrumental phase (see Table 12), in which participants had to perform two responses, each of them reinforced with one outcome ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). Then in the Pavlovian phase participants received trials in which each of two CS was immediately followed by one of the outcomes ( $A \rightarrow O_1$ ;  $B \rightarrow O_2$ ), and another two CSs were preceded by these outcomes ( $O_1 \rightarrow C$ ;  $O_2 \rightarrow D$ ). In the PIT test participants received presentations of each of the CSs, while the rate of  $R_1$  and  $R_2$  performance was measured. According to the S-O-R accounts, if backward pairings produce inhibitory learning, then C and D should suppress the activation of  $O_1$  and  $O_2$ , respectively; thus no specific PIT effect should be observed in these trials. But if C and D acquired excitatory properties, then these cues should produce the specific PIT effect, i.e. more performance of  $R_1$  than  $R_2$  on C trials, and the reverse in D trials. Moreover, even if the backward pairings result in weaker S-O associations compared to those formed in the forward pairings, this should not affect the size of the specific PIT effect produced by these stimuli. This is because it has been shown that the strength of the S-O associations do not affect the specific PIT effect (e.g. Delamater & Oakeshott, 2007), then C and D should produce a similar effect to A and B, which were trained in a forward relation with the outcomes. In a final phase, participants' attention to the task was measured by asking them a series of questions about the relationships between each of the CSs and the

outcomes. In the previous experiments reported in this thesis participants were asked about these relationships during the Pavlovian phase, by answering which outcome was predicted by each of the CSs. However, it was not possible to ask similar questions about the backward pairings (because the CSs were presented after the outcomes) and for this reason the questions were asked at the end of the experiment.

*Table 12.* Design of Experiment 9.

Instrumental	Pavlovian	PIT test	Pavlovian test
R <sub>1</sub> ->O <sub>1</sub>	A->O <sub>1</sub>	A	A
R <sub>2</sub> ->O <sub>2</sub>	B->O <sub>2</sub>	B	B
	O <sub>1</sub> ->C	C	C
	O <sub>2</sub> ->D	D	D

*Note:* A, B, C and D: neutral fractal images; R<sub>1</sub> and R<sub>2</sub>: keyboard responses; O<sub>1</sub> and O<sub>2</sub>: food and drink images.

#### **4.8.1 Method**

*Participants.* Sixteen students from the University of Nottingham participated in this experiment (8 males and 8 females) aged between 19 and 38 years old.

*Apparatus and materials.* Only four images were used as CSs and these were presented in the centre instead of at each side of the screen. A, B, C and D were partially counterbalanced, resulting in 4 different counterbalancing conditions.

*Procedure.* Everything was the same as the previous experiment unless otherwise stated.

*Instrumental training.*

*Pavlovian phase.* Each trial began with the fixation dot and the text "*Press SPACE when you are ready*". Both remained on the screen until participants pressed the space bar. In the *forward* trials a CS was presented and 2s later it was replaced by the corresponding outcome, which remained on the screen for 2s. Then a new trial began. The *backward* trials were identical except that the outcome was presented before the CS. This phase consisted of two blocks, each of them with 4 trials of each of A-O<sub>1</sub>, B-O<sub>2</sub>, O<sub>1</sub>-C and O<sub>2</sub>-D.

*PIT test.* Each trial comprised 2s of the fixation cross (preCS period), followed by one of the CSs for 2s (CS period). Then the CS was replaced by the cross for 1s (ITI period), and then a new trial began. This test was divided in two blocks, each of them including 2 trials of each of A, B, C and D.

*Pavlovian test.* The question "*When you look at this image, what do you think more about?*" was positioned at the top of the screen with one of the CSs below it. A rating scale was presented at the bottom of the screen with the word "*Drink*" and a drink image at the left of the scale, and the word "*Food*" and a food image at the right of the scale. Participants had to select a point on the scale by using the mouse. This test was divided in two blocks, each with one question about each of A, B, C and D.

*Data analysis.* After the calculation of the PIT scores, these scores were grouped according to the type of training - that is, the PIT scores of A and B were grouped as forward trials, and the scores of C and D as backward trials. For the analysis of the Pavlovian test, the responses on the scale were transformed to values from 0 to 100, in which 100 represents the selection of the correct outcome (i.e. O<sub>1</sub> for A/C and O<sub>2</sub> for B/D) and 0 the opposite outcome.

#### 4.8.2 Results

*Instrumental training.* All participants successfully completed this phase (see Table 13). No significant differences were found between the type of response or outcome,  $F_s < 1$ .

*Table 13.* Mean number of R1 and R2 responses and mean number of O1 and O2 deliveries in the instrumental phase of Experiments 9, 10, 11, 12 and 13.

	R <sub>1</sub>	R <sub>2</sub>	O <sub>1</sub>	O <sub>2</sub>
<b>Experiment 9</b>	248	247	49.9	51.3
<b>Experiment 10</b>	247	246	50.5	51.3
<b>Experiment 11</b>	243	259	49.3	52.2
<b>Experiment 12</b>	260.9	270.7	51.4	51.7
<b>Experiment 13</b>	248.1	255.2	50	50

*PIT test.* The PIT scores are presented in Figure 27, which shows a clear specific PIT effect for both type of trials. An ANOVA with block, congruency and trial type (forward, backward) revealed a significant main effect of congruency,  $F(1, 15) = 8.26, p = .012, MSe = 87.72, \eta_p^2 = .36$ . Nothing else was significant, largest  $F(1, 15) = 1.61, p = .224, MSe = 2.68$ .

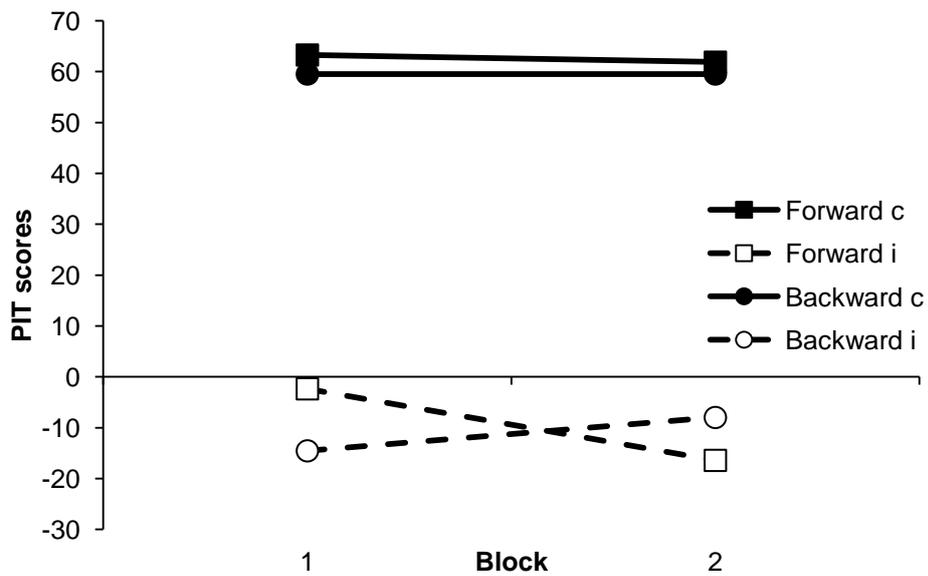


Figure 27. Mean rate of congruent and incongruent responses for forward and backward trials in the PIT Test of Experiment 9.

A corresponding ANOVA was conducted on the data of the preCS period (Table 14), which showed no significant main effect or interaction, largest  $F(1, 15) = 2.72, p = .12, MSe = 2.17$ .

Table 14. Mean preCS response rates in each block of the PIT test of Experiments 9, 10, 11, 12 and 13.

	Congruent		Incongruent	
	1	2	1	2
<b>Experiment 9 Forward</b>	24.8	40.8	37.5	45.5
<b>Experiment 9 Backward</b>	37.5	50.2	40.8	35.6
<b>Experiment 10 Forward</b>	51.5	44.7	39.1	45.7
<b>Experiment 10 Backward</b>	36.2	46.3	42.1	54.5
<b>Experiment 11 Forward</b>	62.9	48.3	55.1	77.6
<b>Experiment 11 Backward</b>	72.1	54.1	44.7	72.4
<b>Experiment 12 Forward</b>	71.3	71.3	82	72.5
<b>Experiment 12 Backward</b>	68.6	72	84.5	77.5
<b>Experiment 13 Forward</b>	108.1	117.8	130.9	134.7
<b>Experiment 13 Backward</b>	120	132.2	114.4	116.6

*Pavlovian test.* The mean participants' ratings to the forward and backward trials were 74 and 79, respectively. The statistical analysis confirmed no difference between these ratings, largest  $F(1, 15) = 1.27$ ,  $p = .28$ ,  $MSe = .02$ . This suggests that participants identified the CS-Outcome relationship with the same accuracy regardless of whether the CSs were trained in a backward or forward relation.

#### 4.8.3 Discussion

The results of the PIT showed that the CSs trained in a backward relation with the outcome produced the specific PIT effect, and that this effect was the same as that produced by the CSs trained in forward conditioning. These results suggest that backward training produced excitatory conditioning as effectively forward training, which is inconsistent with the reports of Delamater et al. (2003) and Laurent et al. (2014), but not with those of Cohen-Hatton and colleagues (2013). However, one interpretation of these results can be made by

using the AESOP model (Wagner & Brandon, 1989). According to this model, an excitatory CS-US association is formed if the processing of the US is still in the A1 state at the moment of CS presentation. This is possible considering that in the backward trials of this experiment the outcome delivery was immediately followed by the CS presentation. If this interpretation is correct, then increasing the temporal distance between the outcome and the CS should increase the likelihood of producing an inhibitory association.

#### **4.9 Experiment 10**

This experiment was identical to Experiment 9, except that a 1-s interval was included between the CS and the outcome presentations in both forward and backward trials. In this period no stimuli were presented on the screen. The inclusion of a similar 1-s interval in the forward trials was meant to maintain the same conditions in both types of conditioning.

*Participants.* Twenty-four students from the University of Nottingham participated in this experiment (4 males and 20 females) aged between 18 and 22 years old. One participant was removed because they only performed one response in instrumental training.

#### 4.9.1 Results

*Instrumental training.* The participants successfully completed this phase (see Table 13). No differences were found between the type of response performed nor the type of outcome received,  $F_s < 1$ .

*PIT test.* The PIT scores are presented in Figure 28. Although the graph suggests that the specific PIT effect was smaller in the forward trials, this was not confirmed by the statistical analysis. An ANOVA with block, congruency and trial type revealed a significant main effect of congruency,  $F(1, 22) = 8.79$ ,  $p = .007$ ,  $MSe = 69.4$ ,  $\eta_p^2 = .29$ , that interacted with block,  $F(1, 22) = 5.34$ ,  $p = .031$ ,  $MSe = 12.76$ ,  $\eta_p^2 = .20$ , and a significant Block x Trial interaction,  $F(1, 22) = 8.42$ ,  $p = .008$ ,  $MSe = 3.93$ ,  $\eta_p^2 = .28$ . Nothing else was significant, largest  $F(1, 22) = 1.6$ ,  $p = .22$ ,  $MSe = 8.49$ . The analysis of the Congruency x Block interaction revealed a significant effect of congruency at both blocks, smallest  $F(1, 22) = 5.87$ ,  $p = .024$  for block 1, and a significant effect of block at congruent,  $F(1, 22) = 5.26$ ,  $p = .032$ , but not at incongruent,  $F < 1$ .

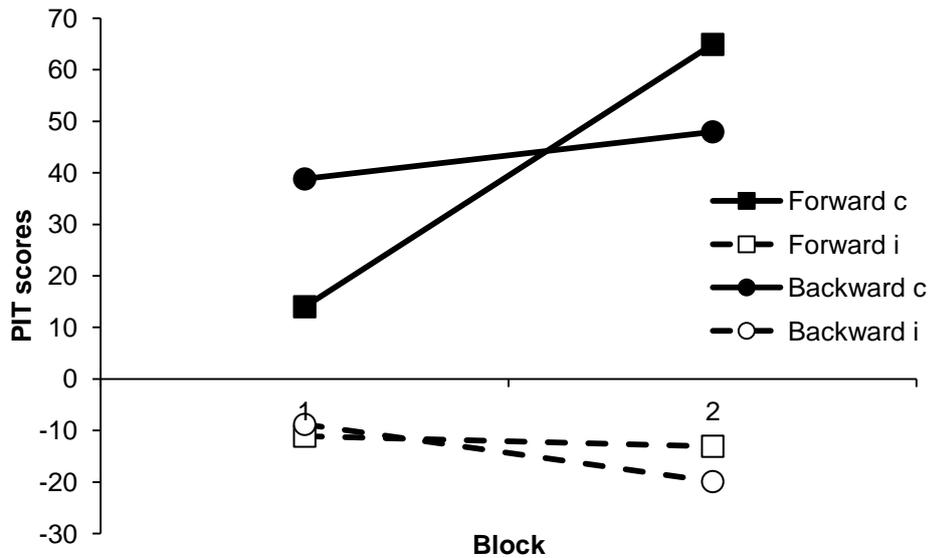


Figure 28. Mean rate of congruent and incongruent responses for forward and backward trials in the PIT Test of Experiment 10.

The mean rates of responding during the preCS period are presented in Table 14, and the statistical analysis revealed no significant main effect or interaction, largest  $F(1, 22) = 2.59, p = .12, MSe = 3.45$ .

*Pavlovian test.* The mean ratings to the forward and backward trials were 85.4 and 87.4 respectively. This difference was not significant,  $F < 1$ .

#### 4.9.2 Discussion

The results of this experiment were identical of those found in Experiment 9. CSs trained in a backward relation produced the

specific PIT effect in the same way as did CSs trained in a forward relation with the outcome. This experiment assessed the idea that a temporal gap between the outcome and the CS would result in inhibitory backward associations, but the results indicate that this modification had no effect on the excitatory properties acquired by the CSs. However, the possibility must be considered that the 1-s interval included between the CS and the outcome was not sufficient to produce an observable difference. It might be that increasing this interval further would produce different results. This idea was assessed in the next experiment.

#### **4.10 Experiment 11**

In this experiment the interval between the CSs and the outcomes was increased to 2s. Also, in order to have a better assessment of learning of the CS-outcome associations, the Pavlovian test was presented at intervals throughout the Pavlovian phase instead of just at the end of the experiment.

##### **4.10.1 Method**

*Participants.* Twenty-four students from the University of Nottingham participated in this experiment (4 males and 20 females) aged

between 18 and 25 years old. One participant was excluded because an experimenter error, i.e. the participant was administered a different version of the task.

*Procedure.*

*Instrumental training.*

*Pavlovian phase.* This phase was divided in four blocks, each of them comprising 2 trials each of A-O<sub>1</sub>, B-O<sub>2</sub>, O<sub>1</sub>-C and O<sub>2</sub>-D. A Pavlovian test identical to those used in the previous two experiments was presented after each of these blocks.

*PIT test.*

#### **4.10.2 Results**

*Instrumental training.* All participants completed this phase successfully (see Table 13). No differences were found either between the type of response performed or type of outcome delivered, largest  $F(1, 22) = 3.76, p = .065, MSe = 24.44$  for the outcome deliveries.

*Pavlovian phase.* The participants' mean ratings in each of the Pavlovian tests are presented in Figure 29. Values close to 100 represent the choice of the correct outcome, while values close to 0 represent the incorrect choice. The Figure suggests that participants were equally accurate in identifying the CS-outcome relationship in both type of trials from the beginning of the phase. An ANOVA with

block (1-4) and trial type (forward, backward) as factors revealed no significant main effect or interaction, largest  $F(3, 66) = 2.39, p = .076, MSe = 242.31$ .

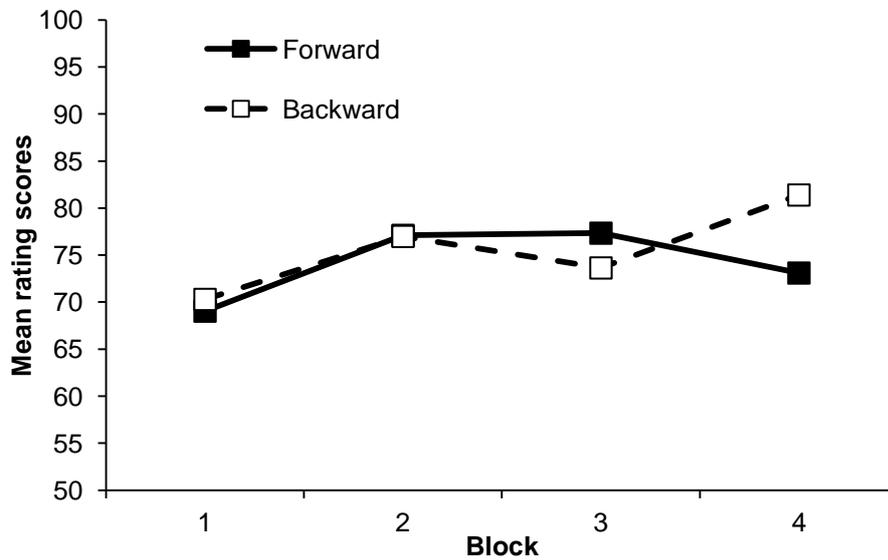


Figure 29. Mean rating scores grouped by type of trial (forward; backward) in the Pavlovian tests of Experiment 11.

*PIT test.* The PIT scores are presented in Figure 30, which suggests that again both type of trials produced the specific PIT effect, although it seems diminished in the first half of the test. An ANOVA with block, congruency and trial type revealed a significant main effect of congruency,  $F(1, 22) = 8.41, p = .008, MSe = 75.8, \eta_p^2 = .28$ , which interacted with block,  $F(1, 22) = 6.16, p = .021, MSe = 32.02, \eta_p^2 = .22$ . Nothing else was significant,  $F_s < 1$ . The analysis of the significant interaction showed an effect of congruency at block 2,  $F(1, 22) = 12.2, p = .002$ , but not at block 1,  $F(1, 22) = 1.41, p = .248$ .

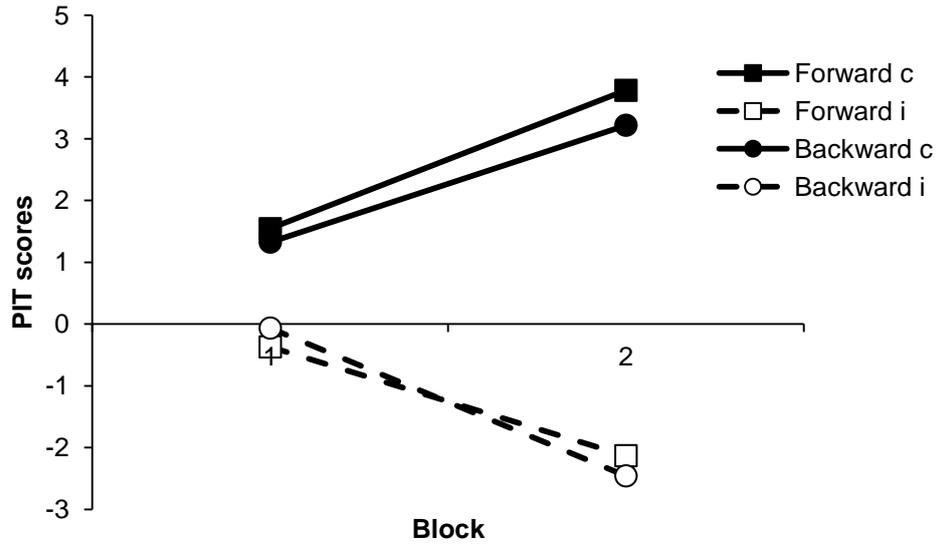


Figure 30. Mean rate of congruent (c) and incongruent (i) responses for forward and backward trials in the PIT Test of Experiment 11.

The mean rates of responding during the pre-CS are presented in Table 14. It seems that in the first block participants performed more congruent than incongruent responses, which could have affected the PIT scores, reducing the specific PIT effect in the first block of the test. This pattern was reversed in the second block of the pre-CS period, i.e. more incongruent than congruent responses. However, the statistical analysis showed no significant main effect or interaction, largest  $F(1, 22) = 4.03, p = .057, MSe = 21.76$  for the Block x Congruency interaction.

### 4.10.3 Discussion

The results of this experiment are consistent with the findings of Experiments 9 and 10. The CSs produced the specific PIT effect regardless of whether these CSs were trained in a forward or backward relation with the outcome. This experiment aimed to assess if the inclusion of a larger temporal gap between the CS and the outcome would result in an inhibitory association, as in the experiments reported by Delamater and others (2003) and Laurent and colleagues (2014). However, the inclusion of this interval eliminated the specific PIT effect in the first block of the test, but for both types of CS.

One interpretation is that in this task backward pairings produced excitatory conditioning in the same way as forward pairings. According to the S-O-R accounts, both types of CSs retrieved an outcome representation in the PIT test, which in turn produced the selective elevation of performance. A second interpretation is in the line of the S-R mechanisms proposed by Cohen-Hatton et al. (2013). According to this account, during the Pavlovian phase an association was formed between the CS and a representation of the response activated by the outcome delivery, in both forward (e.g.  $A \rightarrow O_1 \rightarrow [R_1]$ ) and backward trials (e.g.  $O_1 \rightarrow [R_1] \rightarrow C$ ). Thus, A and C should elicit  $R_1$  and  $R_2$  at test, respectively. However, the problem is that this account predicts a larger specific PIT effect for C than for A, i.e. the CS trained

in a backward relation with the outcome, which was not observed in the experiments reported here. This is because during training the contiguity between C and the evoked  $R_1$  representation is likely to be greater than between A and the  $R_1$  representation, which should result in a stronger S-R association in the backward pairings. This prediction was supported by the results reported by Cohen-Hatton et al. (2013) described in the introduction of this chapter. However, in the experiments presented here the effect of the CSs in the PIT test was the same, regardless of the type of conditioning (forward or backward).

Another possibility is that participants processed the CSs and the outcomes simultaneously, even with the inclusion of the interval. In the Pavlovian phase participants passively watched the screen while the stimuli were presented and the trials were clearly separated from each other. This might have caused the participants to identify each of these trials as one CS-US relationship, without paying attention to the direction of this association. For example, in the backward trials they might have waited until the CS presentation and then thought about the previous outcome in order to remember the relationship (Arcediano, Escobar & Miller, 2003; Jants & Underwood, 1958). This would result in a symmetrical association between the CS-US association for both forward and backward trials, which would explain why both types of CSs produced a comparable specific PIT effect. Similarly, and appealing to the AESOP model (Wagner & Brandon, 1989), it is possible that in each of the trials of the Pavlovian phase the outcome processing was still in the A1 state when the CS was

presented in both backward and forward pairings. As it was described above, this model states that excitatory associations are formed when the CS are processed while the outcome representations are in the A1 state, but if the CS are processed after these representations decay into the A2 state, then the associations are inhibitory. Thus the inclusion of the gap between the outcome and the CS presentation should increase the chances that the outcome processing decays into A2 state at the time the CS is presented. However, it is possible that even in Experiment 11 the gap between the stimuli was not enough for this to occur. Moreover, in the experiments using rats the interval between the outcome delivery and the CS was 10 seconds, while in Experiment 2 was only 2 seconds. This difference is larger considering that humans are likely to retain a memory of the outcomes for a longer period than rats.

#### **4.11 Experiment 12**

The Pavlovian phase of this task was modified in order to foster a differential processing of the CSs trained in a backward and forward relation with the outcomes. In the Pavlovian phase, participants had to perform a response each time the outcomes were presented. If the outcome was a drink image participants had to press the 't' key, but if it was a food image they had to press the 'g' key. The reaction time was measured and reported to the participants as feedback on each of the

trials. Because in the forward trials the CS provided information about the outcome that followed, it should have facilitated the performance of the response. But in the backward trials the CS did not contribute to the response required. Thus, if participants engaged in this task they should learn to distinguish between the CSs paired in a forward relationship with the outcomes and those paired in a backward relationship. This should be reflected in lower reaction times in the forward than in the backward trials.

Also in each trial of the Pavlovian phase two identical CSs were presented, each of them at one side of the fixation dot. The purpose of this was to keep participants' attention on the centre of the screen so they did not have to look for the CS. These identical CSs were also presented in the PIT test in order to keep consistency across the task. In addition to this, the number of Pavlovian trials was increased by half.

A final modification was the replacement of the Pavlovian tests by a questionnaire about the relationships of the CSs and instrumental responses with the outcomes. Instead of being presented with the rating scales, when participants were presented with a CS they had to select between three options, i.e. food, drink or nothing. This test was conducted at the end of the experiment.

#### 4.11.1 Method

*Participants.* Thirty-five students from Brooklyn College participated in this experiment (18 males and 17 females), aged between 18 and 41 years old. Three of them were excluded from the experiment: one interrupted the task during the PIT test, another failed to complete the instrumental training, and another did not perform any response during Pavlovian training. All the students received course credit for their participation.

*Procedure.*

*Instrumental training.*

*Pavlovian phase.* Participants were informed that some neutral images and rewards (drink and food pictures) would appear on the screen. Their goal was to press one key ('t') each time they saw a drink picture, and a different key ('g') when they saw a food picture. They were also informed that the neutral images could appear before or after the rewards, and that later in the experiment they would have to answer questions about all the images.

Each trial began with the fixation dot and the text "T' = Drink" above it in orange letters, and the text "G' = Food" below the dot in blue letter. The texts disappeared after 3s, leaving the dot alone for another 1s. In the *forward* trials this was followed by two identical CS images, each of them presented to one side of the dot. After 3s, the CSs and the dot were replaced by one of the outcomes, which was

positioned in the centre of the screen. During the outcome presentation participants had to press one of the keys ('g' or 't'), receiving feedback after their response. If the response was correct, the text "*Correct! RT =*" with the corresponding reaction time (measured since the outcome onset) appeared on top of the outcome in green letters. But if the response was incorrect or no response was performed after 2s, the text "*Oops! That was wrong*" appeared below the outcome in red letters. In all cases the feedback and the outcome remained on the screen for 1s, after which a new trial began. The same occurred on the backward trials, except that these began with the outcome presentation. Immediately after the 1-s feedback, the CSs were presented on the screen for 3s and then a new trial began. This phase was divided in two blocks, each of them comprising 6 trials each of A->O<sub>1</sub>, B->O<sub>2</sub>, O<sub>1</sub>->C and O<sub>2</sub>->D. A, B, C and D were partially counterbalanced, resulting in 4 different counterbalancing conditions.

*PIT test.* The fixation cross was present throughout this phase. Each trial began with the cross alone on the screen (preCS period) and after 3s two identical CSs were positioned on each side of the cross (CS period), which were removed after 3s. After an additional 1s of the cross alone on the screen (ITI period) a new trial started. The test was divided in two blocks, each of them with 2 trials of each of A, B, C and D.

*Assessment questionnaire.* Participants had to answer a series of questions about the relationship between the CSs, the instrumental responses and the outcomes, by pressing one of three keys (1, 2 or

3). For the Pavlovian relationships the question "*This image was followed/preceded by*" was positioned at the top of the screen and one of the CSs below it. For the instrumental relationships the question was "*The Z/M key was followed by*" and no image was presented. For all the questions, the text "*1) Drink 2) Food 3) Dot/Cross*" appeared at the bottom of the screen. This was divided in two blocks, each of them with 2 trials each of A, B, C, D, R<sub>1</sub> and R<sub>2</sub>.

*Data analysis.* The responses made during the Pavlovian phase were grouped according to trial type (forward, backward). Then they separated by correct and incorrect responses: incorrect responses were defined as those in which participants either pressed the incorrect key, pressed a key during the CS presentation, or did not press any key during the outcome. Only the reaction times for the correct responses were analysed.

In the case of the assessment questionnaire, the responses were grouped into *correct* and *incorrect outcome* responses. For instance, if the question was about A, the correct outcome was O<sub>1</sub> and the incorrect outcome O<sub>2</sub>. Those answers in which participants chose the third option *cross/dot* were excluded from the analysis.

#### 4.11.2 Results

*Instrumental training.* All participants successfully completed this phase (see Table 13). No significant differences were found between the type of response performed nor the type of outcome received by the participants, largest  $F(1, 31) = 1.54, p = .22, MSe = 993.25$ .

*Pavlovian phase.* The mean percentages of correct responses were grouped in two blocks of six trials and presented on the top panel of Figure 31. The graph shows better accuracy in the second half of this phase, regardless of the type of trial. An ANOVA with block (1, 2) and trial type (forward, backward) showed a significant main effect of block,  $F(1, 31) = 4.47, p = .043, MSe = 58.7, \eta_p^2 = .13$ . Nothing else was significant, largest  $F(1, 31) = 1.04, p = .32, MSe = 52.01$ . The mean reaction times were similarly grouped and presented on the bottom panel of Figure 31. This Figure suggests that participants' reaction times decreased during training. It also appears that participants were faster on the backward trials but this was not confirmed by the statistical analysis. The corresponding ANOVA revealed a significant main effect of trial block,  $F(1, 31) = 4.53, p = .041, MSe = 0.006, \eta_p^2 = .13$ . Nothing else was significant, largest  $F_s < 1$ . The fact that participants' reaction times did not differ on both type of trials suggests that they did not use the information provided by the CS to predict the outcomes. If they had, the reaction times for the

forward trials should have been lower than those of the backward trials.

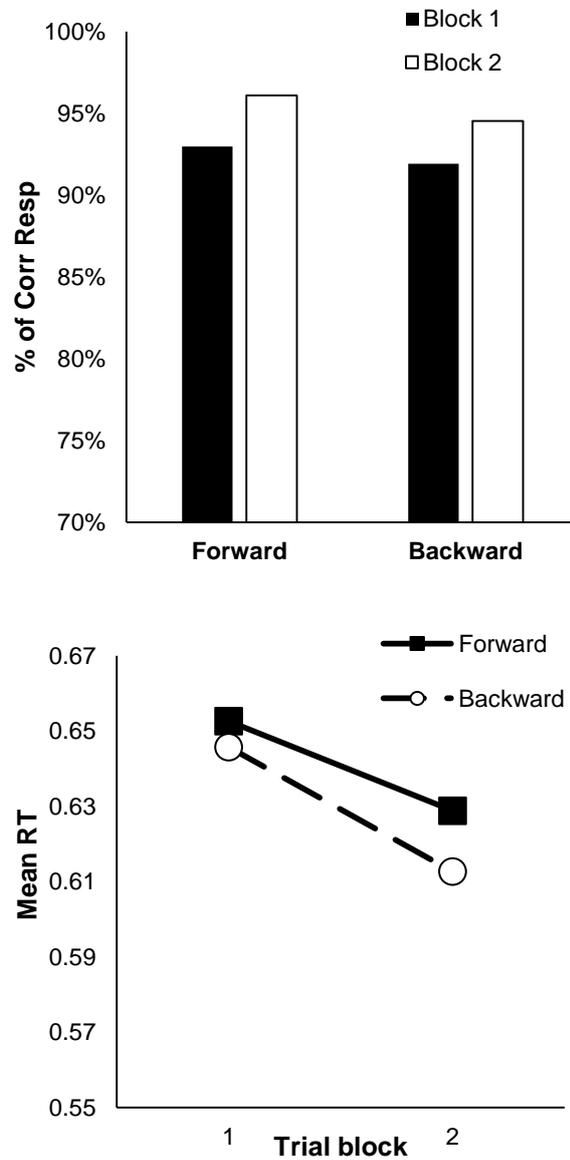


Figure 31. Responses in the forward and backward trials of the Pavlovian phase in Experiment 12. Top panel: the mean percentage of correct responses. Bottom panel: the mean reaction times.

*PIT test.* The PIT scores are presented in Figure 32. Unlike in the previous experiments, the figure suggests that the specific PIT effect was found in the forward trials but it was reduced in the backward trials. However, this was not confirmed by the statistical analysis. An ANOVA with block, congruence and trial type (forward, backward) showed that neither the Trial type x Congruence interaction,  $F(1, 31) = 2.35, p = .14, MSe = 15.5$ , nor the Trial type x Congruence x Block interaction,  $F < 1$ , were significant. It did reveal a significant Block x Trial type interaction,  $F(1, 31) = 5.23, p = .028, MSe = 9.02, \eta_p^2 = .14$ . Nothing else was significant, largest  $F(1, 31) = 3.27, p = 0.08, MSe = 45.7$ . The analysis of the significant interaction showed an effect of trial type at block 2,  $F(1, 31) = 6.64, p = .015$ , but not at block 1,  $F(1, 31) = 1.97, p = .17$ .

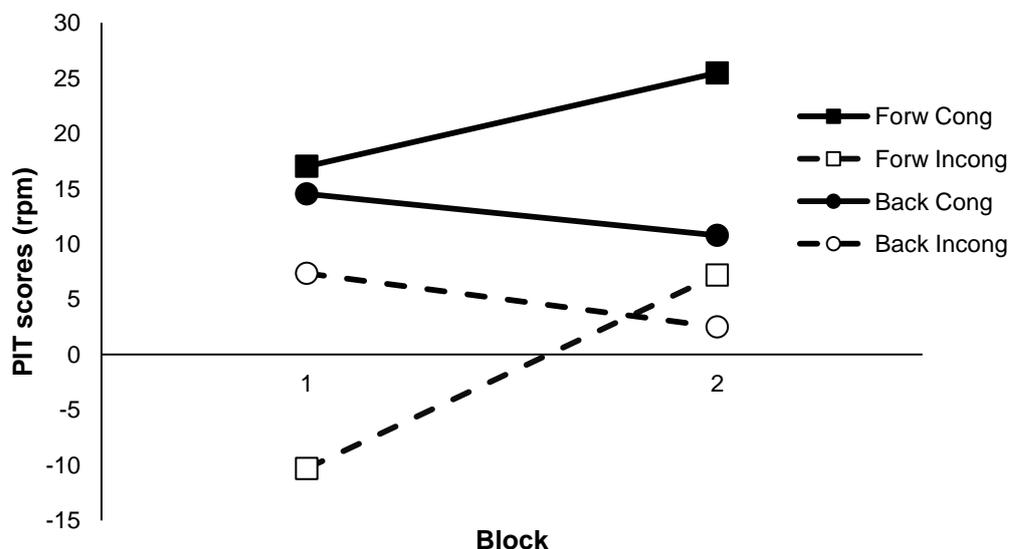
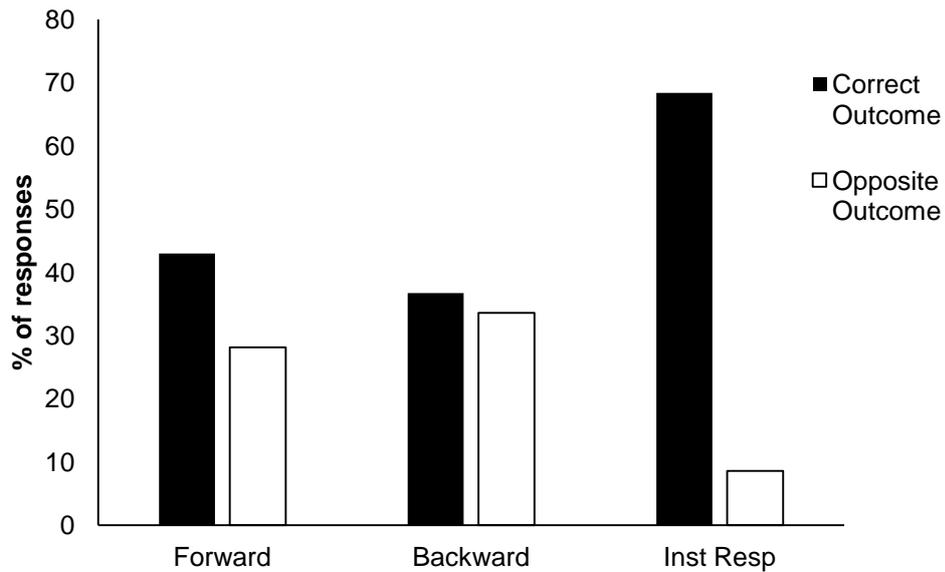


Figure 32. Mean rate of congruent and incongruent responses for forward and backward trials in the PIT Test of Experiment 12.

The mean rates of responding during the preCS period are presented in Table 14. The statistical analysis showed no significant main effect nor interaction, largest  $F(1, 31) = 2.95, p = .096, MSe = 15.15$ .

*Assessment questionnaire.* The mean percentage of correct and incorrect outcome responses are presented in Figure 33. The graph shows more correct than incorrect answers to the questions about the instrumental contingencies. It also suggests that participants were slightly more accurate in the questions about the forward relationships. However, this was not supported by the statistical analysis. An ANOVA with question (forward, backward, instrumental) and response (correct, incorrect) as factors showed a significant main effect of response,  $F(1, 31) = 17.32, p < .001, MSe = 1861, \eta_p^2 = .36$ , that interacted with Question,  $F(2, 62) = 11.44, p < .001, MSe = 1249, \eta_p^2 = .27$ . Simple main effects on the interaction showed a significant effect of response for the instrumental questions,  $F(1, 31) = 60.72, p < .001$ , but not for questions about the forward and backward trials, largest  $F(1, 31) = 2.01, p = .17$ .



*Figure 33.* Mean percentage of correct and incorrect outcome responses in the assessment questionnaire of Experiment 12.

#### 4.11.3 Discussion

Unlike in the previous experiments, the CSs did not produce the specific PIT effect. In this Experiment participants had to perform a response each time they saw an outcome on the screen in the Pavlovian phase. Thus it is possible that participants ignored the CS presentations in this phase, focusing on performing the responses. This idea is supported by the fact that participants' reaction times were not lower on the forward trials, even when the CS predictor of the outcome should have facilitated responding. Additionally, the results of the assessment questionnaire showed a good discrimination of the R-O associations, but no evidence of learning about the Pavlovian relationships.

#### 4.12 Experiment 13

In Experiment 13 the instructions of the Pavlovian phase were slightly modified in order to increase participants' attention to the CS-outcome relationships. Also, the set of food and drink images used in all the previous experiments was replaced by a single picture of one food and one drink image. Because one of the indices of Pavlovian conditioning in this experiment was reaction time, it was thought that having a set of images of the same category might result in more time being required to identify the category to which the image belonged (food or drink). Thus, by using only one image per outcome the reaction times should be reduced.

Additionally, this experiment explored the effect of training each of the instrumental responses separately. As has been described in this thesis, the S-O-R accounts assume that a CS activates a representation of the outcome, and it is this representation that elicits responding. Although some authors have suggested that this is possible via a bidirectional R-O association (e.g. Rescorla, 1994b), others have proposed that O-R associations are also formed during training (e.g. Balleine & Ostlund, 2007). In all the experiments reported here, both instrumental responses have been trained simultaneously, which implies that consistent associations were formed between the responses and the outcomes that followed them (forward R-O associations), but not between the responses and the outcome that preceded them (backward O-R associations). This is because when

two responses are trained in a concurrent training each of the responses is reinforced with only one of the outcomes, i.e.  $R_1 \rightarrow O_1$  and  $R_2 \rightarrow O_2$ , but each of the outcomes might precede the performance of any of both responses, e.g.  $O_1 \rightarrow R_1$ ,  $O_1 \rightarrow R_2$ ,  $O_2 \rightarrow R_1$  and  $O_2 \rightarrow R_2$ . The fact that in all the experiments presented here (except the previous) the specific PIT effect was found is consistent with the idea that an outcome representation elicits performance of the response encoded in R-O associations (e.g. Rescorla, 1994b). If the mechanism behind this phenomenon were based solely on an O-R association then an outcome representation, in the tasks reported here, should have elicited both responses equally; thus no specific PIT should have been observed in any of the experiments.

Nevertheless, this does not necessarily mean that O-R associations do not contribute to the specific PIT effect. It might be possible that this effect is mainly determined by R-O associations but that possible O-R associations also contribute to the phenomenon. If this idea is correct, then consistent O-R associations might increase the size of the specific PIT effect. For instance, activation of an outcome representation might elicit responding more easily when that response was preceded and followed by the same outcome during training, e.g.  $O_1 \rightarrow R_1 \rightarrow O_1$ .

#### **4.12.1 Method**

*Participants.* Seventeen students from Brooklyn College participated in this experiment (5 males and 12 females) aged between 18 and 28 years old. One participant was excluded because he failed to perform any response in the Pavlovian phase.

Procedure.

*Instrumental training.* Participants were instructed to press 'z' to obtain as many pictures of drinks as they could. After they obtained 25 outcome deliveries they were instructed to press 'm' to obtain pictures of food, which also ended after they received 25 outcome. This cycle was then repeated, so participants received 50 presentations of each outcome by the end of this phase.

*Pavlovian phase.* In addition to the instructions presented in the previous experiment, participants were informed that the neutral images might help them to predict the outcome and that they should not press any key during the presentations of these images.

*PIT test.*

*Assessment questionnaire.*

#### 4.12.2 Results

*Instrumental training.* All participants completed this phase successfully (see Table 13). No differences were found between type of response performed nor type of outcome delivered,  $F_s < 1$ .

*Pavlovian phase.* The mean percentages of correct responses are presented on the top panel of Figure 34. As in the previous experiment, statistical analysis failed to detect significant differences in participants' accuracy on both types of trial. An ANOVA with trial type (forward, backward) and trial block (1, 2) showed no significant main effect or interaction, largest  $F(1, 15) = 4$ ,  $p = .064$ ,  $MSe = 69.4$  for the interaction. The mean reaction times are presented on the bottom panel of Figure 34, which suggests lower reaction times in the forward than in the backward trials, but this was not confirmed by the statistical analysis. The corresponding ANOVA showed no significant main effect or interaction, largest  $F(1, 15) = 1.21$ ,  $p = .29$ ,  $MSe = .022$ . Although the results are not significant, they seem to be in the expected direction, i.e. faster performance in the forward than in the backward trials.

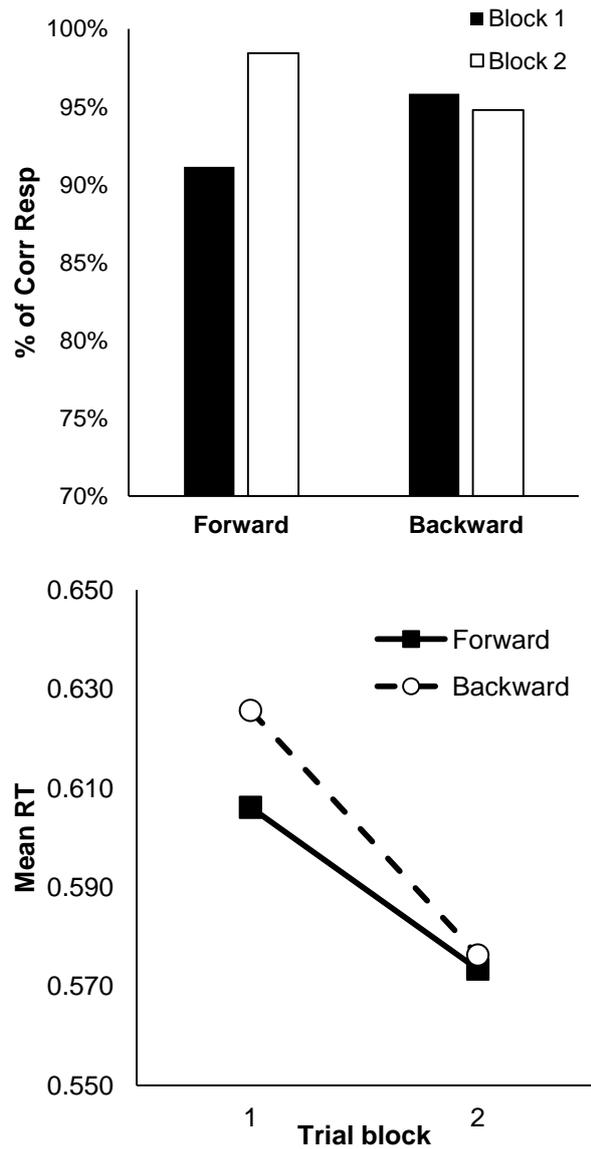


Figure 34. Responses in the forward and backward trials of the Pavlovian phase in Experiment 13. Top panel: the mean percentage of correct responses. Bottom panel: the mean reaction times.

*PIT test.* The PIT scores are presented in Figure 35, which shows a clear difference between congruent and incongruent responses in the forward trials, i.e. the specific PIT effect, but the opposite in the backward trials. An ANOVA with block, congruence and trial type (forward, backward) showed a significant Congruence x Trial type

interaction,  $F(1, 15) = 5.02, p = .041, MSe = 36.7, \eta_p^2 = .25$ . Nothing else was significant, largest  $F(1, 15) = 1.35, p = .26, MSe = 39.34$ . Simple main effects on the significant interaction revealed an effect of congruence on forward trials,  $F(1, 15) = 5.05, p = 0.04$ , but not on backward trials,  $F < 1$ . Unlike in the previous experiments, the CSs trained in a backward procedure did not produce the specific PIT effect.

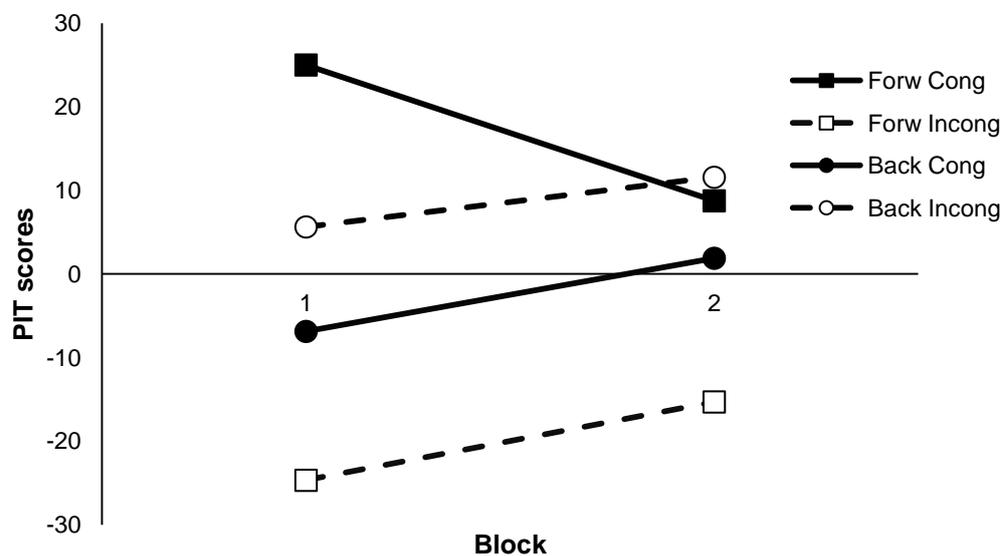
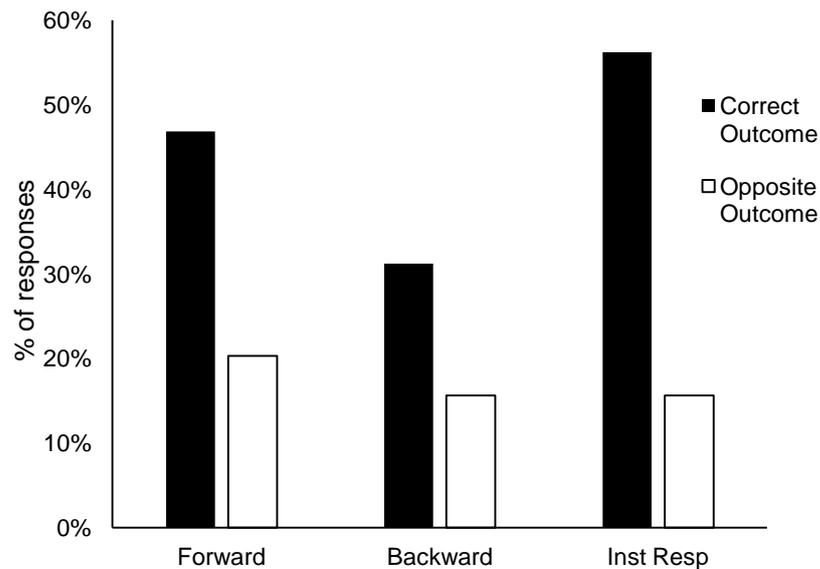


Figure 35. Mean rate of congruent and incongruent responses for forward and backward trials in the PIT Test of Experiment 13.

The analysis of the mean rates of responding during the preCS period (see Table 14) revealed no significant main effect nor interaction, largest  $F(1, 15) = 1.59, p = .23, MSe = 46.44$ .

*Assessment questionnaire.* The mean percentage of correct and incorrect outcome responses are presented in Figure 36, which shows more correct than incorrect responses in the questions referring to the instrumental responses and also about the forward and backward

trials. An ANOVA with question (forward, backward, instrumental) and response (correct, incorrect) as factors showed a significant main effect of response,  $F(1, 15) = 14.51, p = .002, MSe = 1259, \eta_p^2 = .49$ . Nothing else was significant, largest  $F(2, 30) = 1.73, p = .19, MSe = 815.54$ .



*Figure 36.* Mean percentage of correct and incorrect outcome responses in the assessment questionnaire of Experiment 13.

#### 4.12.3 Discussion

The results of the PIT test showed that only the CSs trained in a forward relation with the outcome produced the specific PIT effect. Although the CSs trained in a backward relation produced the opposite tendency, the fact that the difference was not significant does not allow us to conclude that they acquired inhibitory properties. However, these

results suggest that in this procedure the CSs trained in a backward relation with the outcome do not acquire excitatory strength.

It might be argued that in this version of the task participants simply ignored the CSs when they were presented after the outcomes (i.e. backward trials) but the results of the assessment questionnaire indicate that they were equally accurate in identifying the CS-outcome relationships in both forward and backward trials. These results support the idea that participants learned both backward and forward relationships, but only the forward training produced excitatory conditioning.

#### **4.13 Experiment 14**

A perhaps more sensitive comparison between forward and backward training is to train a CS with both types of relationships with the outcomes. In this experiment each CS was both preceded and followed by an outcome and participants had to perform a response ('t' or 'g') in each of these outcome presentations (see Table 15). Thus, after performing the first response they had to pay attention to the CS to use the information provided by the CS to predict the next outcome. This should foster the formation of both backward and forward associations.

For the experimental group the outcome that preceded the CS was different to the outcome that followed. This group received

presentations of two CSs that were preceded by  $O_1$  and followed by  $O_2$  (i.e.  $O_1 \rightarrow A \rightarrow O_2$ ;  $O_1 \rightarrow B \rightarrow O_2$ ) and another two CSs were paired with the outcomes in the reverse order (i.e.  $O_2 \rightarrow C \rightarrow O_1$ ;  $O_2 \rightarrow D \rightarrow O_1$ ). The number of CSs was doubled in order to raise the difficulty of the task, keeping participants interested. Additionally, the number of Pavlovian trials was increased.

If these backward pairings produce excitatory conditioning then at test the CSs should activate a representation of the outcome that preceded and followed the CS during training simultaneously. For instance, A was preceded by  $O_1$  and followed by  $O_2$  so then at test should activate a representation of both  $O_1$  and  $O_2$ . If this occurs then it should result in either an elevation of both  $R_1$  and  $R_2$ , i.e. general PIT, or in an elimination of the specific PIT effect due to response competition. In contrast, if backward pairings result in inhibitory conditioning, in the PIT test A should suppress the activation of  $O_1$  while it activates  $O_2$ . This should result in a larger specific PIT effect than that produced by a CS that was preceded and followed by the same outcome. To assess this idea, a control group in a similar manner to the experimental group, except that the outcomes that preceded and followed the CSs were the same (i.e.  $O_1 \rightarrow A \rightarrow O_1$ ;  $O_1 \rightarrow B \rightarrow O_1$ ;  $O_2 \rightarrow C \rightarrow O_2$ ;  $O_2 \rightarrow D \rightarrow O_2$ ).

Overall if backward pairings produce excitatory conditioning then it is expected that the CSs will either reduce or eliminate the specific PIT effect in the experimental group compared to that seen in the control group. Alternatively, if backward training produces

inhibitory CSs, then the CSs should produce a larger PIT effect in the experimental than in the control group.

*Table 15.* Design of Experiment 14.

	Instrumental		Pavlovian	PIT test	Assessment
Experimental	R <sub>1</sub> ->O <sub>1</sub>	R <sub>2</sub> ->O <sub>2</sub>	O <sub>1</sub> ->A->O <sub>2</sub>	A	A, B, C, D
			O <sub>1</sub> ->B->O <sub>2</sub>	B	R <sub>1</sub> , R <sub>2</sub>
			O <sub>2</sub> ->C->O <sub>1</sub>	C	
			O <sub>2</sub> ->D->O <sub>1</sub>	D	
Control	R <sub>1</sub> ->O <sub>1</sub>	R <sub>2</sub> ->O <sub>2</sub>	O <sub>1</sub> ->A->O <sub>1</sub>	A	A, B, C, D
			O <sub>1</sub> ->B->O <sub>1</sub>	B	R <sub>1</sub> , R <sub>2</sub>
			O <sub>2</sub> ->C->O <sub>2</sub>	C	
			O <sub>2</sub> ->D->O <sub>2</sub>	D	

*Note:* A, B, C and D: neutral fractal images; R<sub>1</sub> and R<sub>2</sub>: keyboard responses; O<sub>1</sub> and O<sub>2</sub>: food and drink images.

#### 4.13.1 Method

*Participants.* Thirty-two students from Brooklyn College participated in this experiment (8 males and 24 females) aged between 19 and 58 years old.

*Procedure.* Everything was the same as in Experiment 13 for both groups unless otherwise stated.

*Instrumental training.*

*Pavlovian phase.* An outcome was presented before and after each CS, and participants had to press 't' for drink images and 'g' for food images for each outcome. This phase was divided in three blocks: in the control group each of these blocks comprised 6 trials each of O<sub>1</sub>->A->O<sub>1</sub>, O<sub>1</sub>->B->O<sub>1</sub>, O<sub>2</sub>->C->O<sub>2</sub> and O<sub>2</sub>->D->O<sub>2</sub>, and in the

experimental group 6 trials each of  $O_1 \rightarrow A \rightarrow O_2$ ,  $O_1 \rightarrow B \rightarrow O_2$ ,  $O_2 \rightarrow C \rightarrow O_1$  and  $O_2 \rightarrow D \rightarrow O_1$ .

*PIT test.*

*Assessment questionnaire.*

*Data analysis.* The data of all the phases were analysed using mixed ANOVAs. For the calculation of the PIT scores, the forward relationship between the CS and the outcome was considered. For example in the case of A in the experimental group ( $O_1 \rightarrow A \rightarrow O_2$ ), the congruent response was  $R_2$  and the incongruent response was  $R_1$ . This was chosen because it has already been established here that the forward training results in CS+s that produce the specific PIT effect. If the backward pairings produce excitatory conditioning, then the specific PIT effect will be less in the experimental than in the control group. But if it produces inhibitory conditioning, the specific PIT effect in the experimental group will be larger than in the control group.

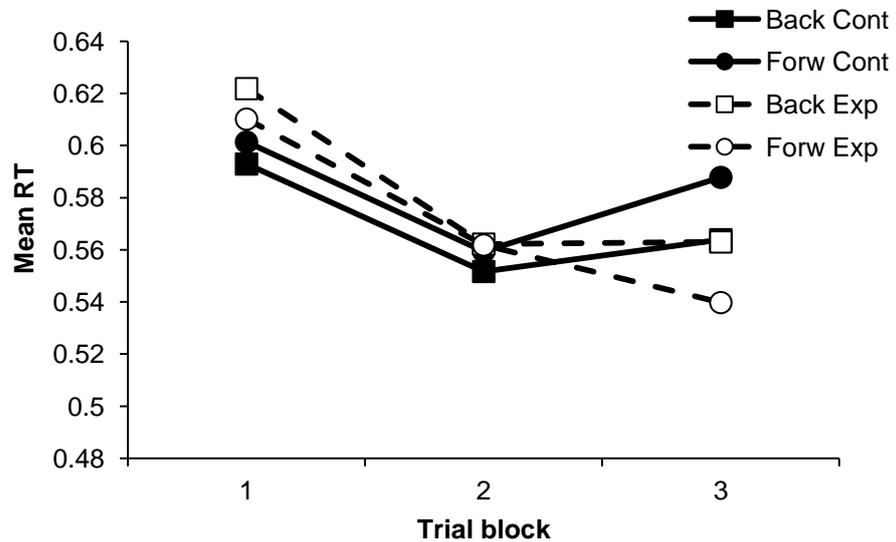
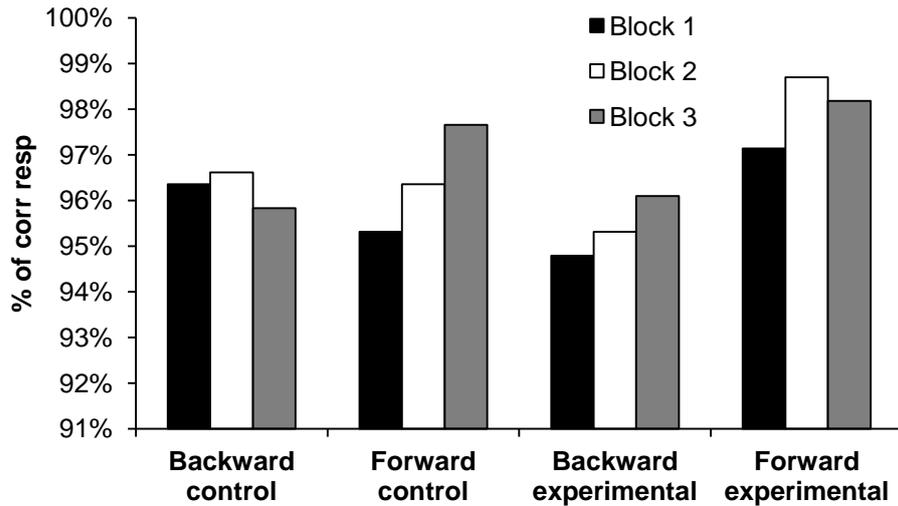
#### **4.13.2 Results**

*Instrumental training.* All participants completed this phase successfully (see Table 16) and no differences were found between the type of response, type of outcome or group,  $F_s < 1$ .

Table 16. Mean number of R1 and R2 responses and mean number of O1 and O2 deliveries for the experimental and control group in the instrumental phase of Experiment 14.

	R <sub>1</sub>	R <sub>2</sub>	O <sub>1</sub>	O <sub>2</sub>
<b>Experimental group</b>	251.8	246.8	50	50
<b>Control group</b>	254.4	251.3	50	50

*Pavlovian phase.* The mean percentages of correct responses were grouped into three blocks of six trials each, and they are presented in the top panel of Figure 37. It seems participants in both groups reached a higher level of accuracy for the forward trials but this was not confirmed by the statistical analysis. An ANOVA with block (1-3), trial type (forward, backward) and group showed no significant main effect or interaction, largest  $F(1, 30) = 3.91, p = .057, MSe = 23.68$  for the main effect of trial type. The mean reaction times were similarly grouped and presented in the bottom panel of Figure 37. As in the previous experiments, the graph suggests that participants' reaction times did not differ on the forward and backward trials. The corresponding ANOVA showed no significant main effect or interaction, largest  $F(1, 60) = 2.72, p = 0.074, MSe = .013$ .



*Figure 37.* Responses in the forward and backward trials of the Pavlovian phase in the experimental and control group of Experiment 14. Top panel: the mean percentage of correct responses. Bottom panel: the mean reaction times.

*PIT test.* The PIT scores for both groups are presented in Figure 38. The graph shows a clear specific PIT effect in both the experimental and the control group, i.e. more congruent than incongruent responses, even though the experimental group had inconsistent

backward and forward relationships. An ANOVA with block, congruence and group as factors revealed a significant main effect of congruence,  $F(1, 30) = 6.14, p = .019, MSe = 82.36, \eta_p^2 = .17$ , which did not interact with group,  $F < 1$ . Also, a significant Block x Group interaction was found,  $F(1, 30) = 4.85, p = 0.035, MSe = 11.09, \eta_p^2 = .14$ . However, simple main effects showed no significant effect of block in either of the groups, largest  $F(1, 30) = 3.77, p = 0.062$  for control group. Nothing else was significant,  $F_s < 1$ .

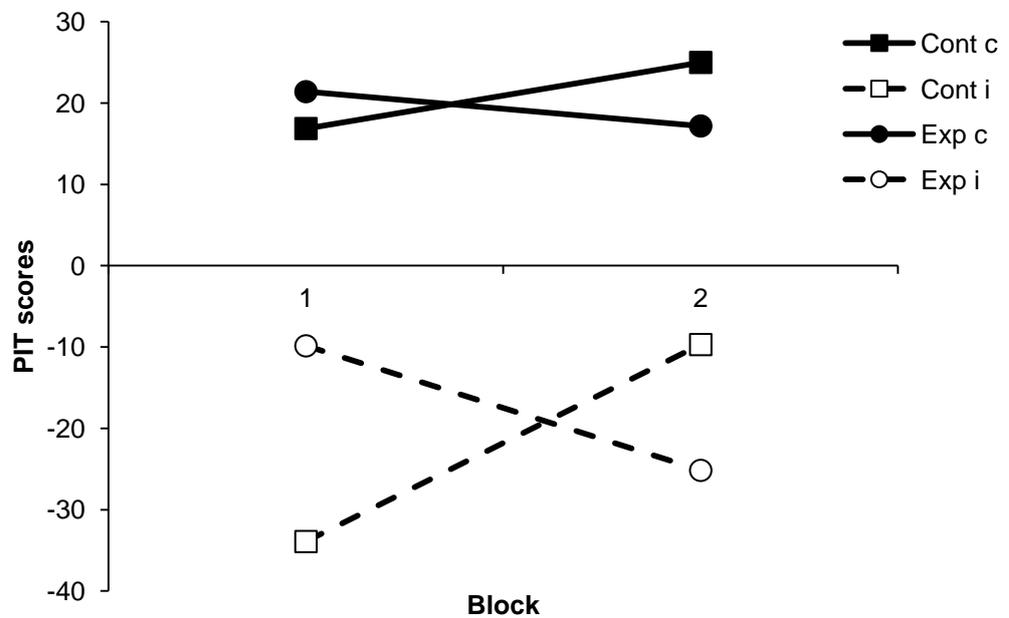


Figure 38. Group mean rates of congruent and incongruent responses for forward and backward trials in the PIT test of Experiment 14.

The analysis of the mean rates of responding during the preCS period (see Table 17) showed no significant main effect or interaction, largest  $F(1, 30) = 1.82, p = .19, MSe = 37.2$ .

Table 17. Group mean preCS response rates in each block of the PIT test of Experiment 14.

	Congruent		Incongruent	
	1	2	1	2
<b>Experimental group</b>	100.2	127.7	90	115.6
<b>Control group</b>	74.84	90.2	105.6	75.94

*Assessment questionnaire.* The percentages of responses are presented in Figure 39. It seems that the percentage of correct responses was higher than that of the incorrect responses, regardless of the type of trial, in both groups. An ANOVA with question (forward, backward, instrumental), response (correct, incorrect) and group as factors revealed a significant main effect of response,  $F(1, 30) = 28.08, p < .001, MSe = 1573, \eta_p^2 = .48$ , which did not interact with group factor,  $F < 1$ . Nothing else was significant, largest  $F(1, 30) = 2.82, p = .07, MSe = 389.3$ .

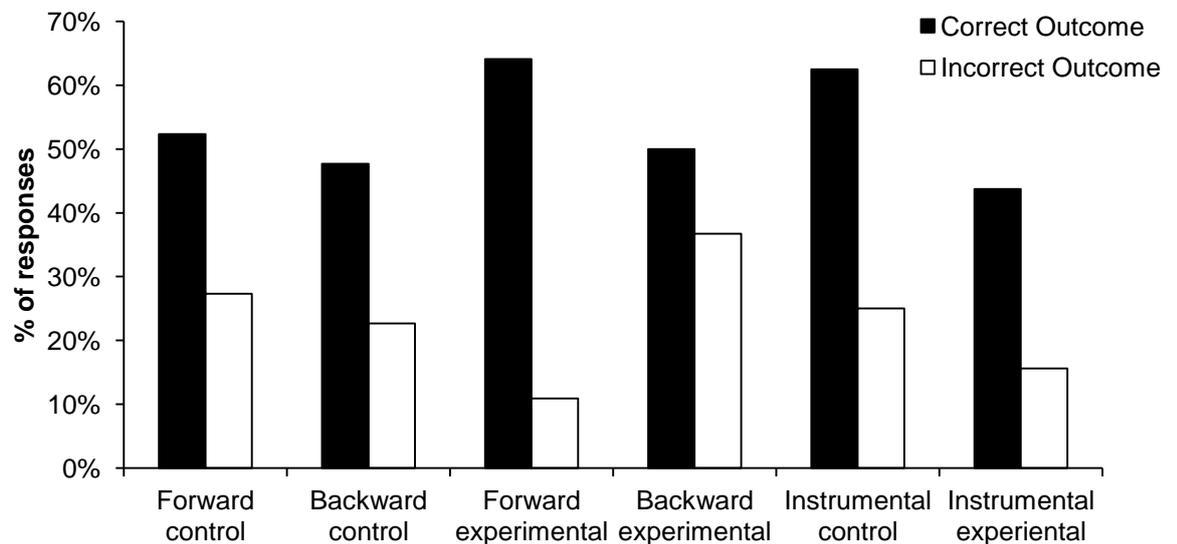


Figure 39. Mean percentage of correct and incorrect outcome responses in the assessment questionnaire of the experimental and control group of Experiment 14.

### 4.13.3 Discussion

The results of the PIT test indicate that the CSs in the experimental group produced the specific PIT effect, and that this effect was driven by the outcome that *followed* the CS during training (forward pairings). Furthermore, the size of this effect was the same as that found in the control group, in which the CSs were trained with the same outcome in backward and forward conditioning. It was hypothesized that if the backward associations became excitatory, the specific PIT effect would be reduced in the experimental group compared to the control. But if these associations were inhibitory, then the specific PIT effect should be larger than in the control group. However, neither of these predictions seems to be correct. The most obvious possibility is that even with this particular arrangement participants did not pay attention to the backward relationship between the CS and the outcome. In the experimental group participants could easily learn, for example, that if  $O_1$  was presented first, the next would be  $O_2$ . Then the information provided by the CS was irrelevant. However, there are two reasons to think that this explanation is incorrect. The first is that if participants ignored the CS, then the forward associations would also be impaired. Clearly this was not the case, otherwise no specific PIT effect would have been observed in the experimental group. The second reason is that the results of the assessment questionnaire indicate that participants learned both the

backward and the forward relationships between the CS and the outcomes.

The safest interpretation of these results is that when forward and backward associations are formed simultaneously, the forward association prevails in the PIT test. However, these results are not sufficient to draw a clear conclusion and further research is needed.

#### **4.14 General Discussion**

The main goal of this series of experiments was to assess if a CS trained in a backward relation with an outcome could produce the specific PIT effect, and in that case also to compare its effect with that produced by a CS trained in a forward relation with an outcome. However, the results of these experiments are inconclusive. In Experiment 9 participants watched the screen while CSs were immediately followed by the outcomes and different CSs were immediately preceded by these outcomes. The results of the PIT test showed that both types of CSs produced a specific PIT effect of the same magnitude. The possibility that in Experiment 9 participants' processing of the CS and the outcomes occurred simultaneously, regardless of whether the pairings were forward or backward, was assessed in Experiment 10 and 11. In these experiments an interval was included between the CS and the outcome presentations, but

again both types of CSs produced the same specific PIT effect, whether this interval was 1s (Experiment 10) or 2s (Experiment 11).

In Experiment 12 a different strategy was adopted, in that participants had to perform a response during the outcome presentation. This modification was meant to help participants to form backward and forward associations instead of processing the CSs and outcomes simultaneously. In the forward trials participants could use the information provided by the CS in order to anticipate the outcome, but in the backward trials the CS did not provide any information about the outcome presentations. The results of the PIT test were not significant, although they suggested that the specific PIT effect was diminished in the case of the CSs trained with backward pairings. Experiment 13 was a replication of Experiment 12, and in this experiment only the CSs trained with forward pairings produced the specific PIT effect. Furthermore, the CSs trained in a backward relationship with the outcomes seemed to produce the opposite effect, although it was not significant. In Experiment 14 the CSs were preceded and followed by either the same or a different outcome, and the contribution of both forward and backward associations was measured in the PIT test. The results suggest that the CSs produced the specific PIT effect based solely on the forward associations. These results suggest that the specific PIT effect is mainly governed by forward S-O associations, while backward associations seemed to have no impact on this effect.

In all of the experiments presented here, the learning of the CS-outcome relationships was assessed, either during conditioning or at the end of the task. In all of the experiments participants seemed to be able to correctly identify both of these relationships. The only exception was Experiment 12, and here no specific PIT effect was found either. However, the type of questions used in Experiments 9, 10 and 11 do not allow to distinguish if participants were also aware of the direction of the associations (forward or backward). Unfortunately it is precisely in these experiments that the possibility of simultaneous conditioning is present. Thus, if it is assumed that in those experiments participants processed both backward and forward associations as in simultaneous conditioning, regardless of the interval inserted between the CSs and the outcomes in Experiments 10 and 11, then it is not surprising that all the CSs produced the specific PIT effect. In the case of Experiment 13, the questionnaire conducted at the end of the task indicate that participants not only learned the CS-outcome relationships but also recognised the directionality of these associations. And it is in this experiment that the backward cues did not produce the specific PIT effect. If these results are taken together with those of Experiment 14, the safest conclusion is that, in this type of task, the backward S-O associations do not contribute to the specific PIT effect. If this is correct, then these results are not sufficient to conclude anything about the excitatory or inhibitory properties of a CS trained in backward conditioning.

## Chapter V

### General Discussion

#### 5.1 Summary of results

##### 5.1.1 Chapter II: Conditioned inhibition in the specific PIT effect

In Chapter II the effect of two CIs, each of them signalling the absence of a different outcome, was assessed on instrumental performance, either by presenting the CIs alone, or in compound with a CS+ that was trained with the same or a different outcome as the CI.

In three experiments, participants were trained to perform two responses concurrently, each of them reinforced with one outcome ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). In the Pavlovian phase, two CIs were established (X, Y), each of them signalling the absence of one of the outcomes ( $A \rightarrow O_1$ ;  $AX^-$ ;  $B \rightarrow O_2$ ;  $BY^-$ ). The inhibitory properties of X and Y were assessed in a summation test in which participants rated the likelihood of the outcome presentations in the presence of a CS+ in compound with either a CI or a control stimulus (pre-exposed or novel control cue). In the PIT test, participants could perform both instrumental responses in the presence and absence of the different Pavlovian stimuli. The main difference between these experiments was the manner in which the CIs were presented in the PIT test. In Experiment 1 CS+s, CIs and control cues were presented alone. In Experiment 2 the CS+s were presented in compound with either the CIs or the control stimuli. In Experiment 3 two PIT tests were conducted, one

exactly the same as in Experiment 2, and in the other the CS+s and CIs were presented alone and together, in such a way that the outcome whose absence was predicted by the CI was different to that signalled by the CS+.

The results of these experiments suggest that a CI does not exert any detectable effect on instrumental performance when presented alone, but is capable of reducing the specific PIT effect produced by the CS+s when these stimuli are presented together in the PIT test. Furthermore, the CI can also reduce the PIT effect produced by a CS+ that was trained with a different outcome to the CI.

The results of the summation test partially confirmed the inhibitory properties of the CIs. In this test, the CIs reduced participants' expectancy of the outcomes produced by a CS+ more than did a novel control stimulus. However, the effect produced by the CIs on participants' expectation was not different to that produced by a pre-exposed stimulus, which was thought to be a more conservative control. The idea that the pre-exposed control stimuli also acquired inhibitory properties, through differential inhibition, was considered. This possibility was explored in Experiment 2, in which instrumental training was conducted before the Pavlovian phase, in order to decrease the excitatory strength of the context and thus reduce the possibility of differential inhibition. However, the effect of the CIs and the pre-exposed stimuli was still not different in the summation test. This and the fact that the CIs but not the pre-exposed control cues reduced the specific PIT effect, indicate that differential inhibition did

not occur. One possibility is that the CI did not acquire enough inhibitory strength to produce a different effect to the pre-exposed stimuli in the summation test, presumably due to a low sensitivity of this test.

An additional finding was that the CIs reduced participants' expectancies of the outcomes in the summation test, relative to a novel stimulus, even when they were presented with a CS+ that was trained with a different outcome to the CI. This was consistent with the results of the PIT test in Experiment 3, in which the specific PIT effect was also reduced by the presentation of a CI trained with a different outcome to the CS+. These results raised the possibility that in these experiments conditioned inhibition was not outcome-specific. However, as was mentioned before, the fact that the CIs were not rated differently than the pre-exposed control stimuli raised the possibility that the summation test was not sensitive enough to detect small differences on participants' expectations. If this is correct then it is not possible to conclude with absolute certainty that conditioned inhibition was not outcome-specific in these experiments.

### **5.1.2 Chapter III: Outcome-specificity of conditioned inhibition in the specific PIT effect**

In Chapter III the PIT task was simplified by training only one CI, e.g. A->O<sub>1</sub>; AX-. This modification was made to reduce any

ambiguity about which outcome's absence was being signalled by the CI. In Experiment 4 the effect of a CI on instrumental performance was assessed by presenting it with an excitatory CS that was either trained with the same outcome as the CI, or with a different outcome. Experiment 5 was identical except that it included single presentations of the CI during training to increase its inhibitory properties. In Experiment 6 the instrumental conditioning phase was conducted before the Pavlovian phase. This was thought to increase the possibility of S-R associations being formed.

The results of these experiments showed that the CS+s produced the specific PIT effect when they were presented in compound with the control stimulus, but this effect was reduced when they were presented with the CI, although this reduction was transient in some of the experiments. Moreover, the CI reduced the specific PIT to a similar degree when it was presented with a CS+ trained with the same outcome (FX) as when it was trained with a different one (GX). These results are consistent with those of the summation test: in Experiments 4, 5 and 6 the participants' expectation of the outcomes was reduced by presenting the CI, relative to presenting the control stimulus, regardless of whether the CS+ was trained with the same or a different outcome to the CI.

Experiments 7 and 8 aimed to assess alternative explanations of the effect of X in the PIT test. In Experiment 7 the idea that X reduced the specific PIT effect for both CS+s due to its ability to elicit competing responses was assessed. In this experiment the

instrumental responses were trained in the absence of the outcomes, and the effect of the CI and the control stimulus was directly assessed in the PIT test. If the CI did not suppress the activation of the outcome representation but instead reduced performance by eliciting competing responses, then it should have reduced the specific PIT effect even if the responses were not reinforced with valuable outcome. However, the results showed that the CIs did not reduce performance more than a pre-exposed control stimulus. Critically, the results of the summation test confirmed the inhibitory properties of X. This evidence rules out the possibility that the CI reduced responding because elicited competing responses. Experiment 8 explored the possibility that during training X might have become associated with a different neutral outcome and that in the PIT test activation of this outcome representation interfered with the retrieval of the outcome representation associated with the instrumental responses, affecting performance. In this experiment X was explicitly paired with a different outcome while C received the same training as in the previous experiments. In the PIT test the CS+s were presented with X and C and the specific PIT effect was measured. The results showed no difference between the effect of X and C on the PIT effect produced by the CS+s, confirming that it was not the ability of X to elicit an alternative outcome representation that was responsible for the results found in Experiment 4, 5 and 6.

### 5.1.3 Chapter IV: Backward conditioning in the specific PIT effect

Chapter IV explored a different method of producing inhibitory CSs, by training these CSs in a backward relation with the outcome presentations, e.g.  $O_1 \rightarrow A$ . It also aimed to provide further evidence relating to the conflicting literature on backward conditioning (e.g. Mahoney & Ayres, 1976; Siegel & Domjan, 1971, 1974) and the effect of CSs trained in this procedure on the specific PIT effect (Cohen-Hatton et al., 2013; Delamater et al., 2003; Laurent et al., 2014). In Experiment 9 two responses were trained, as in the previous experiments ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). In the Pavlovian phase participants watched the screen while two CSs were immediately followed by one of the outcomes ( $A \rightarrow O_1$ ;  $B \rightarrow O_2$ ), and two additional CSs were immediately preceded by one of the outcomes ( $O_1 \rightarrow C$ ;  $O_2 \rightarrow D$ ). In Experiment 10 a 1-second interval was inserted between the CS and the outcome presentations in both types of trials, while in Experiment 11 the interval was increased to 2 seconds. In Experiment 9, 10 and 11 each of the CSs was presented in the PIT test while participants performed both instrumental responses. The results of these experiments showed that both types of CSs, i.e. those trained in a forward or a backward relation to the outcomes, produced a comparable specific PIT effect.

In Experiment 12 and 13 a different strategy was adopted to increase the difference between forward and backward pairings. In the Pavlovian phase participants had to perform a response each time

they saw one of the outcomes and a different response when they saw the other outcome. The rest of the task was the same as in Experiments 9, 10 and 11. The results of Experiment 12 did not reach significance, but the results of Experiment 13 showed that only the CSs trained in a forward relation with the outcomes produced the specific PIT effect, while the CSs trained in a backward manner produced a numerically opposite, although not statistically significant, effect. Experiment 14 aimed to compare the effect of backward and forward associations in the specific PIT effect directly. In this experiment each of the CSs was preceded *and* followed by an outcome. For one group of participants both outcomes were the same, e.g.  $O_1 \rightarrow A \rightarrow O_1$ , but for the other group the outcomes were different, e.g.  $O_1 \rightarrow A \rightarrow O_2$ . The results of the PIT test showed that in both groups the CSs increased the performance of the response that was trained with the outcome that followed, not preceded, the CS, suggesting that the specific PIT effect is mainly determined by the forward associations.

## **5.2 Theoretical implications**

The idea that CIs produce the opposite effect to CS+s because the CIs suppress a representation of the US (Rescorla & Holland, 1977) was thought to serve as a tool to discriminate between the different accounts of PIT. These accounts were divided into two

groups: those that explain the specific PIT effect being mediated by an outcome representation, i.e. S-O-R accounts (e.g. Trapold & Overmier, 1972; Balleine & Ostlund, 2007), and those that do not, e.g. S-R account (Cohen-Hatton et al., 2013). According to the S-O-R accounts, in a PIT test a CS+ activates an outcome representation that elevates performance of those responses reinforced with the same outcome as the CS+, either through a bidirectional R-O association (Asratyan, 1974; Pavlov, 1932; Rescorla, 1994b) or by a backward O-R association (e.g. Balleine & Ostlund, 2007; Ostlund & Balleine, 2007). If the activation of the outcome representation is suppressed by presenting a CI then the specific PIT effect should be reduced. In contrast, the S-R account states that a direct link between the CS and the instrumental responses is formed during training, and that at test the CS elicits these responses without the mediation of an outcome representation. For this reason, a CI that suppresses the activation of an outcome representation should not affect the ability of the CS to produce responding. However, the results of experiments reported in Chapters 2 and 3, in which a CI reduced the specific PIT effect produced by a CS+, are consistent with the idea that an outcome representation mediates the specific PIT effect, as the S-O-R accounts propose.

Nevertheless, the results of the summation and PIT tests in Chapter 2 and 3 also suggest that the effect of the CIs was not outcome-specific. More specifically, in Chapter 3 a CI that signalled the absence of a particular outcome, e.g. X-no O<sub>1</sub>, reduced

participants' expectations of a different outcome, e.g.  $O_2$ , in the summation tests and also reduced the specific PIT effect produced by a CS+ trained with a different outcome, e.g.  $G \rightarrow O_2$ . Although there is conflicting evidence about the outcome-specificity of conditioned inhibition (e.g. LoLordo, 1967), there is evidence suggesting that a CI can reduce responding to a CS+ even if this CS+ is trained with a different outcome (e.g. Nieto, 1984; Pearce, Montgomery and Dickinson, 1981), which is consistent with the results reported here. The main problem with these results is that the mechanism through which the CIs produce this non-specific effect is not clear. One of the interpretations is that the CIs suppressed the activation of the common elements of both outcomes, thus reducing participants' expectations in the summation test. For instance, it is possible that the motivational value of both outcomes was similar, as both were visual representations of appetitive events, i.e. food and drinks. If a CI suppresses the common elements of  $O_1$  and  $O_2$  then it will reduce participants' expectations of both outcomes, which is what was found in the summation tests.

If this is the correct interpretation of the results of the summation tests, then it is hard to understand how the CIs also reduced the specific PIT effect in a non-specific manner, i.e. the CIs reduced the specific PIT effect produced by a CS+ trained with a different outcome. According to the S-O-R accounts, the specific PIT effect must be mediated by the activation of a representation that encodes specific information about the outcome. In other words, a CS

trained to predict  $O_1$  will activate the elements of  $O_1$  that are different to  $O_2$ . Only in this way can this representation elevate performance of the response also trained with  $O_1$ . If the CS+ activates the common elements of the outcomes then it should also elevate performance of a response trained with  $O_2$ , i.e. general PIT.

Different solutions to this problem were proposed in Chapter 3. For instance, one possibility is based on the idea that a CI can elicit a motivational state opposite to that produced by the outcome presentations, which interferes with instrumental performance (Dearing & Dickinson, 1979). According to this, a CS paired with an outcome will elicit a motivational state similar to that produced by the outcome. Also, outcomes from the same motivational modality, e.g. appetitive, will elicit a similar motivational state. However, because during a conditioned inhibition procedure the CI is paired with the absence of an expected outcome, an association is formed between the CI and a motivational state antagonist to that produced by the outcomes. This motivational state would be similar to that produced by an outcome from the opposite motivational valence, e.g. aversive. Then when a CS+ and a CI are presented together at test, the motivational state elicited by the CI counteracts that elicited by the CS+, reducing the specific PIT effect. Critically, the CI will reduce the specific PIT effect produce by CSs that signal different outcomes, as long as these outcomes have the same motivational valence. For instance, in the experiments reported in Chapter 2 and 3 the pictures of foods and drinks were appetitive outcomes that elicit a similar

motivational state. Accordingly, the CI produce a motivational state opposite to that elicited by both outcomes, affecting behaviour in a non-specific manner. This is consistent with the results of the summation tests, in which the CIs reduced participants' expectations to the outcomes produced by different CS+s, even when the CIs signalled the absence of a different outcome than that predicted by the CS+s. Similarly, the specific PIT effect produced by the CS+s at test was reduced by the presence of the CI, regardless of which outcome they signalled, because all the CS+s elicited a similar motivational state that was counteracted by the CI.

If this interpretation is correct, these results do not provide evidence to discriminate between the S-O-R and the S-R accounts. If the motivational state elicited by the CIs was incompatible with that required for the instrumental responses to be performed, then the specific PIT effect should be reduced regardless of whether it is produced by an outcome representation or directly by the CS. However, if this interpretation is correct then single presentations of the CIs at test should reduce instrumental performance more than a neutral control stimulus. This is because while the neutral cue does not elicit any particular motivational state, the CI elicits a state that is incompatible with instrumental performance. This is not consistent with the results of Experiment 1 reported in Chapter 2, in which single presentations of the CIs did not reduce instrumental responding more than a pre-exposed control stimulus.

Another possible explanation, based on the S-O-R accounts, is that although the sensory outcome representation produces the specific PIT effect, the motivational aspects of the outcome also serves to support instrumental performance. That is to say, a CS that only activates a sensory outcome representation will produce a smaller specific PIT effect than will a CS that activates both sensory and motivational outcome representations. This solution is consistent with the idea that the CIs suppressed the motivational components of the outcome representations common to  $O_1$  and  $O_2$ , and also with the idea that instrumental associations (R-O) encode both the sensory and motivational properties of the outcomes. Nevertheless, this suggestion seems to be in contradiction with most of the evidence from outcome devaluation procedures. Experiments that have devalued the outcome before the PIT test, either by pairing the outcome with an aversive state or through satiety, have found that this procedure does not affect the size of the specific PIT effect (e.g. Corbit, Janak & Balleine, 2007; Holland, 2004). However, it is possible that the effect of the CIs is different to that obtained via outcome devaluation. Because outcome devaluation selectively reduces responding in an extinction test, it is conceptualised as being sensory-specific. In contrast, the CIs trained in the experiments reported here seem to affect behaviour in a general manner.

Another possibility is based on the idea that the CSs activate one outcome representation that encodes the sensory aspects and the motivational properties of the outcome. Although the specific PIT effect

is determined by the sensory elements of this representation, inhibition of the motivational elements should slow down the activation of the outcome representation as a whole. As it is likely that the outcomes used in these experiments had the same or similar motivational value, the CIs should be capable of slowing down activation of both outcome representations. For instance, a CS<sub>1</sub> paired with O<sub>1</sub> should activate an outcome representation that encodes specific information about O<sub>1</sub> but also the motivational value of that outcome that is common with O<sub>2</sub>. Similarly, a CS<sub>2</sub> that signals O<sub>2</sub> should activate a sensory outcome representation specific to O<sub>2</sub> but that also encodes the motivational properties common to O<sub>1</sub>. Then, according to this explanation, if CS<sub>1</sub> is presented at test with a CI that signalled the absence of O<sub>1</sub> the activation of the sensory aspects of the outcome will be delayed by the suppression of the motivational elements of this representation, reducing the size of the specific PIT effect. Importantly, if CS<sub>2</sub> is presented with the same CI the activation of the O<sub>2</sub> representation will also be delayed because of the suppression of the motivational aspects of the outcome by the CI, which are common to O<sub>1</sub>. This would explain how in the experiments reported here the CIs reduced specific PIT even when the summation tests revealed that their effect on behaviour was not outcome-specific. This interpretation is not contrary to the results of outcome devaluation procedures. Pairing the outcome with an aversive consequence or shifting the motivational state of the subjects does reduce the motivational value of the outcomes. However, outcome devaluation should not affect the ability

of the CS to activate both the sensory and motivational aspects of the outcome representation. Even if a CS activates a representation of a no longer desirable outcome, the activation of the sensory elements of this representations should not be slowed down.

In Chapter 4 some seemingly contradictory results were reported. In Experiments 9, 10 and 11, CSs that were preceded by the outcomes during training produced a comparable specific PIT effect to those trained in a forward conditioning procedure. But in Experiment 13 only the CSs trained in forward conditioning produced the specific PIT effect. However, there were important differences in the procedures used in these experiments. In Experiments 9, 10 and 11 participants had to passively watch the screen while the CSs and outcomes were presented in the screen, while in Experiment 13 participants had to try to predict the outcome presentations and to perform a response depending on the type of the outcome. In this sense it is possible that in Experiments 9, 10 and 11 participants processed the CS and outcomes simultaneously instead of paying attention to the direction of the associations, resulting in the formation of an excitatory association. Although this possibility was explored by increasing the interval between the CS and the outcomes, producing the same results, it is possible that either the interval was not long enough to produce inhibitory conditioning, or that participants had time to rehearse each of the CS-US associations during the trials, treating the CS-US pairings as simultaneous rather than backward pairings (Arcediano, Escobar & Miller, 2003; Jants & Underwood, 1958).

Regardless of the mechanism by which the CS trained in a backward relation with the outcomes acquired excitatory properties, these results can be predicted by both S-O-R and S-R accounts. In both cases, i.e. if participants processed the stimuli as in a backward or in a simultaneous relation, the S-O-R accounts predict that the CS will activate an outcome representation, producing the specific PIT effect. However, it might be argued that the backward cues should have produced a smaller PIT effect than the forward cues because backward conditioning results in weaker associations than forward. Nevertheless, even if it is correct that both types of training produce associations of different strength, there is evidence suggesting that CSs with different associative strength produce a comparable specific PIT effect (e.g., Delamater & Oakeshott, 2007). Moreover, it has been suggested that after very limited training a CS acquires the ability to evoke the outcome representation, and that simple contiguity between the CS and the outcome is enough to endows the CSs with this ability (Barnet & Miller, 1996; Cole & Miller, 1999; Matzel, Held & Miller, 1988).

In the case of the S-R account, this states that a CS trained in a backward relation with the US should produce a larger specific PIT effect than one trained in a forward relation (Cohen-Hatton et al., 2013). For instance, if two responses are trained, (e.g.  $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ) then each time a CS is followed by an outcome in the following Pavlovian phase, e.g.  $CS_1 \rightarrow O_1$ , the outcome will activate a representation of the response trained in the previous phase, e.g.  $O_1$ -

$>R_1$ , allowing the formation of an S-R association, i.e.  $CS_1 \rightarrow R_1$ . A similar process occurs if the CS is preceded by the outcome, e.g.  $O_2 \rightarrow CS_2$ ; each time  $O_2$  is presented it activates a representation of  $R_2$ , i.e.  $O_2 \rightarrow R_2$ , which results in a  $CS_2 \rightarrow R_2$  association. However, the critical difference between the forward and backward trials is the temporal distance between the CS and the response representation evoked by the outcome delivery. While in the forward pairings of the CS and US this representation is activated after the CS is presented ( $CS_1 \rightarrow O_1 \rightarrow R_1$ ), in the backward pairings of the CS and US this representation should be active closer in time to the CS presentation ( $O_2 \rightarrow CS_2/R_2$ ), resulting in a stronger S-R association in the backward pairings compared to the forward. Thus, if these experiments produced excitatory backward associations the S-R account cannot easily explain the results of the PIT tests. However, if participants processed the pairings of the CSs and the outcomes as simultaneous trials rather than backward, then both forward and backward pairings of CS and US should have produced similar S-R associations. If this is correct, then the S-R account, as the S-O-R accounts, predicts that both types of CSs should produce a comparable specific PIT effect, which is consistent with the results reported.

However, the results of Experiment 13, in which participants had an active role in trying to predict the outcome deliveries, showed that the CSs that were trained in a backward relation with the outcomes did not produce the specific PIT effect, while the CSs trained in forward conditioning did. It was thought that the procedure

used in this experiment increased participants' attention to the directionality of the Pavlovian associations, and the results can be explained by the S-O-R accounts in two possible ways. One possibility is that the backward trials produced an inhibitory association between the CS and the outcome, so that at test the CS suppressed rather than activated the outcome representation. However, this idea is not entirely consistent with the animal literature showing that inhibitory backward associations produce a reverse specific PIT effect. It has been reported that a CS trained in a backward relation with the outcome increase the response that has been trained with an outcome different than that paired with the CS, which is a reverse specific PIT effect, using rats (Delamater, 2003) and mice (Laurent et. al., 2015) as subjects. A second possibility is that backward associations were formed, but these association do not contribute to the specific PIT effect. In other words, a CS cannot retrieve a representation of the outcome that is encoded in a backward association. The results of Experiment 14, in which CSs that were preceded by one outcome but followed by a different one, e.g.  $O_2 \rightarrow CS_1 \rightarrow O_1$ , support this latter suggestion. Here the CSs produced the specific PIT effect according to the forward, not the backward, relationship, i.e.  $CS_1: R_1 > R_2$  ( $R_1 \rightarrow O_1$ ;  $R_2 \rightarrow O_2$ ). If the backward pairings result in the formation of an excitatory association with the outcomes, i.e. excitatory backward associations, then at test these CSs should have activated a representation of the outcome that preceded but also a representation of the outcome that followed them during training, cancelling the

specific PIT effect. In contrast, if backward pairings produce an inhibitory backward associations, then at test the CSs should have suppressed the representation of the outcome that preceded them while activating a representation of the outcome that followed them in training, resulting in a larger specific PIT effect. However, the specific PIT effect produced by these CSs was comparable to that produced by the CSs in a control group, in which the CSs were preceded and followed by the same outcome, e.g.  $O_1 \rightarrow CS_1 \rightarrow O_1$ .

A final note regards the instrumental associations that are part of the specific PIT effect. Some of the versions of the S-O-R account suggest that the specific PIT effect is produced by O-R associations (e.g. Trapold & Overmier, 1972; Balleine & Ostlund, 2007). According to these versions, during instrumental training the outcome that precedes a response forms an O->R association and in the PIT test the CS elevates responding through activation of this outcome representation. Nevertheless, the fact that in most of the experiments presented here (Chapters 2 and 3) the instrumental responses were trained concurrently in the same session, contradicts that idea. This is because each of the responses was reinforced with one outcome ( $R_1 \rightarrow O_1$  and  $R_2 \rightarrow O_2$ ), but they could have been preceded by any of the two outcomes ( $O_1 \rightarrow R_1$ ,  $O_1 \rightarrow R_2$ ,  $O_2 \rightarrow R_1$ ,  $O_2 \rightarrow R_2$ ), resulting in multiple O-R association. Thus, if the specific PIT effect found in these experiments was caused by an S-O-R mechanism, the most likely explanation is the outcome representation elevated responding via a bidirectional R-O association (e.g. Rescorla, 1994b).

### 5.3 Limitations and future research

Most of the discussion presented here is based on the concepts proposed by Konorski (1948; 1967), according to which a CS representation can form associations with different US representations, i.e. sensory and motivational properties of the US. However, in the experiments presented here the outcomes were pictures of food and drinks instead of real outcomes. Although it is clear that the food and drink pictures had unique elements (otherwise no specific PIT effect should have been found), it can be argued that these symbolic outcomes had a low motivational value that were common to both types of images, which might reduce the validity of the interpretation described here. In this sense future research could be conducted by using outcomes with a higher motivational value that allow us to directly compare between the different explanations of specific PIT. For instance, one of the possible explanations of the results presented here is that a CI acts by suppressing the activation of an outcome representation that encodes the common elements of the outcomes, e.g. motivational representation. One possible way to assess that interpretation is to use outcomes with real motivational value. In this sense, outcomes from different motivational valence could be used, e.g. appetitive and aversive. Some researchers have already used real outcomes to study PIT in humans (e.g. Colagiuri & Lovibond, 2015, Watson et al. 2014), but to my knowledge none of them have addressed conditioned inhibition nor backward conditioning

as a means of analysing the specific PIT effect in the way employed here.

Another issue concerns the measurement of Pavlovian conditioning with human participants. Different strategies were used in these experiments, such as rating scales, asking participants to predict the outcomes trial by trial, etc., but it is possible that none of them really captured the strength of the Pavlovian associations. This is even more important in the experiments that used backward and forward conditioning. Although the results of the Pavlovian tests in the experiments reported in Chapter IV suggest that participants had some knowledge of the S-O associations, it was not possible to determine if there were differences between the associative strength of the cues trained in backward and forward relation with the outcome. Thus one of the challenges for future research is to develop a better assessment of Pavlovian conditioning for human studies.

The results presented in this thesis are not entirely consistent with some of the literature on PIT. For instance, it has been found that in mice conditioned inhibition results in a CI that when presented in a PIT test elevates performance of a response that has been trained with a different outcome, i.e. reverse specific PIT effect, suggesting that inhibition is outcome specific (Laurent et al., 2015). In contrast, the experiments presented here indicate that conditioned inhibition is not outcome specific. This was found in the summation tests conducted in each of the experiments using conditioned inhibition. Also, it was found that a CI not only reduced the specific PIT effect

produced by a CS+ trained with the same outcome, but also that produced by a CS+ predictor of a different outcome, suggesting that a CI also affects specific PIT effect in a general manner. Furthermore, the CI did not produce a reverse specific PIT effect in any of the experiments presented in this thesis, it only reduced the size of the specific PIT effect produced by a CS+ compared to a neutral control stimulus. Although it is expected that the results might differ due to different procedures and species, the contrast between these results require further research. If conditioned inhibition does diminish the specific PIT effect in a general manner, as the experiments presented here suggest, then this could have important repercussions in applied psychology. For instance, one of the problems in addiction from a clinical perspective is that drug-related CSs can increase the level of drug consumption. The fact that the specific PIT effect seems to be resistant to outcome devaluation and extinction is a challenge for therapies focused on addiction and rehabilitation. Thus if a CI can reduce the specific PIT effect produced by drug-related stimuli it could contribute to the efficacy of this type of therapy. Furthermore, if the effect of these CIs is not specific to the outcome of training, then it would be possible to train a stimulus to signal the absence of a consequence that is not necessarily the target drug, and then use this stimulus to reduce drug consumption. For these reasons further research is necessary to assess if the results reported here can be generalised or not.

The results found using backward and forward pairings are not consistent with the literature either. While some researchers have found that backward training produces excitatory conditioning (e.g. Ayres, Haddad & Albert, 1987; Burkhardt, 1980), other have found that it produces inhibitory conditioning (e.g. Maier, Rapaport & Wheatley 1976; Moscovitch & LoLordo, 1968). In some studies using a PIT task it has been found that a CS trained in a backward relation with the outcomes produces the opposite of specific PIT (Delamater et al., 2003; Laurent et al., 2014), but in others it has been found to increase the size of the specific PIT effect (Cohen-Hatton et al., 2013). In contrast, the experiments reported here suggest that backward associations do not contribute to the specific PIT effect. However, unlike the experiments reported in this thesis, all the experiments described have used animals as subjects and, to my knowledge, there are no other experiments that have assessed backward conditioning using a PIT task in humans. In this sense replications are needed in order to validate the results reported here.

#### **5.4 Conclusion**

This thesis used a conditioned inhibition procedure and backward conditioning in a PIT task with human participants. The aim of these experiments was to improve understanding of the mechanism that underlies the specific PIT effect in humans. This phenomenon is

important from both a theoretical and an applied perspective. Understanding the processes involved in the specific PIT effect might help us to improve our knowledge about the interactions between Pavlovian and instrumental conditioning, and it might also help us to treat disorders that can be explained, at least partially, by the specific PIT phenomenon, such as eating disorders and addictions. Overall the findings suggest that the S-O-R accounts might be capable of explaining the specific PIT effect but not without further assumptions. However, although the results reported here are not entirely consistent with some of the animal literature, they provide novel evidence in humans and they hopefully will serve to future studies on the specific PIT effect.

## References

- Adams, C. D. (1980). Post-conditioning devaluation of an instrumental reinforcer has no effect on extinction performance. *The Quarterly Journal of Experimental Psychology*, 32(3), 447-458.
- Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly journal of experimental psychology*, 33(2), 109-121.
- Allman, M. J., DeLeon, I. G., Cataldo, M. F., Holland, P. C., & Johnson, A. W. (2010). Learning processes affecting human decision making: An assessment of reinforcer-selective Pavlovian-to-instrumental transfer following reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, 36(3), 402-408.
- Asratyan, E. A. (1974). Conditional reflex theory and motivational behavior. *Acta Neurobiologiae Experimentalis*, 34(1), 15-31.
- Ayres, J. J., Haddad, C., & Albert, M. (1987). One-trial excitatory backward conditioning as assessed by conditioned suppression of licking in rats: Concurrent observations of lick suppression and defensive behaviors. *Animal Learning & Behavior*, 15(2), 212-217.
- Baker, A. G. (1977). Conditioned inhibition arising from a between-sessions negative correlation. *Journal of Experimental Psychology: Animal Behavior Processes*, 3(2), 144-155.

- Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4), 407-419.
- Balleine, B. W., & Ostlund, S. B. (2007). Still at the Choice-Point. *Annals of the New York Academy of Sciences*, 1104(1), 147-171.
- Baxter, D. J., & Zamble, E. (1982). Reinforcer and response specificity in appetitive transfer of control. *Animal Learning & Behavior*, 10(2), 201-210.
- Blundell, P., Hall, G., & Killcross, S. (2001). Lesions of the basolateral amygdala disrupt selective aspects of reinforcer representation in rats. *The Journal of Neuroscience*, 21(22), 9018-9026.
- Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, 114(1), 80-89.
- Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & memory*, 11(5), 485-494.
- Bouton, M. E., & Bolles, R. C. (1979). Role of conditioned contextual stimuli in reinstatement of extinguished fear. *Journal of Experimental Psychology: Animal Behavior Processes*, 5(4), 368-378.

- Burkhardt, P. E. (1980). One-trial backward fear conditioning in rats as a function of US intensity. *Bulletin of the Psychonomic Society*, 15(1), 9-11.
- Cohen-Hatton, S. R., Haddon, J. E., George, D. N., & Honey, R. C. (2013). Pavlovian-to-instrumental transfer: Paradoxical effects of the Pavlovian relationship explained. *Journal of Experimental Psychology: Animal Behavior Processes*, 39(1), 14-23.
- Colagiuri, B., & Lovibond, P. F. (2015). How food cues can enhance and inhibit motivation to obtain and consume food. *Appetite*, 84, 79-87.
- Cole, R. P., & Miller, R. R. (1999). Conditioned excitation and conditioned inhibition acquired through backward conditioning. *Learning and Motivation*, 30(2), 129-156.
- Colwill, R. M., & Motzkin, D. K. (1994). Encoding of the unconditioned stimulus in Pavlovian conditioning. *Animal Learning & Behavior*, 22(4), 384-394.
- Colwill, R. M., & Rescorla, R. A. (1985). Postconditioning devaluation of a reinforcer affects instrumental responding. *Journal of Experimental Psychology: Animal Behavior Processes*, 11(1), 120-132.
- Colwill, R. M., & Rescorla, R. A. (1986). Associative structures in instrumental learning. *The psychology of learning and motivation*, 20, 55-104.

- Colwill, R. M., & Rescorla, R. A. (1988). Associations between the discriminative stimulus and the reinforcer in instrumental learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 14(2), 155-164.
- Colwill, R. M., & Rescorla, R. A. (1990). Effect of reinforcer devaluation on discriminative control of instrumental behavior. *Journal of Experimental Psychology: Animal Behavior Processes*, 16(1), 40-47.
- Corbit, L. H., & Balleine, B. W. (2005). Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer. *The Journal of Neuroscience*, 25(4), 962-970.
- Corbit, L. H., & Balleine, B. W. (2011). The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. *The Journal of Neuroscience*, 31(33), 11786-11794.
- Corbit, L. H., & Janak, P. H. (2007). Ethanol-Associated Cues Produce General Pavlovian-Instrumental Transfer. *Alcoholism: Clinical and Experimental Research*, 31(5), 766-774.
- Corbit, L. H., Janak, P. H., & Balleine, B. W. (2007). General and outcome-specific forms of Pavlovian-instrumental transfer: the effect of shifts in motivational state and inactivation of the

ventral tegmental area. *European Journal of Neuroscience*, 26(11), 3141-3149.

Corbit, L. H., Muir, J. L., & Balleine, B. W. (2001). The role of the nucleus accumbens in instrumental conditioning: evidence of a functional dissociation between accumbens core and shell. *The Journal of Neuroscience*, 21(9), 3251-3260.

Cunningham, C. L., Fitzgerald, R. D., & Francisco, D. L. (1977). Excitatory and inhibitory consequences of explicitly unpaired and truly random conditioning procedures on heart rate in rats. *Animal Learning & Behavior*, 5(2), 135-142.

De Borchgrave, R., Rawlins, J. N. P., Dickinson, A., & Balleine, B. W. (2002). Effects of cytotoxic nucleus accumbens lesions on instrumental conditioning in rats. *Experimental Brain Research*, 144(1), 50-68.

Dearing, M. F., & Dickinson, A. (1979). Counterconditioning of shock by a water reinforcer in rabbits. *Animal Learning & Behavior*, 7(3), 360-366.

Delamater, A. R. (1995). Outcome-selective effects of intertrial reinforcement in a Pavlovian appetitive conditioning paradigm with rats. *Animal Learning & Behavior*, 23(1), 31-39.

Delamater, A. R. (1996). Effects of several extinction treatments upon the integrity of Pavlovian stimulus-outcome associations. *Animal Learning & Behavior*, 24(4), 437-449.

- Delamater, A. R. (2004). Experimental extinction in Pavlovian conditioning: behavioural and neuroscience perspectives. *Quarterly Journal of Experimental Psychology Section B*, 57(2), 97-132.
- Delamater, A. R. (2007). The Role of the Orbitofrontal Cortex in Sensory-Specific Encoding of Associations in Pavlovian and Instrumental Conditioning. *Annals of the New York Academy of Sciences*, 1121(1), 152-173.
- Delamater, A. R., & Oakeshott, S. (2007). Learning about multiple attributes of reward in Pavlovian conditioning. *Annals of the New York Academy of Sciences*, 1104(1), 1-20.
- Delamater, A. R., Sosa, W., & LoLordo, V. M. (2003). Outcome-specific conditioned inhibition in Pavlovian backward conditioning. *Animal Learning & Behavior*, 31(4), 393-402.
- Di Ciano, P., & Everitt, B. J. (2003). Differential control over drug-seeking behavior by drug-associated conditioned reinforcers and discriminative stimuli predictive of drug availability. *Behavioral neuroscience*, 117(5), 952.
- Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 308(1135), 67-78.
- Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, 22(1), 1-18.

- Dickinson, A., & Balleine, B. (2002). Steven's handbook of experimental psychology: learning, motivation and emotion. *The role of learning in the operation of motivational systems*, 3, 497-534.
- Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning & Behavior*, 23(2), 197-206.
- Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, 35(1), 35-51.
- Dickinson, A., Smith, J., & Mirenowicz, J. (2000). Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behavioral Neuroscience*, 114(3), 468.
- Estes, W. K. (1943). Discriminative conditioning. I. A discriminative property of conditioned anticipation. *Journal of Experimental Psychology*, 32(2), 150-155.
- Everitt, B. J., Dickinson, A., & Robbins, T. W. (2001). The neuropsychological basis of addictive behaviour. *Brain Research Reviews*, 36(2), 129-138.
- Ganesan, R., & Pearce, J. M. (1988). Effect of changing the unconditioned stimulus on appetitive blocking. *Journal of*

*Experimental Psychology: Animal Behavior Processes*, 14(3), 280-291.

Garbusow, M., Schad, D. J., Sommer, C., Jünger, E., Sebold, M., Friedel, E. & Rapp, M. A. (2014). Pavlovian-to-instrumental transfer in alcohol dependence: a pilot study. *Neuropsychobiology*, 70(2), 111-121.

Glasner, S. V., Overmier, J. B., & Balleine, B. W. (2005). The role of Pavlovian cues in alcohol seeking in dependent and nondependent rats. *Journal of studies on alcohol*, 66(1), 53-61.

He, Z., Cassaday, H. J., Park, S. B. G., & Bonardi, C. M. (2012). When to hold that thought: an experimental study showing reduced inhibition of pre-trained associations in schizophrenia. *PloS one*, 7(7), e42175.

Hearst, E., & Franklin, S. R. (1977). Positive and negative relations between a signal and food: Approach-withdrawal behavior to the signal. *Journal of Experimental Psychology: Animal Behavior Processes*, 3(1), 37-52.

Hall, J. F. (1984). Backward conditioning in Pavlovian type studies. *Pavlovian Journal of Biological Science* 19, 163-168.

Heth, C. D. (1976). Simultaneous and backward fear conditioning as a function of number of CS-UCS pairings. *Journal of Experimental Psychology: Animal Behavior Processes*, 2(2), 117-129.

- Hoffman, J. W., & Fitzgerald, R. D. (1982). Bidirectional heart rate responses in rats associated with excitatory and inhibitory stimuli. *Animal Learning & Behavior*, 10(1), 77-82.
- Hogarth, L. (2012). Goal-directed and transfer-cue-elicited drug-seeking are dissociated by pharmacotherapy: evidence for independent additive controllers. *Journal of Experimental Psychology: Animal Behavior Processes*, 38(3), 266-278.
- Hogarth, L., & Chase, H. W. (2011). Parallel goal-directed and habitual control of human drug-seeking: implications for dependence vulnerability. *Journal of Experimental Psychology: Animal Behavior Processes*, 37(3), 261-276.
- Hogarth, L., Retzler, C., Munafo, M. R., Tran, D. M., Troisi, J. R., Rose, A. K., Jones, A. & Field, M. (2014). Extinction of cue-evoked drug-seeking relies on degrading hierarchical instrumental expectancies. *Behaviour Research and Therapy*, 59, 61-70.
- Holland, P. C. (1977). Conditioned stimulus as a determinant of the form of the Pavlovian conditioned response. *Journal of Experimental Psychology: Animal Behavior Processes*, 3(1), 77-104.
- Holland, P. C. (1981). Acquisition of representation-mediated conditioned food aversions. *Learning and Motivation*, 12(1), 1-18.

- Holland, P. C. (2004). Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, 30(2), 104-117.
- Holland, P. C. (2008). Cognitive versus stimulus-response theories of learning. *Learning & Behavior*, 36(3), 227-241.
- Holland, P. C., & Gallagher, M. (2003). Double dissociation of the effects of lesions of basolateral and central amygdala on conditioned stimulus-potentiated feeding and Pavlovian-instrumental transfer. *European Journal of Neuroscience*, 17(8), 1680-1694.
- Holland, P. C., & Lamarre, J. (1984). Transfer of inhibition after serial and simultaneous feature negative discrimination training. *Learning and Motivation*, 15(3), 219-243.
- Holland, P. C., & Rescorla, R. A. (1975). The effect of two ways of devaluing the unconditioned stimulus after first-and second-order appetitive conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, 1(4), 355-363.
- Holman, J. G., & Mackintosh, N. J. (1981). The control of appetitive instrumental responding does not depend on classical conditioning to the discriminative stimulus. *The Quarterly Journal of Experimental Psychology*, 33(1), 21-31.
- Holmes, N. M., Marchand, A. R., & Coutureau, E. (2010). Pavlovian to instrumental transfer: a neurobehavioural

perspective. *Neuroscience & Biobehavioral Reviews*, 34(8), 1277-1295.

Honey, R. C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, 15(4), 338-346.

Konorski, J. (1948). Conditioned reflexes and neuron organization.

Konorski, J. (1967). Integrative activity of the brain: An interdisciplinary approach. *Chicago: University of Chicago*.

Krank, M. D. (2003). Pavlovian Conditioning With Ethanol: Sign-Tracking (Autoshaping), Conditioned Incentive, and Ethanol Self-Administration. *Alcoholism: Clinical and Experimental Research*, 27(10), 1592-1598.

Kruse, J. M., Overmier, J. B., Konz, W. A., & Rokke, E. (1983). Pavlovian conditioned stimulus effects upon instrumental choice behavior are reinforcer specific. *Learning and Motivation*, 14(2), 165-181.

Laurent, V., Wong, F. L., & Balleine, B. W. (2015).  $\delta$ -Opioid receptors in the accumbens shell mediate the influence of both excitatory and inhibitory predictions on choice. *British journal of pharmacology*, 172(2), 562-570.

LeBlanc, K. H., Ostlund, S. B., & Maidment, N. T. (2012). Pavlovian-to-instrumental transfer in cocaine seeking rats. *Behavioral Neuroscience*, 126(5), 681-689.

- LoLordo, V. M. (1967). Similarity of conditioned fear responses based upon different aversive events. *Journal of Comparative and Physiological Psychology*, 64(1), 154-158.
- LoLordo, V. M., & Fairless, J. L. (1985). Pavlovian conditioned inhibition: The literature since 1969. In *Information processing in animals: Conditioned inhibition*, ed. R.R. Miller, N.E. Spear, pp. 1-49. Hillsdale, N.J.: Earlbaum.
- Lovibond, P. F. (1981). Appetitive Pavlovian-instrumental interactions: Effects of inter-stimulus interval and baseline reinforcement conditions. *The Quarterly Journal of Experimental Psychology*, 33(4), 257-269.
- Lovibond, P. F. (1983). Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, 9(3), 225-247.
- Lovibond, P. F., & Colagiuri, B. (2013). Facilitation of voluntary goal-directed action by reward cues. *Psychological science*, 24(10), 2030-2037.
- Mackintosh, N. J. (1974). *The Psychology of Animal Learning*. Academic Press.
- Mackintosh, N. J. (1983). *Conditioning and Associative Learning* (p. 316). Oxford: Clarendon Press.

- Mackintosh, N. J., & Dickinson, A. (1979). Instrumental (type II) conditioning. In Dickinson, A. and Boakes, R. A. (Eds), *Mechanisms of learning and motivation*. Hillsdale, N. J.: Lawrence Erlbaum Associates.
- Mahoney, W. J., & Ayres, J. J. (1976). One-trial simultaneous and backward fear conditioning as reflected in conditioned suppression of licking in rats. *Animal Learning & Behavior*, 4(4), 357-362.
- Maier, S. F., Rapaport, P., & Wheatley, K. L. (1976). Conditioned inhibition and the UCS-CS interval. *Animal Learning & Behavior*, 4(2), 217-220.
- Martinovic, J., Jones, A., Christiansen, P., Rose, A. K., Hogarth, L., & Field, M. (2014). Electrophysiological responses to alcohol cues are not associated with Pavlovian-to-instrumental transfer in social drinkers. *PloS one*, 9(4).
- McLaren, I. P. L., & Dickinson, A. (1990). The conditioning connection. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 329(1253), 179-186.
- McNish, K. A., Betts, S. L., Brandon, S. E., & Wagner, A. R. (1997). Divergence of conditioned eyeblink and conditioned fear in backward Pavlovian training. *Animal Learning & Behavior*, 25(1), 43-52.

- Miller, R. R., Hallam, S. C., Hong, J. Y., & Dufore, D. S. (1991). Associative structure of differential inhibition: Implications for models of conditioned inhibition. *Journal of Experimental Psychology: Animal Behavior Processes*, *17*(2), 141-150.
- Milton, A. L., Schramm, M. J., Wawrzynski, J. R., Gore, F., Oikonomou-Mpegeti, F., Wang, N. Q., Samuel, D., Economidou, D. & Everitt, B. J. (2012). Antagonism at NMDA receptors, but not  $\beta$ -adrenergic receptors, disrupts the reconsolidation of pavlovian conditioned approach and instrumental transfer for ethanol-associated conditioned stimuli. *Psychopharmacology*, *219*(3), 751-761.
- Moscovitch, A., & LoLordo, V. M. (1968). Role of safety in the Pavlovian backward fear conditioning procedure. *Journal of Comparative and Physiological Psychology*, *66*(3p1), 673-678.
- Nadler, N., Delgado, M. R., & Delamater, A. R. (2011). Pavlovian to instrumental transfer of control in a human learning task. *Emotion*, *11*(5), 1112-1123.
- Nieto, J. (1984). Transfer of conditioned inhibition across different aversive reinforcers in the rat. *Learning and Motivation*, *15*(1), 37-57.
- Ostlund, S. B., & Balleine, B. W. (2007). Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. *The Journal of Neuroscience*, *27*(18), 4819-4825.

- Overmier, J. B., Bull, J. A., & Pack, K. (1971). On instrumental response interaction as explaining the influences of Pavlovian CS+ s upon avoidance behavior. *Learning and Motivation*, 2(2), 103-112.
- Pavlov, I. P. (1927). *Conditioned Reflexes. An Investigation of the physiological activity of the cerebral cortex*. London: Oxford Univ. Press.
- Pavlov, I. P. (1928). *Lectures on conditioned Reflexes*. New York: International.
- Pavlov, I. P. (1932). The reply of a physiologist to psychologists. *Psychological Review*, 39, 91-127.
- Pearce, J. M., Montgomery, A., & Dickinson, A. (1981). Contralateral transfer of inhibitory and excitatory eyelid conditioning in the rabbit. *The Quarterly Journal of Experimental Psychology*, 33(1), 45-61.
- Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1), 8-13.
- Prévost, C., Liljeholm, M., Tyszka, J. M., & O'Doherty, J. P. (2012). Neural correlates of specific and general Pavlovian-to-Instrumental Transfer within human amygdalar subregions: a high-resolution fMRI study. *The Journal of Neuroscience*, 32(24), 8383-8390.

- Randich, A., & LoLordo, V. M. (1979). Associative and nonassociative theories of the UCS preexposure phenomenon: implications for Pavlovian conditioning. *Psychological Bulletin*, 86(3), 523-548.
- Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review*, 74(1), 71-80.
- Rescorla, R. A. (1969a). Pavlovian conditioned inhibition. *Psychological Bulletin*, 72(2), 77-94.
- Rescorla, R. A. (1969b). Conditioned inhibition of fear resulting from negative CS-US contingencies. *Journal of Comparative and Physiological Psychology*, 67(4), 504-509.
- Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43(3), 151-160.
- Rescorla, R. A. (1991). Associations of multiple outcomes with an instrumental response. *Journal of Experimental Psychology: Animal Behavior Processes*, 17(4), 465.
- Rescorla, R. A. (1992a). Response-outcome versus outcome-response associations in instrumental learning. *Animal Learning & Behavior*, 20(3), 223-232.
- Rescorla, R. A. (1992b). Response-independent outcome presentation can leave instrumental RO associations intact. *Animal Learning & Behavior*, 20(2), 104-111.

- Rescorla, R. A. (1993a). Preservation of response-outcome associations through extinction. *Animal Learning & Behavior*, 21(3), 238-245.
- Rescorla, R. A. (1993b). Inhibitory associations between S and R in extinction. *Animal Learning & Behavior*, 21(4), 327-336.
- Rescorla, R. A. (1994a). Control of instrumental performance by Pavlovian and instrumental stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, 20(1), 44.
- Rescorla, R. A. (1994b). Transfer of instrumental control mediated by a devalued outcome. *Animal Learning & Behavior*, 22(1), 27-33.
- Rescorla, R. A. (1995). Full preservation of a response–outcome association through training with a second outcome. *The Quarterly Journal of Experimental Psychology*, 48(3), 252-261.
- Rescorla, R. A. (1996). Preservation of Pavlovian associations through extinction. *The Quarterly Journal of Experimental Psychology: Section B*, 49(3), 245-258.
- Rescorla, R. A., & Colwill, R. M. (1989). Associations with anticipated and obtained outcomes in instrumental learning. *Animal Learning & Behavior*, 17(3), 291-303.
- Rescorla, R. A., & Heth, C. D. (1975). Reinstatement of fear to an extinguished conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, 1(1), 88-96.

- Rescorla, R. A., & Holland, P. C. (1977). Associations in Pavlovian conditioned inhibition. *Learning and Motivation*, 8(4), 429-447.
- Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological review*, 74(3), 151-182.
- 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.
- Rosas, J. M., Paredes-Olay, M. C., García-Gutiérrez, A., Espinosa, J. J., & Abad, M. J. (2010). Outcome-specific transfer between predictive and instrumental learning is unaffected by extinction but reversed by counterconditioning in human participants. *Learning and Motivation*, 41(1), 48-66.
- Saddoris, M. P., Stamatakis, A., & Carelli, R. M. (2011). Neural correlates of Pavlovian-to-instrumental transfer in the nucleus accumbens shell are selectively potentiated following cocaine self-administration. *European Journal of Neuroscience*, 33(12), 2274-2287.
- Shurtleff, D., & Ayres, J. J. (1981). One-trial backward excitatory fear conditioning in rats: Acquisition, retention, extinction, and spontaneous recovery. *Animal Learning & Behavior*, 9(1), 65-74.

- Siegel, S., & Domjan, M. (1971). Backward conditioning as an inhibitory procedure. *Learning and Motivation*, 2(1), 1-11.
- Siegel, S., & Domjan, M. (1974). The inhibitory effect of backward conditioning as a function of the number of backward pairings. *Bulletin of the Psychonomic Society*, 4(2), 122-124.
- Solomon, R. L., & Turner, L. H. (1962). Discriminative classical conditioning in dogs paralyzed by curare can later control discriminative avoidance responses in the normal state. *Psychological Review*, 69(3), 202-218.
- Spetch, M. L., Wilkie, D. M., & Pinel, J. P. (1981). Backward conditioning: A reevaluation of the empirical evidence. *Psychological Bulletin*, 89(1), 163-175.
- Tait, R. W., & Saladin, M. E. (1986). Concurrent development of excitatory and inhibitory associations during backward conditioning. *Animal Learning & Behavior*, 14(2), 133-137.
- Terrace, H. S. (1973). Classical conditioning. *The study of behavior: Learning, motivation, emotion, and instinct*, 70-112.
- Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, 2(4).
- Tiffany, S. T. (1990). A cognitive model of drug urges and drug-use behavior: role of automatic and nonautomatic processes. *Psychological review*, 97(2), 147-168.

- Trapold, M. A., & Overmier, J. B. (1972). The second learning process in instrumental learning. In A. H. Black & W. F. Prokasy (Eds), *Classical conditioning II: Current research and theory* (pp. 427-452). New York, NY: Appleton-Century-Crofts.
- Troisi, I. I., & Joseph, R. (2010). Pavlovian-instrumental transfer of the discriminative stimulus effects of nicotine and ethanol in rats. *The Psychological Record*, 56(4), 499-512.
- Wagner, A. R. & Brandon, S. E. (1989). Evolution of a structured connectionist model of Pavlovian conditioning (AESOP). In S.B. Klein & R.R. Mowrer (Eds.), *Contemporary learning theories: Pavlovian conditioning and the status of traditional learning theory* (pp. 149-189). Hillsdale, NJ: Erlbaum.
- Wasserman, E. A., & Miller, R. R. (1997). What's elementary about associative learning?. *Annual Review of Psychology*, 48(1), 573-607.
- Watson, P., Wiers, R. W., Hommel, B., & de Wit, S. (2014). Working for food you don't desire. Cues interfere with goal-directed food-seeking. *Appetite*, 79, 139-148.
- Williams, D. A., Dyck, D. G., & Tait, R. W. (1986). Excitatory backward conditioning in conditioned punishment and conditioned suppression in rats. *The American journal of psychology*, 367-384.

- Williams, D. A., Overmier, J. B., & LoLordo, V. M. (1992). A reevaluation of Rescorla's early dictums about Pavlovian conditioned inhibition. *Psychological Bulletin*, 111(2), 275-290.
- Williams, D. A., Travis, G. M., & Overmier, J. B. (1986). Within-compound associations modulate the relative effectiveness of differential and Pavlovian conditioned inhibition procedures. *Journal of Experimental Psychology: Animal Behavior Processes*, 12(4), 351-362.
- Zamble, E. (1969). Conditioned motivational patterns in instrumental responding of rats. *Journal of comparative and physiological psychology*, 69(3), 536-543.
- Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology*, 86(5), 837-845.

# Appendix A: Task instructions

## Chapter 1

### Experiment 1

**General Instructions:** “This is a simple game in which your goal is to obtain as many rewards as you can. To do this, you will need to discover the relationships between different images during the game. When you are ready please press space to continue. Good luck!”

**Instrumental phase:** “In this part of the game you need to learn the relationship between button presses and different rewards (images of food or juice). When the “+” is present on the screen, you can press either the 'Z' or the 'M' button in order to produce one of the rewards but sometimes nothing will happen so you have to keep pressing! It is important to be focussed on the relationship between each button and reward. The more you press, the more likely you are to obtain a reward. Try to obtain as many rewards as you can. When you are ready, please press space bar to continue.”

**Pavlovian phase:** “During this part of the experiment you just have to pay attention to the images on the screen. You will see a black dot on the screen and sometimes different figures will appear. Some of these figures will be followed by rewards while others will not, so you have to discover the relationships between the figures and the images.

After these instructions you do not have to press any button, just pay attention. When you are ready, please press space button to continue.”

**Pavlovian test:** “Now some images will be presented and you will have to determine, using a scale from 1 to 100, how likely is that this image will be followed by food or juice. Read the questions carefully and answer using the mouse. After this, you will continue watching some images and you have to pay attention in the relationship between them. When you are ready press SPACE to continue.”

**Pavlovian phase (continuation):** “Now again you have to pay attention to the images on the screen. When you are ready press SPACE to continue.”

**Break:** “Now you can take a brief break. After this, you will continue responding similar questions about the relationship between the stimuli. When you are ready press SPACE to continue.”

**Summation test:** “Now again, some images will be presented and you will have to determine, using a scale from 1 to 100, how likely is that this image will be followed by food or juice. Read the questions carefully and answer using the mouse. When you are ready press SPACE to continue.”

**Instrumental retraining:** “Now again you need to learn the relationship between button presses and different rewards (images of

food or juice). When the “+” is present on the screen, you can press either the 'Z' or the 'M' button in order to produce one of the rewards but sometimes nothing will happen so you have to keep pressing! It is important to be focussed on the relationship between each button and reward. The more you press, the more likely you are to obtain a reward. Try to obtain as many rewards as you can. When you are ready, please press space bar to continue.”

**PIT test:** “In this part of the experiment, you will see either a “+” sign or one of the figures. Now you have to press either the 'Z' or the 'M' button in order to obtain the rewards. You can press the buttons as many times as you want. When you are ready, please press space bar to continue.”

**End:** “That was all. Thank you for your participation.”

## **Experiment 2**

**General instructions:** “This is a simple game in which your goal is to obtain as many rewards as you can. To do this, you will need to discover the relationships between different images during the game. When you are ready please press space to continue. Good luck!”

**Instrumental phase:** “In this part of the game you need to learn the relationship between button presses and different rewards (images of food or juice). When the “+” is present on the screen, you can press

either the 'Z' or the 'M' button in order to produce one of the rewards but sometimes nothing will happen so you have to keep pressing! It is important to be focussed on the relationship between each button and reward. The more you press, the more likely you are to obtain a reward. Try to obtain as many rewards as you can. When you are ready, please press space bar to continue.”

**Pavlovian phase:** “During this part of the experiment you will see a black dot followed by different figures. Some of these figures will be followed by rewards while others will not, so you have to discover the relationships between the figures and the rewards. At the same time, you will have to answer the question “Which reward will appear now?” using the numbers 1, 5 or 9 on the keyboard. At the beginning you will have to guess, but you will receive feedback on your response so you can learn to respond correctly. When you are ready, please press space button to continue.”

**Break:** “Now you can take a brief break. After this, you will continue responding similar questions about the relationship between the stimuli. When you are ready press SPACE to continue.”

**Summation test:** “Now some images will be presented and you will have to determine, using a scale from 1 to 100, how likely is that these images will be followed by food or juice. Read the questions carefully and answer using the mouse, clicking in the line at the point you consider correct. When you are ready press SPACE to continue.”

**Instrumental retraining:** “Now again you need to learn the relationship between button presses and different rewards (images of food or juice). When the “+” is present on the screen, you can press either the 'Z' or the 'M' button in order to produce one of the rewards but sometimes nothing will happen so you have to keep pressing! It is important to be focussed on the relationship between each button and reward. The more you press, the more likely you are to obtain a reward. Try to obtain as many rewards as you can. When you are ready, please press space bar to continue.”

**PIT test:** “In this part of the experiment, you will see either a “+” sign or one of the figures. Now you have to press either the 'Z' or the 'M' button in order to obtain the rewards. You can press the buttons as many times as you want. When you are ready, please press space bar to continue.”

**End:** “That was all. Thank you for your participation.”

### **Experiment 3**

**General instructions:** “This is a simple game in which your goal is to discover the relationships between different images, actions and rewards. When you are ready please press SPACE BAR to continue. Good luck!”

**Pavlovian phase:** “During this part of the game you will see a black dot followed by different images. Some of the images will be

followed by FOOD, some by JUICE and some by NO REWARD. Your task is to predict which reward will appear using the numbers 1 (FOOD), 5 (JUICE) and 9 (NO REWARD) on the keyboard. At the beginning you will have to guess, but you will receive feedback on your response so you can learn to predict correctly. When you are ready, please press space bar to continue.”

**Summation test:** “Now you must estimate how likely it is that each display will be followed by FOOD or JUICE. You will make your judgement using the scale. The right extreme means you think it is very LIKELY the reward will appear, while the left extreme means you think it very UNLIKELY it will appear. The centre point means there is a 50/50 chance that the reward will appear. Please read the questions carefully, and click the mouse over the line at the point you consider correct; then click in the confirmation icon. Only use the extreme ends of the scale if you are absolutely certain that the reward WILL or WILL NOT occur. When you are ready press SPACE to continue.”

**Break:** “Now you can take a brief break. When you are ready press SPACE to continue.”

**Instrumental phase:** “In this part of the game you need to learn the relationship between key presses and the different rewards (pictures of food or juice). When the “+” is present on the screen, you can press either the 'Z' or the 'M' key in order to produce one of the rewards but sometimes nothing will happen so you have to keep pressing! It is important to be focussed on the relationship between each key and

reward. The more you press, the more likely you are to obtain a reward. Try to obtain as many rewards as you can. When you are ready, please press the space bar to continue.”

**PIT test A:** “In this part of the game, you will see either a “+” sign or one of the images that you saw at the beginning of the game. Now you have to press either the 'Z' or the 'M' key in order to obtain the rewards. You can press the keys as many times as you want. When you are ready, please press the space bar to continue.”

**Instrumental retraining:** “When the “+” is present on the screen, you can press either the 'Z' or the 'M' key in order to produce one of the rewards but sometimes nothing will happen so you have to keep pressing! It is important to be focussed on the relationship between each key and reward. The more you press, the more likely you are to obtain a reward. Try to obtain as many rewards as you can. When you are ready, please press space bar to continue.”

**PIT test B:** “Again you will see either a “+” sign or one of the images that you saw at the beginning of the game. Now you have to press either the 'Z' or the 'M' key to obtain the rewards. You can press the buttons as many times as you want. When you are ready, please press the space bar to continue.”

**End:** “That was all. Thank you for your participation.”

## Chapter 3

### Experiment 4

**General instructions:** “This is a simple game in which your goal is to discover the relationships between different images, actions and rewards. Sometimes you will have to pay attention to the images on the screen and answer some questions, while other times you will have to press some keys in order to obtain rewards. Keep in mind that accurate response will produce more rewards. When you are ready please press SPACE BAR to continue. Good luck!”

**Pavlovian phase:** “During this part of the experiment you will see a black dot followed by different figures. Some of these figures will be followed by rewards while others will not, so you have to discover the relationships between the figures and the rewards. At the same time, you will have to answer the question “Which reward will appear now?” using the numbers 1, 5 or 9 on the keyboard. At the beginning you will have to guess, but you will receive feedback on your response so you can learn to respond correctly. When you are ready, please press space button to continue.”

**Summation test:** “Now some images will be presented and you will have to determine, using a scale from 'very UNLIKELY' to 'very LIKELY', how likely is that these images will be followed by food or drink. At this point you will have to guess, but later in the experiment

you will have more information to make a better judgment. Read the questions carefully and answer using the mouse, clicking in the line at the point you consider correct. When you are ready press SPACE to continue.”

**Instrumental training:** “In this part your goal is to obtain different types of rewards. To achieve this, you need to learn the relationship between button presses ('Z' and 'M' buttons) and different rewards (images of food or a drink). One of the buttons will produce food rewards and the other a drink reward. When the “+” is present on the screen, you can press either the 'Z' or the 'M' button in order to produce one of the rewards but nothing will happen if you press 'Z' or 'M' during the reward presentations. Your goal is to get at least 50 of each so it is important to be focussed on the relationship between each button and reward. You will not get rewards all the times, so you have to keep trying! When you are ready please press space to continue. Good luck!”

**PIT test:** “In this part of the experiment, you will see either a “+” sign or one of the figures. Now you have to press either the 'Z' or the 'M' button in order to obtain the rewards. You can press the buttons as many times as you want. When you are ready, please press space bar to continue.”

**End:** “That was all. Thank you for your participation.”

## **Experiment 5**

**General instructions:** "This is a simple game in which your goal is to discover the relationships between different images, actions and rewards. Sometimes you will have to pay attention to the images on the screen and answer some questions, while other times you will have to press some keys in order to obtain rewards. Keep in mind that accurate response will produce more rewards. When you are ready please press SPACE BAR to continue. Good luck!"

**Pavlovian phase:** "During this part of the experiment you will see a black dot followed by different figures. Some of these figures will be followed by rewards while others will not, so you have to discover the relationships between the figures and the rewards. At the same time, you will have to answer the question "Which reward will appear now?" using the numbers 1, 5 or 9 on the keyboard. At the beginning you will have to guess, but you will receive feedback on your response so you can learn to respond correctly. When you are ready, please press space button to continue."

**Summation test:** "Now some images will be presented and you will have to determine, using a scale from 1 to 100, how likely is that these images will be followed by food or drink. At this point you will have to guess, but later in the experiment you will have more information to make a better judgment. Read the questions carefully

and answer using the mouse, clicking in the line at the point you consider correct. When you are ready press SPACE to continue."

**Instrumental phase:** "In this part your goal is to obtain different types of rewards. To achieve this, you need to learn the relationship between button presses ('Z' and 'M' buttons) and different rewards (images of food or drink). One of the buttons will produce food rewards and the other drink rewards. When the "+" is present on the screen, you can press either the 'Z' or the 'M' button in order to produce one of the rewards but nothing will happen if you press 'Z' or 'M' during the reward presentations. Your goal is to get at least 50 of each so it is important to be focussed on the relationship between each button and reward. You will not get rewards all the times, so you have to keep trying- you can press as many times or as quickly as you see fit! When you are ready please press space to continue. Good luck!"

**PIT test:** "In this part of the experiment, you will see either a "+" sign or one of the figures. Now you have to press either the 'Z' or the 'M' button in order to obtain the rewards, just as in the first phase of the experiment. You can press the buttons as many times or as quickly as you see fit. When you are ready, please press space bar to continue."

**End:** "That was all. Thank you for your participation."

## Experiment 6

**General instructions:** "This is a simple game in which your goal is to discover the relationships between different images, actions and rewards. Sometimes you will have to pay attention to the images on the screen and answer some questions, while other times you will have to press some keys in order to obtain rewards. Keep in mind that accurate response will produce more rewards. When you are ready please press SPACE BAR to continue. Good luck!"

**Instrumental phase:** "In this part your goal is to obtain different types of rewards. To achieve this, you need to learn the relationship between button presses ('Z' and 'M' buttons) and different rewards (images of food or drink). One of the buttons will produce food rewards and the other drink rewards. When the "+" is present on the screen, you can press either the 'Z' or the 'M' button in order to produce one of the rewards but nothing will happen if you press 'Z' or 'M' during the reward presentations. Your goal is to get at least 50 of each so it is important to be focussed on the relationship between each button and reward. You will not get rewards all the times, so you have to keep trying- you can press as many times or as quickly as you see fit! When you are ready please press space to continue. Good luck!"

**Pavlovian instructions:** "During this part of the experiment you will see a black dot followed by different figures. Some of these figures will be followed by rewards while others will not, so you have to

discover the relationships between the figures and the rewards. At the same time, you will have to answer the question “Which reward will appear now?” using the numbers 1, 5 or 9 on the keyboard. At the beginning you will have to guess, but you will receive feedback on your response so you can learn to respond correctly. When you are ready, please press space button to continue."

**Summation test:** "Now some images will be presented and you will have to determine, using a scale from 1 to 100, how likely is that these images will be followed by food or drink. At this point you will have to guess, but later in the experiment you will have more information to make a better judgment. Read the questions carefully and answer using the mouse, clicking in the line at the point you consider correct. When you are ready press SPACE to continue."

**PIT test:** "In this part of the experiment, you will see either a “+” sign or one of the figures. Now you have to press either the 'Z' or the 'M' button in order to obtain the rewards, just as in the first phase of the experiment. You can press the buttons as many times or as quickly as you see fit. When you are ready, please press space bar to continue."

**End:** “That was all. Thank you for your participation.”

## **Experiment 7**

**General instructions:** "This is a simple game in which your goal is to discover the relationships between different images, actions and rewards. Sometimes you will have to pay attention to the images on the screen and answer some questions, while other times you will have to press some keys in order to obtain rewards. Keep in mind that accurate response will produce more rewards. When you are ready please press SPACE BAR to continue. Good luck!"

**Pavlovian phase:** "During this part of the experiment you will see a black dot followed by different figures. Some of these figures will be followed by rewards while others will not, so you have to discover the relationships between the figures and the rewards. At the same time, you will have to answer the question "Which reward will appear now?" using the numbers 1, 5 or 9 on the keyboard. At the beginning you will have to guess, but you will receive feedback on your response so you can learn to respond correctly. When you are ready, please press space button to continue."

**Summation test:** "Now some images will be presented and you will have to determine, using a scale, how likely is that these images will be followed by food or juice. Read the questions carefully and answer putting the mouse over the line at the point you consider correct, and then click in the confirmation icon. When you are ready press SPACE to continue."

**Instrumental phase:** "Now in this part you have to press the keys 'Z' and 'M'. You will see a "+" on the screen and you will have a counter for each key, so your goal is to press the buttons until you obtain 50 of each. You will not increase the counters all the times, so you have to keep trying- you can press as many times or as quickly as you see fit! When you are ready please press space to continue. Good luck!"

**PIT test:** "Now again you have to press either the 'Z' or the 'M' button, but this time you will not see the counters. You will see a "+" on the screen but some of the previous figures may appear. As before, you can press the buttons as many times or as quickly as you see fit. When you are ready, please press space bar to continue."

## **Experiment 8**

**General instructions:** "This is a simple game in which your goal is to discover the relationships between different images, actions and rewards. Sometimes you will have to pay attention to the images on the screen and answer some questions, while other times you will have to press some keys in order to obtain rewards. Keep in mind that accurate response will produce more rewards. When you are ready please press SPACE BAR to continue. Good luck!"

**Pavlovian phase:** "During this part of the experiment you will see a black dot followed by different figures. Some of these figures will

be followed by rewards while others will not, so you have to discover the relationships between the figures and the rewards. At the same time, you will have to answer the question "Which reward will appear now?" using the numbers 1, 5 or 9 on the keyboard. At the beginning you will have to guess, but you will receive feedback on your response so you can learn to respond correctly. When you are ready, please press space button to continue."

**Instrumental phase:** "In this part your goal is to obtain different types of rewards. To achieve this, you need to learn the relationship between button presses ('Z' and 'M' buttons) and different rewards (images of food or drink). One of the buttons will produce food rewards and the other drink rewards. When the "+" is present on the screen, you can press either the 'Z' or the 'M' button in order to produce one of the rewards but nothing will happen if you press 'Z' or 'M' during the reward presentations. Your goal is to get at least 50 of each so it is important to be focussed on the relationship between each button and reward. You will not get rewards all the times, so you have to keep trying- you can press as many times or as quickly as you see fit!"

**PIT test:** "In this part of the experiment, you will see either a "+" sign or some of the figures. Now you have to press either the 'Z' or the 'M' button in order to obtain the rewards, just as in the previous phase of the experiment. You can press the buttons as many times or as quickly as you see fit. When you are ready, please press space bar to continue."

## Chapter 4

### Experiment 9, 10 and 11

**General instructions:** "This is a simple game in which your goal is to discover the relationships between different images, actions and rewards. Sometimes you will have to pay attention to the images on the screen, while other times you will have to press some keys in order to obtain rewards. Keep in mind that accurate response will produce more rewards. When you are ready please press SPACE BAR to continue. Good luck!"

**Instrumental phase:** "In this part your goal is to obtain different types of rewards. To achieve this, you need to learn the relationship between button presses ('Z' and 'M' buttons) and different rewards (images of food or drink). One of the buttons will produce food rewards and the other drink rewards. When the "+" is present on the screen, you can press either the 'Z' or the 'M' button in order to produce one of the rewards but nothing will happen if you press 'Z' or 'M' during the reward presentations. Your goal is to get at least 50 of each so it is important to be focussed on the relationship between each button and reward. You will not get rewards all the times, so you have to keep trying- you can press as many times or as quickly as you see fit! When you are ready please press space to continue. Good luck!"

**Pavlovian phase:** "During this part of the experiment you will see a black dot followed by different figures. Your goal is to pay attention to the relationship between these figures. When you are ready, please press space button to continue."

**PIT test:** "In this part of the experiment, you will see either a "+" sign or one of the figures. Now you have to press either the 'Z' or the 'M' button in order to obtain the rewards, just as in the first phase of the experiment. You can press the buttons as many times or as quickly as you see fit. When you are ready, please press space bar to continue."

**Pavlovian test :** "Now you will see some images and you will have to determine, using a scale, to which of the rewards that image is related. Read the questions carefully and answer putting the mouse over the line at the point you consider correct, and then click in the confirmation icon. When you are ready press SPACE to continue."

## **Experiment 12**

**General instructions:** "This is a simple game in which your goal is to discover the relationships between different images, actions and rewards (images of FOOD and DRINKS). You will have to pay attention to the screen and press some keys to obtain or to predict rewards. When you are ready please press SPACE BAR to continue. Good luck!"

**Instrumental phase:** "In this part your goal is to press different keys ('Z' and 'M' keys) to obtain different types of rewards (images of DRINKS or FOODS). It will be helpful to learn the relationships between your key presses and the different rewards. A "+" will appear on the screen throughout this part of the game. Please keep your eyes focused on this. You can press either the 'Z' or the 'M' key in order to produce the rewards but nothing will happen if you press 'Z' or 'M' during the reward presentations. Your goal is to get 50 of each reward but you will not get rewards all the time, so please keep trying- you can press as many times or as quickly as you see fit! Two counters will let you know how many of each reward you have earned. When you are ready please press space to continue."

**Pavlovian phase:** "Now you will see a black dot centered on the screen. Once again please keep your eyes focused on this. At some point either a neutral image or a reward (DRINK or FOOD) will appear on the screen and if it is a DRINK you have to press 'T', but if it is a FOOD you have to press 'G'. Your goal is to press as quickly and accurately as you can. Please do not press any key during the neutral images. Sometimes the neutral images will appear before the rewards but other times after them. You may find these neutral images useful in predicting which type of reward may occur at any given time, so please pay attention to the neutral images as well as the rewards because later on you will have to answer some questions about these. When you are ready, please press space button to continue."

**PIT test:** "In this part of the experiment, you will see either a "+" sign or one of the neutral images and you have to press either the 'Z' or the 'M' key in order to obtain the rewards, just as in the first phase of the game. You can press the keys as many times or as quickly as you see fit. When you are ready, please press space bar to continue."

**Assessment questionnaire:** "Finally, you will see a few more questions about the relationships between the neutral images, the keys and the rewards. This time, please answer by using the number on the keyboard. When you are ready press SPACE to continue"

### **Experiment 13**

**General instructions:** "This is a simple game in which your goal is to discover the relationships between different images, actions and rewards (images of FOOD and DRINKS). You will have to pay attention to the screen and press some keys to obtain or to predict rewards. When you are ready please press SPACE BAR to continue. Good luck!"

**Instrumental phase (first response):** "In this part a "+" will appear on the screen. Please keep your eyes focused on this. You have to press the 'Z' key to obtain one of the rewards (image of DRINK) and your goal is to obtain as much as you can. You will not obtain rewards each time you press 'Z' key, so you need to keep pressing! A counter will let you know how many DRINK rewards you have earned. When you are ready please press space to continue."

**Instrumental instructions (second response):** Now again a “+” will appear on the screen. Please keep your eyes focused on this. Now you have to press the ‘M’ key to obtain the other reward (image of FOOD) and your goal again is to obtain as much as you can. You will not obtain rewards each time you press ‘M’ key, so you need to keep pressing! A counter will let you know how many FOOD rewards you have earned. When you are ready please press space to continue.

**Pavlovian phase:** "Now you will see a black dot centered on the screen. Once again please keep your eyes focused on this. Neutral images and rewards (DRINK or FOOD) will appear on the screen. When a DRINK appears you have to press ‘T’, and when FOOD appears you have to press ‘G’. Your goal is to press as quickly and accurately as you can. The neutral images may help you to predict the rewards, so please do not press any key during these images. In addition, you have to pay attention to the relationship between the neutral images and rewards because later on you will have to answer some questions about these. When you are ready, please press space button to continue."

**PIT test:** "In this part of the experiment, you will see either a “+” sign or one of the neutral images and you have to press either the 'Z' or the 'M' key in order to obtain the rewards, just as in the first phase of the game. You can press the keys as many times or as quickly as you see fit. When you are ready, please press space bar to continue."

**Assessment questionnaire:** "Finally, you will see a few more questions about the relationships between the neutral images, the keys and the rewards. This time, please answer by using the number on the keyboard. When you are ready press SPACE to continue"

## **Experiment 14**

**General instructions:** "This is a simple game in which your goal is to discover the relationships between different images, actions and rewards (images of FOOD and DRINKS). You will have to pay attention to the screen and press some keys to obtain or to predict rewards. When you are ready please press SPACE BAR to continue. Good luck!"

**Instrumental instructions (first response):** "In this part a "+" will appear on the screen. Please keep your eyes focused on this. You have to press the 'Z' key to obtain one of the rewards (image of DRINK) and your goal is to obtain as much as you can. You will not obtain rewards each time you press 'Z' key, so you need to keep pressing! A counter will let you know how many DRINK rewards you have earned. When you are ready please press space to continue."

**Instrumental instructions (second response):** "Now again a "+" will appear on the screen. Please keep your eyes focused on this. Now you have to press the 'M' key to obtain the other reward (image of FOOD) and your goal again is to obtain as much as you can. You

will not obtain rewards each time you press 'M' key, so you need to keep pressing! A counter will let you know how many FOOD rewards you have earned. When you are ready please press space to continue.

**Pavlovian phase:** "Now you will see a black dot centered on the screen. Once again please keep your eyes focused on this. Neutral fractal images and rewards (DRINK or FOOD) will appear on the screen. When a DRINK appears you have to press 'T', and when FOOD appears you have to press 'G'. Your goal is to press as quickly and accurately as you can. In addition, please pay close attention to the relationship between the neutral images and rewards because later on you will have to answer some questions about these. Please do not press any key during the neutral images."

**PIT test:** "In this part of the experiment, you will see either a "+" sign or one of the neutral images and you have to press either the 'Z' or the 'M' key in order to obtain the rewards, just as in the first phase of the game. You can press the keys as many times or as quickly as you see fit. When you are ready, please press space bar to continue."

**Assessment questionnaire:** "At the beginning of the task the neutral fractal images and the keys ('z' and 'm') were paired with DRINK and/or FOOD rewards, but later these rewards did not occur. Now please answer a few questions about the relationships that occurred at the BEGINNING of the task using the numbers '1', '2' and '3' on the keyboard. After each response, please indicate the degree

of confidence that you have in your answer. When you are ready press  
SPACE to continue."

## Appendix B

### Counterbalancing Tables

Table B-1. Counterbalancing conditions in Experiments 1, 2 and

3.

	X	Y	C	N1	H	N2	F	G
1	1	2	3	4	5	6	7	8
2	2	1	5	6	3	4	7	8
3	1	2	4	3	6	5	7	8
4	2	1	6	5	4	3	7	8
5	3	5	1	4	2	6	7	8
6	5	3	2	6	1	4	7	8
7	3	5	4	1	6	2	7	8
8	5	3	6	2	4	1	7	8
9	4	6	1	3	2	5	7	8
10	6	4	2	5	1	3	7	8
11	4	6	3	1	5	2	7	8
12	6	4	5	2	3	1	7	8
13	1	2	3	4	5	6	8	7
14	2	1	5	6	3	4	8	7
15	1	2	4	3	6	5	8	7
16	2	1	6	5	4	3	8	7
17	3	5	1	4	2	6	8	7
18	5	3	2	6	1	4	8	7
19	3	5	4	1	6	2	8	7
20	5	3	6	2	4	1	8	7
21	4	6	1	3	2	5	8	7
22	6	4	2	5	1	3	8	7
23	4	6	3	1	5	2	8	7
24	6	4	5	2	3	1	8	7

**Note.** Numbers 1 to 24 represent the counterbalancing conditions,

and numbers 1-8 represent unique fractal images.

**Table B-2. Counterbalancing conditions in Experiment 4.**

	A	B	X	C	H	F	G
1	1	2	4	3	7	5	6
2	2	1	4	3	7	5	6
3	1	2	3	4	7	5	6
4	2	1	3	4	7	5	6
5	1	2	4	3	7	6	5
6	2	1	4	3	7	6	5
7	1	2	3	4	7	6	5
8	2	1	3	4	7	6	5
9	1	2	4	7	3	6	5
10	2	1	4	7	3	6	5
11	1	2	4	7	3	6	5
12	2	1	4	7	3	6	5
13	1	2	7	4	3	5	6
14	2	1	7	4	3	5	6
15	1	2	7	4	3	5	6
16	2	1	7	4	3	5	6
17	1	2	3	7	4	6	5
18	2	1	3	7	4	6	5
19	1	2	3	7	4	6	5
20	2	1	3	7	4	6	5
21	1	2	7	3	4	5	6
22	2	1	7	3	4	5	6
23	1	2	7	3	4	5	6
24	2	1	7	3	4	5	6

**Note.** Numbers 1 to 24 represent the counterbalancing conditions, and numbers 1-7 represent unique fractal images. For participants with conditions from 1 to 12, O<sub>1</sub> was a food image and O<sub>2</sub> a drink image, and the reverse for participants with conditions from 13 to 24.

**Table B-3. Counterbalancing conditions in Experiment 5, 6, 7 and 8.**

	A	B	C	X	F	G
1	1	2	3	4	5	6
2	2	1	3	4	5	6
3	1	2	4	3	5	6
4	2	1	4	3	5	6
5	1	2	3	4	6	5
6	2	1	3	4	6	5
7	1	2	4	3	6	5
8	2	1	4	3	6	5

**Note.** Numbers 1 to 8 represent the counterbalancing conditions, and numbers 1-6 represent unique fractal images. This set of conditions was repeated once, and for one set  $O_1$  was a food image and  $O_2$  a drink image, and the reverse for the other set.

**Table B-4. Counterbalancing conditions in Experiment 9, 10 and 11.**

	A	B	C	D
1	1	2	3	4
2	2	1	4	3
3	3	4	1	2
4	4	3	2	1

**Note.** This set of conditions was repeated three times. For two of these sets  $O_1$  was a food image and  $O_2$  a drink image, and the reverse for the other two sets.

**Table B-4. Counterbalancing conditions in Experiment 12, 13 and 14.**

	A	B	C	D
1	1	2	3	4
2	1	2	4	3
3	2	1	3	4
4	2	1	4	3
5	3	4	1	2
6	3	4	2	1
7	4	3	1	2
8	4	3	2	1

**Note.** This set of conditions was repeated once. For one set  $O_1$  was a food image and  $O_2$  a drink image, and the reverse for the other set.