The Effect of Conditioned Inhibition on the Specific Pavlovian-Instrumental Transfer Effect

Daniel Alarcón & Charlotte Bonardi

School of Psychology, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom

Correspondence: Daniel Alarcón, School of Psychology, University of Nottingham, University Park, Nottingham, NG7 2RD, UK; email: lpxdeala@exmail.nottingham.ac.uk; cmb@psychology.nottingham.ac.uk

Acknowledgements: We would like to thank Ms Ester Benzaquén for help in collecting some of the data, and Professor Andrew Delamater for allowing us to collect data in his laboratory at Brooklyn College - CUNY.

This work was funded by the Chilean Government (BECAS-CHILE scholarship).

Running Head: Inhibition and PIT

Abstract

Four experiments examined the effect of Pavlovian conditioned inhibition on specific Pavlovian-instrumental transfer (PIT) in human participants. The task comprised an instrumental phase in which two responses (R_1, R_2) were each paired with one of two outcomes (O_1, O_2: R_1-->O_1, R_2-->O_2), and a Pavlovian phase, in which two CSs, CS_1 and CS_2 each signalled one of the two outcomes (CS_1-->O_1, CS_2-->O_2). In Experiments 1-2 a conditioned inhibitor, X, predicted the omission of one of the outcomes (e.g. CS_1-->O_1, CS_1X-->nothing). In a subsequent test, performance of R_1 and R_2 was examined in the presence of CS_1 and CS_2. A specific PIT effect was observed: R_1 was performed more than R_2 during CS_1, and R_2 more than R_1 during CS_2. This PIT effect was significantly reduced by the presence of the inhibitor X in Experiment 1, in which the Pavlovian phase followed the instrumental phase, and in Experiment 2 in which it preceded it. No such effect was observed when X was presented in the absence of any expectation of the outcomes during the PIT test (Experiment 3a), or when X was trained as a signal for an alternative outcome (Experiment 3b). These results are consistent with the suggestion that the specific PIT effect occurs through a stimulus-outcome-response (S-O-R) mechanism, according to which the CS evokes a representation of the outcome which in turn elicits the response (e.g. CS_1-->O_1-->R_1). The conditioned inhibitor suppresses performance of the response by suppressing activation of the outcome representation.

Keywords: conditioned inhibition, specific PIT, stimulus-outcome-response.

One key feature of addiction is that addicts experience craving and relapse even after periods of detoxification (Sanchis-Segura & Spanagel, 2008). This has been attributed in part to Pavlovian conditioning, in which a neutral *conditioned stimulus* (CS) reliably signals the occurrence of a motivationally significant *unconditioned stimulus* (US) (e.g., Robbins & Everitt, 1999; Stewart, de Wit & Eikelboom, 1984). This training results in the CS eliciting a conditioned response (CR) indicating anticipation of the US, and also taking on some of the US's motivational properties. Thus CSs associated with a drug come to attract attention, elicit approach responses (e.g. Brown & Jenkins, 1968), acquire secondary reinforcing properties such that subjects will work to obtain them (e.g. Fantino, 1977), and become able to invigorate behaviour (e.g. Bindra, 1968; Saunders & Robinson, 2013). In the context of addiction, this means CSs for the drug can elicit drug-seeking behaviour - raising the possibility that this mechanism could play a central role in relapse.

This effect of CSs on instrumental behaviour is studied in the *Pavlovian-instrumental transfer* (PIT) task (Holmes, Marchand & Coutureau, 2010). For example, in a *specific PIT* task two different outcomes, O_1 and O_2, are paired with two different CSs, S_1 and S_2, and two different responses, R_1 and R_2 (i.e. S_1-->O_1 S_2-->O_2, R_1-->O_1 R_2-->O_2). When the subject is permitted to perform R_1 and R_2 in the presence of S_1 and S_2, each CS selectively increases the rate of the response associated with the *same outcome*: thus during S_1 R_1 is performed more than R_2, and during S_2 the opposite. This specific PIT effect has been widely studied in studies with both humans and animals (e.g., Colagiuri & Lovibond, 2015; Corbit & Janak, 2007; Glasner, Overmier & Balleine, 2005; Hogarth, Dickinson, Wright, Kouvaraki & Duka, 2007) - yet the mechanism underlying it is still unclear. One influential explanation of specific PIT is the stimulus-outcome-response (S-O-R) account. One version of this account[1] relies on the assumption that the R-->O association formed during the instrumental phase can operate bidirectionally (cf., Mackintosh & Dickinson, 1979), such that presentation of O can elicit performance of R, just as R activates the representation of O

[1] An alternative approach will be taken up below (Trapold & Overmier, 1972).

(Balleine & Ostlund, 2007; Cohen-Hatton, Haddon, George & Honey, 2013; Gilroy, Everett & Delamater, 2014). Thus in a specific PIT task S_1 will activate the representation of O_1 and hence elicit performance of R_1, while S_2 will elicit R_2 by activating the representation of O_2.

But it has been argued that some aspects of this account are paradoxical (cf. Cohen-Hatton et al., 2013): if activation of O is critical for S to elicit R, any modification to the S-O association, or to the value of O, before test might be expected to affect specific PIT. Yet extinction of the S-O association does not typically influence the ability of the CS to produce specific PIT (e.g., Delamater, 1996; Hogarth et al., 2014). Although there is good evidence that S-O associations survive extinction treatments (Bouton, 1993), their effectiveness is still reduced, and so some reduction of specific PIT might be expected. Moreover, outcome devaluation selectively depresses instrumental responding for the outcome, yet does not reduce specific PIT in animals (e.g., Holland, 2004). The former result suggests that outcome devaluation, like specific PIT, is partly mediated by the outcomes' unique sensory properties so, as Cohen-Hatton et al. (2013) noted, it is puzzling that one is affected while the other is not (e.g. Balleine & Ostlund, 2007). However, evidence on this issue in human participants is mixed: while Hogarth & Chase (2011) found specific PIT remained intact in humans after outcome devaluation, Allman, DeLeon, Cataldo & Johnson (2010) did not.

Understanding the specific PIT effect and how to reduce it has both theoretical and practical implications. Thus, given the inconsistency of existing evidence on this issue in humans, the present experiments employed a different approach. In a Pavlovian conditioned inhibition task a CS is always followed by the US unless it is accompanied by a second, stimulus (i.e. A --> US AX --> no US; cf. Rescorla, 1969). This results in X signalling the *absence* of the expected outcome, and allows it to counteract the tendency of A to produce the CR. In these cases X is termed a *conditioned inhibitor* (CI), and is assumed to operate by suppressing the activation of the US representation which mediates the ability of the CS to elicit the CR (Rescorla & Holland, 1977; cf. Rescorla, 1969). According to the S-O-R account, specific PIT occurs because at test the CS activates the representation of the

outcome with which it is associated, thus eliciting the response reinforced with that outcome. Thus the S-O-R account predicts that a conditioned inhibitor should reduce the CS's ability to activate the outcome representation, and thus reduce specific PIT. The experiments were designed to test this prediction. Work in animals suggests that conditioned inhibition attenuates specific PIT in rats (Delamater, Lolordo & Sosa, 2003; Laurent, Wong & Balleine, 2014), but to our knowledge no such effects have yet been reported in humans (although see Colagiuri & Lovibond, 2015 for related findings in a general transfer task).

Experiment 1

The task was a computer game comprising a Pavlovian phase, an instrumental training phase and a transfer test (see Table 1). Neutral fractal images were used as conditioned stimuli, and the outcomes were images of foods and drinks (O_1, O_2). In the Pavlovian stage participants were instructed to learn the relationships between the CSs and the outcomes. During initial *pretraining* a single cue, A, and a compound, AB, were followed by one of the outcomes (i.e. A-->O_1, AB-->O_1), to establish A as a signal for O_1. In the *inhibitory training* phase that followed, non-reinforced presentations of A in compound with a new stimulus, X, were added, to establish X as an inhibitor for O_1 (AX-). These trials were accompanied by nonreinforced presentations of two further novel stimuli in a compound, CH-. C and H were treated in exactly the same way as X (being presented the same number of times in a nonreinforced compound) - differing only in that, unlike X, they did not signal the omission of an expected outcome (He et al., 2011; 2012). In the *test excitor training* phase which followed, F was paired with O_1, before the inhibitory properties of X were assessed in a *summation test.* Here F was presented in compound with the neutral control cue, C, or the inhibitor, X (i.e. FC and FX), and participants had to rate the extent to which each of the compounds predicted O_1. If X was a conditioned inhibitor it would suppress the activation of O_1 produced by the excitatory F; thus expectation of O_1 would be lower during FX than FC.

Then participants received *instrumental training,* in which they pressed the *z* and *m* keys (R_1, R_2) to obtain the food and drink USs (O_1 and O_2), with each key being reinforced

with a different outcome (R_1-->O_1 and R_2-->O_2). Finally subjects were given a *PIT test*, during which they continued making R_1 and R_2 while FC and FX were presented. No outcomes were delivered during this phase, and the rate of R_1 and R_2 during stimulus presentations was recorded. As C was never paired with either O_1 or O_2, or their unexpected omission, it should not affect the ability of F predict O_1. Thus we anticipated that standard specific PIT - more performance of R_1 than R_2 - would be observed during FC. The question was whether X would reduce the magnitude of this difference: if the S-O-R account is correct and if X can reduce the ability of F to evoke O_1 then the PIT effect should be reduced.

We also examined whether X, which signalled the absence of O_1, might also influence activation of the O_2 representation. There is some controversy over the extent to which conditioned inhibitors are outcome-specific. Konorski (1967) made a distinction between preparatory and consummatory responses, which respectively reflect motivational and sensory aspects of the US (cf., Wagner & Brandon, 1989). Thus if a conditioned inhibitor suppresses activation of the sensory features that distinguish the outcome from others of the same class, then it will be specific to CSs predicting the outcome whose absence it signalled; but if it acts on the motivational component of the outcome, it will be effective with any CS that signals an outcome from the same motivational class. While some have reported that conditioned inhibitors are not specific to the outcome's sensory properties (LoLordo, 1967; Nieto, 1984; Pearce, Montgomery & Dickinson, 1981), others have found evidence for outcome-specific inhibitory associations (Delamater, Lolordo & Sosa, 2003; Laurent, Wong & Balleine, 2015; see also Kruse, Overmier, Konz & Rokke, 1983). Thus, as O_1 and O_2 were of the same motivational valence, differing only in their sensory properties, we also examined whether the inhibitory properties of X would be evident with a CS that predicted O_2. A second test excitor, G, was paired with O_2, and GX and GH were also presented during the summation and PIT tests. If X can suppress activation of the O_2 as well as the O_1 representation, then expectation of O_2 would be lower during GX than GH in the summation test. We also tested whether X could reduce the size of any PIT effect shown to

G. We anticipated specific PIT - in the presence of G in compound with the neutral control cue H, subjects should perform R_2, which had like G signalled O_2 in training, more than R_1. The question was whether the size of this effect would be reduced during GX.

Table 1 about here

Method

Participants: 24 students aged between 18 - 24 years (5 males, 19 females) from the University of Nottingham and Brooklyn College, New York, participated. Psychology undergraduates received course credit, others a £4 inconvenience allowance.

Apparatus and Materials: The task was programmed in PsychoPy (Peirce, 2007) on a computer with a 20 inch screen. General instructions were given at the start, and specific ones before each phase. Seven fractal images were used as CSs, and 8 food and 8 drink images as outcomes; no outcome was a white square; all measured 8 x 8 cm. Each time an outcome was scheduled, a random image was selected from the corresponding set (food or drink). Images were presented in three, equally-sized positions on the left, centre and right of the screen; CSs were presented in the left and right positions, and outcome/no outcome images in the centre. Single CSs were positioned equally often on the left or right, while for CS compounds each CS component was presented an equal number of times on each side.

During the Pavlovian phase the fixation point was a black dot (3 mm in diameter), and in the instrumental phase and PIT test it was a black cross (10 x 10 mm); both were presented in the middle of the screen. In the summation test a rating scale was positioned at the bottom of the screen with an anchor label at each end. The responses were pressing the keys *z* and *m* (instrumental phase and PIT test), pressing the *1*, *5* and *9* keys (Pavlovian phase), and mouse clicking (summation test). Participants also had to press the space bar to start each phase after reading the instructions. For half the subjects O_1 was drink and O_2

food, and for the remainder the reverse. For all subjects *z* was reinforced with drink and *m* with food; thus for half the subjects *z* served as R_1 and *m* as R_2,, and for the remainder the reverse. C, H and X were counterbalanced with each other, as were F and G, and A and B. These constraints dictated that the experiment include a minimum of 24 subjects.

Procedure

The experiment was conducted in a quiet room. Participants were given an information sheet, and after any questions they were asked to complete a consent form.

Pavlovian stage: Participants were instructed to learn the relationships between different images and rewards, and answer questions about them. There were three phases of trials: the first, *pretraining*, comprised A-O_1 and AB-O_1 pairings (4 of each); the second, *inhibitory training* A-O_1, AB-O_1 pairings (8 of each) and AX- and CH- pairings (12 of each); and the third, *test excitor training,* F-O_1 and G-O_2 pairings (8 of each). Each phase comprised one trial block; the order of the different trial types was semi-random within each trial block.

Each trial was preceded by a 2-s fixation point, after which a CS or CS compound was presented, with the text "*Which reward will appear now?*" at the top of the screen; participants had to indicate which outcome they expected to occur, according to text presented at the bottom of the screen. In the pretraining and inhibitory phases this was "*1) Food 5) Nothing*" was when O_1 was food and " *5) Nothing 9) Drink* " when O_1 was drink; in the test excitor phase it was "*1) Food 5) Nothing 9) Drink*". Both the CS(s) and text remained on the screen until the participant had responded, after which the text disappeared and the corresponding outcome was presented alongside the CS(s). If the participant was correct, the feedback phrase "*Correct!*" was presented above the outcome in green; otherwise, the phrase "*Oops! That was wrong*" appeared below the outcome in red. After two seconds the CS(s), outcome and feedback were replaced by the fixation point, initiating a new trial.

Summation test: Participants were told to rate the likelihood that FC, FX, GH and GX would be followed by one of the outcomes. On each trial the text "*How likely is it that this image will*

be followed by" was presented at the top of the screen, and "*FOOD*" or "*DRINK*" below it (depending on whether F or G was present). At the bottom a rating scale with the text "*very UNLIKELY*" on the left and "*very LIKELY*" on the right was presented. Test compounds were presented until the participant clicked on a point on the scale, at which point the next test compound appeared. This test was divided into 2 blocks, each with 2 presentations of each of the four test compounds presented in semi-random order (16 trials in total).

Instrumental phase: Participants were instructed to press the *z* or *m* keys to discover the relationship between these responses and the food and drink images (i.e. R_1 followed by O_1, and R_2 by O_2) and to obtain at least 50 rewards of each type. Responding on each key was followed by presentation of its respective outcome according to a variable ratio (VR) 5 schedule. The fixation cross was present throughout this phase, except during outcome presentations, which lasted 0.8 seconds unless *z* or *m* was pressed, in which case the outcome was immediately replaced by the fixation cross. There were two response counters, one in each of the top corners of the screen; one comprised the word 'Food' and '=0' next to it, both in blue, and the other the word 'Drink' with '=0' next to it, both in orange; each time the participant received an outcome, the corresponding counter incremented by 1. The phase ended when both counter values had each reached at least 50.

PIT test: Participants received the instructions: "*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.*" The fixation cross was present throughout this phase. Each trial comprised a 2-s preCS period followed by a 2-s presentation of one of the test compounds (FC, FX, GH or GX). Responses were recorded separately during preCS and CS periods. There were four blocks of trials, each comprising two of each test compound presented in a semi-random order.

Statistical analysis: Data were analysed using ANOVA, significant two-way interactions with simple main effects analysis using the pooled error term, and significant three-way

interactions with further two-way ANOVAs. η_p^2 and its 95% confidence interval (CI) were given for significant effects and interactions in the ANOVAs. If direct support for the null hypothesis was more relevant to the question of interest, we conducted a Bayesian analogue of the paired t-test (Rouder et al., 2009; http://pcl.missouri.edu/bayesfactor). This analysis allows quantification of the evidence for the null hypothesis in a way that more standard significance testing does not (Wagenmakers, 2007). It assumes effect sizes that equal zero under the null hypothesis, and are different from zero under the alternative hypothesis, and yields a Bayes factor which indicates how much more likely the null hypothesis is than the alternative; scaled JZS Bayes factors are reported. Values of >=3 (indicating the data are 3 times more likely to occur under the null hypothesis than under the alternative) may be taken as evidence for the null hypothesis (Jeffreys, 1961). In cases where the Bayes factor associated with an interaction was computed, we employed the F value of the interaction (as all analyses were within-subject, all interactions can be reduced to a 2-way comparison with an identical *F* and *p*). In the PIT test both preCS and CS scores were grouped as *congruent* or *incongruent*, according to whether the response and the test excitor in the CS compound had signalled the same outcome; e.g., R_1 and F were both reinforced with O_1, so for FC and FX, R_1 was *congruent* and R_2 *incongruent* (and the opposite for GH and GX). Then *corrected scores* were obtained for each response by computing the rate (in responses per minute) during each CS compound and subtracting the rate during the corresponding preCS period; this reflected the degree to which CS presentation elevated performance of that response above baseline. As the PIT test was conducted in extinction, we anticipated that effects would be transient; but to ensure our observation window was large enough we gave four test blocks in all studies. Data from the first two blocks are reported unless otherwise stated.

Results

Pavlovian stage: In all studies we applied a criterion to ensure that by the end of training subjects had learned about the inhibitor X, and also associated the two test excitors F and G

with the correct outcomes. We excluded any subject scoring 50% correct or less for (i) AX trials in the second half of the *inhibitory training* phase, and/or (ii) F and G trials in the second half of the *test excitor training* phase. Two subjects were excluded in this study.

The mean number of correct responses for each of the four trial types in inhibitory training were computed in four trial blocks (each of 2A, 2AB, 3AX and 3CH trials); data for A and AB trials in pretraining were computed in a similar manner. This gave a total of six blocks, data from which are shown in the upper panel of Figure 1. Participants were clearly responding at a high level of accuracy by the last trial block. ANOVA performed on data from the four blocks of the inhibitory training phase, with trial type (A, AB, AX, CH) and trial block as factors, revealed a significant interaction, $F(9, 189) = 5.13$, $p < .001$, $MSe = .018$, $\eta_p^2 = .20$, C.I. = [.07, .26]; there were significant effects of trial type on blocks 3 and 5, $F(3, 63) = 10.96$, $p < .001$, $MSe = .023$ and $F(3, 63) = 2.92$, $p = .041$, $MSe = .023$ respectively, but not blocks 4 or 6, largest $F(3, 63) = 1.65$, $p = .19$, $MSe = .023$. Tukeys tests revealed accuracy was lower on AX trials than on A or AB trials in block 3, but no differences on block 5. This transient lack of accuracy on AX trials could reflect the fact that A had previously always preceded O_1 - so trials on which AX was followed by nothing violated previous learning.

Data for F and G are shown in the lower panel of Figure 1, computed in four trial blocks each of 2F and 2G trials. ANOVA with trial type and block as factors revealed a significant trial type x block interaction, $F(3, 63) = 13.53$, $p < .001$, $MSe = .029$, $\eta_p^2 = .39$, C.I. = [.18, .52]: accuracy was lower to F than to G on block 1, $F(1, 21) = 36.51$, $p < .001$, $MSe = .040$, but not on any subsequent block, Fs < 1. One possible explanation for this could be that the introduction of O_2 led subjects to suppress their prior learning that stimuli were followed by either O_1 or nothing, which reduced accuracy on F-O_1 trials. Alternatively it could reflect a US preexposure effect slowing learning about O_1, which had been experienced extensively, relative to learning about the newly introduced O_2.

Figure 1 about here

Summation test: Ratings for how likely the outcome was to occur were converted to a scale running from 0 ('very UNLIKELY') to 1 ('very LIKELY'). Mean rating scores for FX FC GX GH are shown in Figure 2 (upper panel). O_1 was rated less likely during FX than during FC, and O_2 less likely during GX than GH. ANOVA with CS (F, G) and trial type (X, C/H) as factors confirmed this, revealing a significant effect of trial type $F(1, 21) = 7.33$, $p = 0.013$, $MSe = 0.053$, $\eta_p^2 = .26$, C.I. = [.01, .50]. Neither the effect of CS, $F(1, 21) = 3.99$, $p = 0.059$, $MSe = 0.031$, nor the interaction, $F < 1$, were significant. This suggests that X was acting as a CI, counteracting the ability of both F and G to signal their respective outcomes. The fact that X, which had signalled the absence of O_1, was also effective with G which had signalled O_2, implies the inhibitory effect of X was not specific to O_1 in this test. To confirm this we conducted a Bayesian analysis, which yielded a value of 3.63, meaning we can accept the null hypothesis.

Figure 2 about here

Instrumental training phase: All subjects learned to make the two responses.

PIT test. The mean corrected response rates on congruent and incongruent trials with FC, GH, FX and GX are shown in Figure 2 (lower panel); the data are averaged over both test blocks (rates of congruent and incongruent responding for each CS compound in each test block are shown in Table 2). The specific PIT effect - greater responding on congruent than on incongruent trials - seemed to be strong with FC, but numerically reversed with the inhibitory compound FX. A specific PIT effect was also evident for GH, but appeared only slightly reduced on GX trials. ANOVA with congruency (congruent or incongruent), CS (F, G), trial type (X, C/H) and trial block as factors confirmed that the critical interaction between congruency and trial type was significant, $F(1, 21) = 4.79$, $p = .04$, $MSe = 17.55$, $\eta_p^2 = .19$, C.I. = [0, .44]; congruency also interacted with block, $F(1, 21) = 4.64$, $p = .043$, $MSe = 13.41$, $\eta_p^2 = .18$, C.I. = [0, .43]; nothing else was significant, largest $F(1, 21) = 4.03$, $p = .058$, $MSe = 29.31$. Further analysis of the critical congruency x trial type interaction confirmed an effect

of congruency on FC and GH trials, $F(1, 21) = 6.85$, $p = .016$, $MSe = 29.31$, but not on FX and GX trials, $F < 1$. There was also a congruency effect on block 2, $F(1, 21) = 6.02$, $p = .023$, $MSe = 29.31$, but not block 1, $F < 1$. Despite the pattern seen in Figure 2, the interaction between congruency, trial type and CS was not significant, $F(1, 21) = 1.90$, $p = .18$, $MSe = 24.47$, providing no support for the idea that reduction in PIT was more marked with F than G. We also conducted a Bayesian analysis, which yielded a Bayes factor of 1.96. This is inconclusive, meaning we cannot reject the null hypothesis in this case.

Table 2 about here

In order to ensure that the baseline preCS scores from which the corrected scores were derived did not differ, we also computed rates of preCS responding (see supplementary materials). Pooled over blocks these were 40.9, 48.8, 49.9 and 47.4 for FX, FC, GX and GH respectively. ANOVA with congruency (congruent or incongruent), CS (F, G), trial type (X, C/H) and trial block as factors revealed a main effect of congruency, $F(1, 21) = 7.65$, $p = 0.012$, $MSe = 1.12$, $\eta_p^2 = .27$, C.I. [.01, .51]. Nothing else was significant, largest $F(1, 21) = 5.75$, $p = .13$, $MSe = 2.34$. The effect of congruency arose because preCS responding was *higher* on congruent (49.1 rpm) than on incongruent trials (44.4 rpm); a higher preCS rate on congruent trials would reduce the resultant corrected score and thus, if anything, obscure the specific PIT effect that we observed in the corrected response rates. Importantly, the critical congruency x trial type interaction, and all higher order interactions involving these factors, were not significant, $Fs < 1$. Thus this preCS difference does not compromise our key finding, that the difference in congruent and incongruent responding was greater on FC and GH than on FX and GX trials.

Discussion

These results suggest that a CI can reduce the specific PIT effect produced by an excitatory CS. PIT was present when F was presented with the control stimulus C, but not

during FX - in fact subjects responded numerically more on incongruent than on congruent trials during FX. This is consistent with the S-O-R account, which attributes the PIT effect to the ability of F to activate the representation of the outcome it predicts: to the extent that the conditioned inhibitor X prevents that activation, PIT should be reduced. The result parallels those reported in rats by Laurent, Wong & Balleine (2015); in their Experiment 2 they used a Pavlovian conditioned inhibition procedure, in which two CSs, each paired with a different outcome, were each nonreinforced when presented with one of two inhibitory cues (A-->O_1, AX-, B-->O_2, BY-). They also found that, when each CS was presented with the inhibitory cue it was trained with, responding was higher on incongruent than on congruent trials (cf. Delamater, LoLordo & Sosa, 2003). Our results complement theirs by showing a parallel result in human participants, in conjunction with an independent summation test confirming that participants were significantly less likely to expect outcome O_1 on FX than on FC trials (He et al., 2011; 2012). One purpose of Experiment 2 was to replicate this finding.

Laurent et al. (2015) also found that when each CS was presented with the alternative inhibitory cue, a normal specific PIT effect was observed. This is exactly what would be predicted by the S-O-R account: specific PIT relies on S_1 being able to activate R_1 rather than R_2 because it is associated with O_1 rather than O_2 - the effect relies on each stimulus being associated with the features that differentiate the two outcomes, rather than the ones that they share. In this sense our results were paradoxical, as there was less indication that X's properties were specific to O_1 whose absence it signalled during training. In the summation test X was *also* able to reduce the subjects' expectation of O_2 in the presence of G, and the Bayesian analysis performed on the critical interaction confirmed that the inhibitory effect of X on F and G did not differ. However, the results of the PIT test were less clear in this respect; numerically the effect of X appeared less robust with G than with F, and the Bayesian analysis did not permit rejection of the null hypothesis. Thus, the suggestion that X was less effective with G than with F in the PIT test is neither supported nor refuted by these data. Experiment 2 also aimed to try and resolve this issue.

Experiment 2

Experiment 2 was designed to replicate the key result of Experiment 1, the reduction of specific PIT produced by a conditioned inhibitor, and also to explore the generality of the effect. An alternative interpretation of specific PIT has been proposed, suggesting it depends on a *direct* association between S and R (Cohen-Hatton et al., 2013). If R-->O associations operate bidirectionally, during instrumental training presentation of O_1 becomes able to elicit R_1, so that when S_1 is presented followed by O_1, this outcome will activate the R_1 representation, allowing the formation of a direct S_1-->R_1 association.[2] Thus at test S_1 can activate R_1 directly without need for mediation by O - thus explaining why specific PIT is unaffected by extinction of the S-O association, or devaluation of O, between PIT training and test. Cohen-Hatton et al., (2013) have reported evidence in rats supporting this interpretation - but it is not clear they can explain our results. If the ability of F to elicit R_1 at test relies on a *direct* F-R_1 association, X suppressing activation of O should have no effect on specific PIT. But the fact there is evidence for both mechanisms suggests they might work in parallel, with one or other predominating depending on the procedural details. It is possible that the procedure employed in our Experiment 1 favoured the S-O-R mechanism, meaning we would not observe an effect of X after training which fostered S-R link formation. For example, if there were two potential contributors to the PIT effect, one of which (S-O-R) were susceptible to conditioned inhibition and the other (S-R) not, then if S-R formation is weak, the only determinant of PIT would be S-O-R, and any reduction produced by conditioned inhibition would eliminate specific PIT; but if the S-R association were strong, conditioned inhibition could only reduce, but not eliminate specific PIT. Thus it is important to establish that the effect reported in Experiment 1 is general enough that it can be observed under conditions in which S-R learning is promoted. In fact two features of our procedure are especially relevant in this respect. First, in the critical studies (Experiments 2, 3 & 4) reported

[2] A parallel process can be invoked when instrumental training occurs *after* the Pavlovian conditioning stage. The Pavlovian association is also assumed to be bidirectional, so that during instrumental training O can activate S while the R representation is active, which is also assumed to result in formation of an S-R association.

by Cohen-Hatton et al. (2013) Pavlovian training always *followed* instrumental training, whereas in our experiment it preceded it. When Pavlovian training follows instrumental training, the S-R association forms because O evokes a representation of R during S. As S is extended in time, it is likely to be accompanied by a *simultaneous* evocation of R, fostering a strong S-R link. But if Pavlovian training *precedes* instrumental training, the S-R link forms because O evokes a representation of S during R - and as R will be relatively brief, it is more likely to be *followed* by evocation of S - favouring formation of an R-S rather than an S-R association. Thus the S-R link will be weaker when instrumental training comes *after* the Pavlovian stage, as it did in our study, than when it comes before. Second, in Cohen-Hatton et al. (2013)'s studies Pavlovian training occurred *immediately* after instrumental training, while in ours the summation test occurred *between* these two stages. As the summation test was conducted in extinction, it could have attenuated the F-O_1 link, weakening the degree to which O_1 could evoke F - which in turn would attenuate formation of the critical S-R links. Thus in Experiment 2 instrumental training occurred before the Pavlovian phase.

Participants first received instrumental training, which was immediately followed by training with test excitors F and G. Inhibitory training followed, and then the summation and PIT tests. The inhibitory training was also modified, as previous work suggests that the A-->O_1 AX- procedure used in Experiment 1 might not be optimal for producing CI, as within-compound associations could form between X and A on AX trials, so when X is presented at test it can activate A, and thus evoke O_1, counteracting X's inhibitory properties (Williams, Travis & Overmier, 1986). Thus we added trials on which X was presented alone during inhibitory training (Table 3), which should extinguish within-compound associations between X and A. To match its training with that of X, C was also presented alone during inhibitory training, and served as the sole control stimulus for the summation and PIT tests.

Table 3 about here

Method

Participants: 32 students from the University of Nottingham aged between 18 and 32 (12 males and 20 females) participated in this experiment. Application of the criterion resulted in the exclusion of two participants, giving a total of 30 participants.

Procedure

Unless otherwise stated procedure and analysis of the data from each stage was identical to that of the previous experiment. As H was not employed as a test stimulus in this experiment it was not counterbalanced with C and X, so that counterbalancing constraints dictated that the experiment include a minimum of 16 subjects.

Instrumental phase

Pavlovian phase: Test excitor training was conducted immediately after instrumental training, and was followed by the pretraining and inhibitory training phases. Inhibitory training was identical to that of Experiment 1, except for the addition of six X- and six C- trials.

Summation test: Identical to that of Experiment 1 except for the replacement of GH by GC.

PIT test: Identical to that of Experiment 1 except for the replacement of GH by GC.

Data treatment: Transient differences in the rates of preCS responding during the PIT test were observed in the first trial block; thus the data reported below were drawn from three trial blocks, at which point no critical differences in preCS responding were significant. Otherwise this was identical to that of the previous experiment.

Results

Pavlovian phase: Acquisition in the Pavlovian phase was similar to that of Experiment 1. The mean ratings for A were 0.65, 0.95, 0.93, 0.92, 0.97 and 0.93, and for AB 0.78, 1.00, 0.85, 0.92, 0.83 and 0.92 for each of the six blocks. Mean ratings for AX trials were 0.60, 0.96, 0.93 and 0.97, and those for CH 0.86, 0.99, 0.98 and 0.98 for each of the inhibitory training blocks. ANOVA performed on the inhibitory training data, with trial type and trial block as

factors, revealed a significant interaction, $F(9, 261) = 7.71$, $p < .001$, $MSe = .026$, $\eta_p^2 = .21$, C.I. = [.11, .27]; there was a significant effect of trial type on blocks 1 and 3, $F(3, 87) = 17.57$, $p < .001$, $MSe = .036$ and $F(3, 87) = 3.62$, $p = .016$, $MSe = .0136$ but not on any other block, largest $F(3, 87) = 1.02$, $p = .39$, $MSe = .036$. Tukeys tests showed accuracy on AX trials was lower than on the other trial types in block 1, and that on block 3 accuracy on AB trials was lower than on all other trial types. Mean ratings for F were 0.60, 0.96, 0.93 and 0.97, and for G 0.86, 0.99, 0.98 and 0.98. ANOVA on these data with trial type and block as factors revealed a significant effect of block, $F(3, 87) = 25.84$, $p < .001$, $MSe = .046$, $\eta_p^2 = .47$, C.I. = [.30, .57]; nothing else was significant, largest $F(3, 87) = 1.48$, $p = .225$, $MSe = .047$. Here there was no difference in accuracy to F and G in the first trial block - supporting the idea that the difference in Experiment 1 arose because O_2 had just been introduced. In this study participants received instrumental training before F and G trials, and so would have been equally exposed to O_1 and O_2 at the start of this phase.

Summation test: The summation test data are presented in Figure 3 (upper panel), and are similar to those of Experiment 1. ANOVA with CS (F,G) and trial type (C, X) as factors revealed a significant effect of trial type, $F(1, 29) = 4.92$, $p = 0.04$, $MSe = 0.034$, $\eta_p^2 = .15$, C.I. = [.00, .37]; nothing else was significant, largest $F(1, 29) = 3.03$, $p = 0.09$, $MSe = 0.054$ for the effect of CS. The interaction between CS and trial type was not significant, $F < 1$; the associated Bayes factor was 4.72, supporting the null hypothesis. Thus X was again no more effective with F than with G in the summation test.

Figure 3 about here

PIT test: The mean rates of congruent and incongruent responding for each test compound, averaged over all three trial blocks, are shown in Figure 3 (lower panel; the mean rates of congruent and incongruent responding for each CS compound by trial block are shown in Table 2.). The PIT effect seen with F again appeared substantially less marked when X was present. In this experiment the effect with G also seemed somewhat reduced: although

responding on incongruent trials was broadly similar for GC and GX, congruent responding appeared lower for GX than for GC. ANOVA with congruency, CS (F, G) trial type (C, X) and block as factors revealed a significant effect of congruency, $F(1, 29) = 10.27$, $p = 0.003$, $MSe = 84.18$, $\eta_p^2 = .26$, C.I. = [.03, .48], which interacted significantly with trial type, $F(1, 29) = 5.48$, $p = 0.026$, $MSe = 19.74$, $\eta_p^2 = .16$, C.I. = [.00, .38]; nothing else was significant, largest $F(2, 58) = 2.96$, $p = 0.06$, $MSe = 18.37$. Exploration of the critical interaction between congruency and trial type revealed a significant effect of congruency on C trials, $F(1, 29) = 9.41$, $p = 0.005$, $MSe = 84.18$ but not on X trials, $F(1, 29) = 2.15$, $p = 0.154$, $MSe = 84.18$. The interaction between congruency, trial type and CS was again not significant, $F < 1$, consistent with the suggestion that X was able to reduce the specific PIT with both F and G. This was further supported by the Bayes analysis, which yielded a Bayes factor of 3.38.

A corresponding ANOVA performed on preCS responding for the three trial blocks (Table 4) revealed a significant three-way interaction between block, congruency and trial type, $F(2, 58) = 3.52$, $p = .036$, $MSe = 25.80$, $\eta_p^2 = .11$, C.I. = [.00, .25] and also an interaction between block and CS, $F(2, 58) = 3.29$ $p = .044$, $MSe = 13.42$, $\eta_p^2 = .10$, C.I. = [.00, .24]; nothing else was significant, largest $F(1,29) = 4.11$ $p = .052$, $MSe = 19.16$. The three-way interaction was explored by conducting separate ANOVAs on the data from each trial block with congruency and trial type as factors. This revealed a significant interaction in block 1, $F(1, 29) = 7.97$ $p = .009$, $MSe = 29.62$, $\eta_p^2 = .22$, C.I. = [.04, .37] but nothing was significant in blocks 2 or 3, largest $F(1, 29) = 1.03$ $p = .32$, $MSe = 11.79$. This effect on block 1 reflected a transiently low rate of preCS responding on incongruent trials (especially for GC). It is, however, not likely that these transient preCS differences were driving the effect in the corrected scores: if they had been, one would expect to see an interaction of the critical congruency x trial type interaction with block. We did not: the corrected scores revealed a smaller PIT on X trials across all three test blocks.

Table 4 about here

Discussion

As in Experiment 1, X reduced the magnitude of the specific PIT effect seen with F. Thus, despite adapting the procedure to foster formation of S-R associations, X was still able to abolish PIT, attesting to the generality of this effect. This is consistent with the S-O-R account, according to which specific PIT depends on S activating O. To the extent that a conditioned inhibitor can prevent this, the specific PIT effect should be reduced.

Once again we provided independent confirmation of X's conditioned inhibitory properties in the summation test. Moreover, this test again indicated that X reduced the ability of G to predict O_2 as effectively as that of F to predict O_1, suggesting its effects were not outcome-specific. In this experiment this result was also supported by the results of the PIT test: the specific PIT supported by G was also numerically reduced by X, and there was no evidence that the effects of X on G and F differed from the Bayesian analysis. This is inconsistent with the results reported by Laurent et al. (2014), who found that the conditioned inhibitors did not reduce specific PIT when compounded with CSs that had signalled the alternative outcome. There were many differences in procedure between the two studies, however, that could account for this discrepancy; Laurent et al. (2014) employed rats, real outcomes, and trained two inhibitors, one with each outcome. In contrast we employed human participants, images of outcomes, and trained only an inhibitor for only one of the outcomes. Any of these factors could have been responsible for the difference in results. The failure of the inhibitor to have outcome-specific effects in the PIT test is also paradoxical in terms of an S-O-R interpretation of the specific PIT effect; this will be taken up below.

Experiments 3a and 3b

We have argued that the conditioned inhibitor X reduced specific PIT by suppressing activation of the outcome representation. But there are other potential explanations. For

example, the inhibitory training might have resulted in X becoming associated with a competing response which interfered with performance of R_1 and R_2 during the PIT test. Alternatively it might have resulted in X becoming associated with some other stimulus: for example, during Pavlovian training participants were presented with the various CS compounds and had to predict which outcome would follow, by selecting the options of *'Food'* (or *'Drink'*) or *'Nothing'*. Thus on A trials they had to select *'Food'* (or *'Drink'*), and on AX trials *'Nothing'*. If they learned that X predicted the word *'Nothing'*, this representation could be evoked during the PIT test and interfere with the ability of F to elicit the response - whether directly through an S-R association, or indirectly via activation of the O_1 representation. Neither explanation requires the assumption that X is a conditioned inhibitor - but both must assume that X is more able to form these associations than C, and it is not entirely clear why this should be the case; for example, participants also had to select the word *'Nothing'* on CH and C trials as well as on AX and X trials. Nonetheless, the purpose of the final two experiments was to evaluate these alternative possibilities experimentally.

Experiment 3a

This was a variant of Experiment 1, except that in the PIT test participants were tested with X and C in the *absence of any expectation of the outcome*. If the attenuation of specific PIT in the previous studies was due to X eliciting a competing response, then it should suppress instrumental responding more than C regardless of whether or not the outcome is anticipated. But if X's effect on PIT was mediated by its effect on the outcome representation, no such effect should be observed. To achieve this, no outcomes were delivered during instrumental training, and participants were simply told to perform R_1 and R_2 to increment the score on the counter that had registered numbers of outcomes delivered in the previous studies; second, F and G were omitted from the PIT test, participants being tested with X and C alone. Finally, X and C were presented alone during inhibitory training, just as in Experiment 2.

Method

Participants: 32 students from the University of Nottingham aged between 18-25 years (8 males and 24 females) participated in this study; one was excluded after application of the performance criterion.

Procedure

Identical to Experiment 1, except: (i) in instrumental training the outcomes were omitted, and the following instructions were given: "*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!*" (ii) the PIT test comprised two test block comprised two presentations of X and two of C in a semi-random order (iii) inhibitory training included nonreinforced presentations of C and X, exactly as in Experiment 2 (iv) the counterbalancing was as in Experiment 2.

Data Treatment Identical that that of Experiment 1 unless otherwise stated.

Results

Pavlovian phase: Mean ratings for A were 0.77, 0.87, 0.87, 0.90, 0.91 and 0.94, and for AB 0.74, 0.84, 0.97, 0.86, 0.87 and 0.92. Mean ratings across inhibitory training were 0.80, 0.83, 0.89 and 0.97 for AX, and 0.95, 0.96, 0.99 and 0.99 for CH. ANOVA performed on data from inhibitory training, with trial type and trial block as factors, revealed significant effects of stimulus, $F(3, 90) = 6.00$, $p < .001$, $MSe = .036$, $\eta_p^2 = .17$, C.I. = [.03, .28] and of block, $F(3, 90) = 4.01$, $p < .01$, $MSe = .028$, $\eta_p^2 = .12$, C.I. = [.01, .23]; the interaction was not significant, $F(9, 270) = 1.86$, $p = .06$, $MSe = .034$. Tukey's tests to explore the effect of stimulus revealed lower accuracy on AX than on both CH and A trials, and also lower accuracy on A than on CH trials. Mean ratings for F were 0.71, 0.85, 0.82 and 0.82, and for G 0.92, 0.90, 0.89 and 0.95; ANOVA revealed a significant main effect of CS $F(1, 30) =$

13.38, $p < .001$, $MSe = .059$, $\eta_p^2 = .31$, C.I. = [.06, .51] - subjects were less accurate with F than G; the effect of block and the interaction were not significant, largest $F(3, 90) = 1.36$, $p = .26$, $MSe = .048$.

Summation test: In contrast to the previous experiments, the data from the summation test were not significant: the mean ratings for FC, FX, GC and GX were 0.4, 0.37, 0.46 and 0.45 respectively, and ANOVA with CS (F,G) and trial type (C, X) as factors revealed only a significant effect of CS, $F(1, 30) = 14.45$, $p = 0.001$, $MSe = 0.014$, $\eta_p^2 = .33$, C.I. = [.07, .53]; nothing else was significant, largest $F(1, 30) = 1.54$, $p = 0.22$, $MSe = 0.014$. It is not clear why performance was so poor, but as we wanted to establish whether an *inhibitory* X was able to elicit a competing response more than the control cue C, it was critical that a significant CI effect was obtained. Thus all participants rating US occurrence as more likely on FX than on FC trials, *and* on GX trials than on GC trials, were excluded. This led to the exclusion of five more participants; the resultant data are presented in Figure 4 (upper panel). ANOVA with CS (F,G) and trial type (C, X) as factors revealed a significant effect of trial type, $F(1, 25) = 5.22$, $p = 0.031$, $MSe = 0.013$, $\eta_p^2 = .17$, C.I. = [.00, .41]; there was also a significant effect of CS, $F(1, 25) = 6.90$, $p = 0.014$, $MSe = 0.013$, $\eta_p^2 = .17$, C.I. = [.01, .45]; the interaction was not significant, $F<1$, and the associated Bayes factor was 4.69. Thus, just as in Experiments 1 and 2, the inhibitor X was no more effective with F than with G.

PIT test: Corrected response rates during C and X averaged over the two trial blocks are presented in Figure 4 (lower panel); all scores are averaged across R_1 and R_2. Neither stimulus appeared to have any effect on conditioned responding, and there was no evidence that X differed from C in this respect; ANOVA with CS (C, X) and block as factors revealed only a significant effect of block, $F(1, 25) = 5.35$, $p = 0.03$, $MSe = 11.36$, $\eta_p^2 = .18$, C.I. = [.00, .41]; nothing else was significant, Fs < 1. The Bayes factor associated with the difference in corrected scores for C and X trials was 4.81, allowing us to accept the null hypothesis of no difference between X and C. The mean rates of preCS responding were,

for X and C respectively, 133.6 and 141.4 for block 1, and 155.2 and 159.2 for block 2. ANOVA with CS and block as factors revealed nothing significant, largest $F(1, 25) = 2.7$, $p = 0.11$, $MSe = 16.71$ for the effect of block; neither the effect of CS nor the interaction were significant, Fs < 1. (Parallel analyses performed on the data *before* application of the additional performance criterion produced the same results.)

Figure 4 about here

Discussion

These results provided no support for the suggestion that X elicited a competing response, allowing it to interfere with performance of R_1 and R_2 more than C. This implies that the results of Experiments 1 and 2 cannot be attributed to such a mechanism.

Experiment 3b

Experiment 3b employed the same Pavlovian training procedure as Experiment 3a, except that in inhibitory training AX presentations were replaced by trials in which X was presented with a novel stimulus, D, and both DX and X presentations were followed by the word *'Nothing'* presented on the screen, overlaid on the white square that had previously accompanied no outcome presentations (see Table 5). CH and C were followed by the white square alone, as in previous experiments. This should encourage selective formation of an association between X and the word *'Nothing'*. Then a PIT test was given, as in Experiments 1 and 2. If X's ability to evoke an alternative outcome competed with the tendency of F and G to elicit their respective responses, and this was responsible for X's selective ability to reduce the PIT effect, then the same effect should be observed here. The summation test was omitted in the present experiment, as there was no conditioned inhibition to assess.

Table 5 about here

Method

Participants: 32 participants from the University of Nottingham aged 18-29 years old (8 males and 24 females) took part. Four failed to achieve the performance criterion (here applied to DX rather than AX trials) and were excluded, leaving 28 participants.

Procedure

Identical to Experiment 1, except (i) during Pavlovian training AX trials were replaced by trials in which X was accompanied by a novel stimulus D; X and C were also presented alone, exactly as in Experiment 2, and on X and XD trials the white square that was presented on no outcome trials had the word *'Nothing'* presented in the middle of it in black font, occupying approximately 20% of its total area (ii) the summation test was omitted.

Data Treatment Identical to that of Experiment 1 unless otherwise stated.

Results

Pavlovian phase: The mean ratings for A were 0.79, 0.88, 0.95, 0.98, 0.98 and 0.93, for AB 0.79, 0.88, 0.93, 0.96, 0.95 and 0.91, for DX 098, 0.95, 0.98 and 0.99, and for CH 0.99, 0.98, 0.96 and 0.98. ANOVA with trial type (A, AB, DX, CH) and trial block as factors revealed nothing significant, largest $F(3, 81) = 1.92$, $p = 0.13$, $MSe = .018$. Mean ratings for F were 0.79, 0.88, 0.77 and 0.80, and for G 0.95, 0.86, 0.89 and 0.88, and ANOVA revealed a main effect of CS $F(1, 27) = 10.21$, $p < .004$, $MSe = .039$, $\eta_p^2 = .27$, C.I. = [.04, .49] indicating lower accuracy with F; nothing else was significant, largest $F(3, 81) = 1.47$, $p = .23$, $MSe = .057$.

PIT test: Figure 5 reveals a robust PIT effect of broadly similar magnitude for all compounds, albeit slightly smaller for FX (rates of congruent and incongruent responding for each CS compound by block are shown in Table 2); but ANOVA with congruency, CS (F, G) trial type (C, X) and block as factors revealed only a significant effect of congruency, $F(1, 27) = 13.25$, $p = 0.001$, $MSe = 67.1$, $\eta_p^2 = .33$, C.I. = [.07, .54]; nothing else was significant, largest $F(1, 27) = 2.70$, $p = 0.112$, $MSe = 12.94$. Importantly, neither the Congruency x Trial interaction

nor the Congruency x Trial x CS interaction were significant (*Fs* < 1, Bayes factors 4.98 and 3.89 respectively); thus there was no evidence that the specific PIT effects during X and C differed. Mean rates of preCS responding were, for FX, FC, GX and GC respectively, 53.30, 51.16, 51.16 and 49.02 rpm respectively (see supplementary materials); ANOVA revealed nothing significant, largest $F(1, 27) = 1.81$, $p = 0.19$, $MSe = 7.11$.

Figure 5 about here

Discussion

The results did not support the view that the ability of X to reduce the magnitude of specific PIT was caused by X predicting an alternative outcome. In the present study we tried to ensure formation of an association between X and an alternative outcome by explicitly pairing X, but not C, with the word *'Nothing'*. According to the argument that is being tested, this should result in a selective suppression of PIT by X that is, if anything, even greater than that seen in Experiments 1 and 2. This was not observed: PIT was of equal magnitude for both C and X. We must therefore conclude that the reduction of PIT observed in Experiments 1 and 2 was due to X's properties as a CI.

General Discussion

These experiments aimed to throw further light on whether the specific PIT effect may be attributed to an S-O-R mechanism in humans, by using a conditioned inhibition paradigm. In Experiments 1 and 2 we trained two Pavlovian CSs, F and G, and two instrumental responses, R_1 and R_2, each associated with one of two outcomes, O_1 and O_2 respectively. When F and G were each presented in compound with a preexposed control stimulus, specific PIT was observed - higher levels of R_1 during F, and of R_2 during G; but when F was presented in compound with X, which had been trained as a conditioned inhibitor signalling the absence of O_1, specific PIT was abolished. We interpreted this in terms of a specific version of the S-O-R theory, according to which specific PIT is produced because the CS activates the outcome representation, which in turn elicits the response that

produced that outcome. If X's conditioned inhibitory properties suppress activation of O, this should dampen the effectiveness of this associative chain, and reduce specific PIT. The results are less consistent with the S-R account (Cohen-Hatton et al. 2013), according to which O, although mediating formation of the S-R association that underlies specific PIT, plays a role *only* during training; thus the presence of a CI at test should not influence expression of specific PIT. Given that evidence has been generated in terms of both explanations of specific PIT it is likely to be multiply determined, which could cast doubt on the generality of our findings. Nonetheless in Experiment 2, in which we attempted to maximise the formation of S-R associations, X was still able to eliminate specific PIT.

The results of the summation tests provided independent confirmation that X was a conditioned inhibitor. In both experiments X passed the summation test: when presented with a CS that had signalled an outcome, participants were less likely to expect that outcome in the presence of the CI, compared to when a neutral control stimulus was present. This provides independent evidence that X could reduce the expectation of the outcome produced by the CS. In Experiment 3 we also evaluated two alternative explanations of X's effects on PIT - that they were mediated by X's selective ability to elicit a competing response (Experiment 3a), or to evoke the representation of an alternative outcome (Experiment 3b). Neither of these explanations was supported. In sum our findings suggest that suppression of the outcome representation could underlie the effect of the inhibitor on the specific PIT effect (see Rescorla & Holland, 1977).

We have confined our attention to one version of the S-O-R account, according to which the ability of O to elicit R relies on the backward operation of the R-->O association. But an alternative is that during instrumental training, expectation of the sensory properties of the outcome produced by the response becomes part of the stimulus complex that elicits it (Trapold & Overmier, 1972): thus the ability of O to elicit R relies on an O-->R, rather than an R-->O, association (Gilroy, Everett & Delamater, 2014). However, it is unlikely that such a mechanism could operate in these experiments, as the responses were trained concurrently

- anticipation of O_1 would have been as likely to precede R_1 as of R_2, resulting in both O_1-->R_1 *and* O_2-->R_1 associations. Thus F's ability to activate O_1 could not produce the specific PIT we observed. Nonetheless, evidence in support of this mechanism has been reported, and it is likely both play a role in instrumental performance (Balleine & Ostlund, 2007).

A paradoxical aspect of our results is the fact that the effects of the inhibitor X did not appear to be outcome-specific. In the summation test X, signalling the absence of O_1, was also able to act as a conditioned inhibitor with a CS that signalled O_2, and in Experiment 2 X also appeared to reduce the PIT effect otherwise produced by a signal for O_2. This seems inconsistent with the S-O-R account of specific PIT, according to which it must depend on the sensory properties that differentiate the two outcomes: CS_1 can only activate R_1 more than R_2 if it can selectively activate the sensory components of O_1 that are absent in O_2 - activating the motivational properties that O_1 and O_2 share would result in performance of both responses. So if the inhibitor X can only suppress activation of the features *common* to both outcomes, both responses should be equally suppressed - and specific PIT should remain intact. Moreover, applying the principles of associative theory to this interpretation of the outcome representations raises further issues. As the motivational and sensory components of each outcome co-occur, associative principles dictate that they should become associated. Thus activation of O_1 will activate the sensory aspects of O_2 via the motivational component they share; but if presenting O_1 activates sensory components of O_2, how can specific PIT occur? If CS_1 activates O_1 and hence O_2, both R_1 and R_2 should be elicited, and produce a general, rather than a selective, elevation of instrumental responding.

The associative analysis may also be able to solve these paradoxes. Conceptualising the two outcomes as comprising common motivational (X) and unique sensory properties (A and B) allows them to be represented as the stimulus compounds AX and BX (Figure 6 panel a). (Our outcomes did not have a strong motivational component, but would have some components in common and others that differentiated them, which is all that is critical for this argument.) Exposure to the two outcomes thus means that A and X will become

associated, as will B and X. But during training there is considerable intermixed exposure with the two outcomes - which we know from studies on perceptual learning results in inhibitory links between the compounds' unique elements A and B (Dwyer & Mackintosh, 2002), such that A will inhibit activation of B and vice versa (Figure 6 panel b). This *mutual inhibition* would ensure that presentation of O_1 will not be able to activate the unique sensory components of O_2 or vice versa. This allows PIT to occur: by virtue of these inhibitory links CS_1 can activate the sensory properties of O_1, but *not* those of O_2. So although CS_1 can activate the motivational aspects of both outcomes, eliciting both responses, it can only activate the sensory properties of O_1, so R_1 will predominate (Figure 6 panel c). Moreover, even if the inhibitor X can only directly suppress activation of the motivational component of O_1, this will also reduce activation of the sensory component of O_1 indirectly, via the associative link between them (Figure 6 panel d). Equally when CS_2 is presented, which activates the common motivational properties and unique sensory properties of O_2, by directly suppressing the former, X will be able to indirectly suppress the latter - again reducing specific PIT. Such an analysis, by incorporating principles of perceptual learning into existing conceptualisations of the associative mechanisms underlying specific PIT (e.g. Balleine & Ostlund, 2007), could allow such an associative model to explain how an inhibitor could reduce the PIT effect even though its inhibitory properties are not outcome-specific.

Figure 6 about here

We believe this study is the first attempt to examine the effects of a conditioned inhibitor on specific PIT in human participants (although see Colagiuri & Lovibond, 2015 for related findings in a general transfer task). The technique proved to be useful in evaluating the S-O-R account of PIT; but it could also have clinical relevance. Pavlovian cues influence our behaviour in many ways - such as the role of Pavlovian cues associated with drugs in craving and relapse (Conklin & Tiffany, 2002; LeBlanc, Ostlund & Maidment, 2012). The fact that specific PIT seems insensitive to outcome devaluation (Holland, 2004) and S-O extinction (Delamater, 1996) presents a challenge for therapy. But if a cue trained as a

conditioned inhibitor can abolish, or at least reduce, the effect of a CS on behaviour, it could help in treating disorders such as drug addiction. Moreover, if our results may be taken as evidence that a CI signalling the absence of one outcome may reduce the specific PIT produced by a CS that signalled another, then perhaps a cue trained as a conditioned inhibitor, even with a consequence different to the drug, could help to reduce the excitatory effects of drug-related cues on addictive behaviour.

References

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*, 402-408.

Balleine, B. W., & Ostlund, S. B. (2007). Still at the Choice‐Point. *Annals of the New York Academy of Sciences*, *1104*, 147-171.

Bindra, D. (1968). Neuropsychological interpretation of the effects of drive and incentive-motivation on general activity and instrumental behavior. *Psychological Review*, *75*, 1-22.

Bouton, M.E. (1993). Context, time and memory retrieval inthe interference paradigms of Pavlovian conditioning. *Psychological Bulletin*, *114*, 80-99.

Brown, P.L., & Jenkins, H.M. (1968). Auto-shaping of the pigeon's keypeck. *Journal of the Experimental Analysis of Behavior*, *11*, 1-8.

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*, 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.

Conklin, C. A., & Tiffany, S. T. (2002). Applying extinction research and theory to cue‐exposure addiction treatments. *Addiction*, *97*, 155-167.

Corbit, L. H., & Janak, P. H. (2007). Ethanol‐Associated Cues Produce General Pavlovian‐Instrumental Transfer. *Alcoholism: Clinical and Experimental Research*, *31*, 766-774.

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

Delamater, A.R., Lolordo, V.M, & Sosa, W. (2003). Outcome-specific conditioned inhibition in Pavlovian backward conditioning. *Learning and Behavior*, *31*, 393-402.

Dwyer, D.M., & Mackintosh, N.J. (2002). Alternating exposure to two compound flavours creates inhibitory associations between their unique features. *Animal Learning and Behavior*, *30*, 201-207.

Fantino, E. (1977). Conditioned reinforcement: Choice and information. In W.K. Honig & J.E.R Staddon (Eds.) *Handbook of Operant Behavior*. Prentice-Hall, Englewood Cliffs, N.J.

Gilroy, K.E., Everett, E.M., & Delamater, A.R. (2014). Response-outcome versus outcome-response associations in Pavlovian-to-instrumental transfer: Effects of instrumental training context. *International Journal of Comparative Psychology*, *27*, 585-597.

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 and Drugs*, *66*, 53-61.

He, Z., Cassaday, H. J., Howard, R. C., Khalifa, N., & Bonardi, C. (2011). Impaired Pavlovian conditioned inhibition in offenders with personality disorders. *Quarterly Journal of Experimental Psychology*, *64*, 2334-2351.

He, Z., Cassaday, H.J., Park, S.B.G, & Bonardi, C. (2012) When to hold that thought: Reduced inhibition of pre-potent associations in schizophrenia *Plosone, 7,* 1-9.

Hogarth, L., Dickinson, A., Wright, A., Kouvaraki, M., & Duka, T. (2007). The role of drug expectancy in the control of human drug seeking. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*, 484.

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*, 261.

Hogarth, L., Retzler, C., Munafò, M.R., Tran, D.M.D., Troisi II, J.R., Rose, A.K., Jones, A., & Field. M. (2014). Extinction of cue-evoked drug-seeking relies in degrading hierarchical instrumental expectancies. *Behaviour Research and Therapy*, *59*, 61-70.

Holland, P. C. (2004). Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*, 104-117.

Holmes, N.M., Marchand, A.R., & Coutureau, E. (2010). Pavlovian to instrumental transfer: A neurobehavioural perspective. *Neuroscience and Biobehavioral Reviews*, *34*, 1277-1295.

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

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

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*, 165-181.

Laurent, V., Wong, F.L., & Balleine, B.W. (2014) δ - Opioid receptors in the accumbens shell mediate the influence of both excitatory and inhibitory predictions on choice. *British Journal of Pharmacology*, *172*, 562-570.

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

LoLordo, V.M. (1967). Similarity of conditioned fear responses based on different aversive events. *Journal of Comparative and Physiological Psychology*, *64*, 154-158.

Mackintosh, N.J., & Dickinson, A. (1979). Instrumental (Type II) conditioning. In A. Dickinson & R.A. Boakes (Eds.) *Mechanisms of learning and motivation.* PP.143-167. Hillsdale, N.J.: Erlbaum.

Nieto, J. (1984). Transfer of conditioned inhibition across different aversive reinforcers in the rat. Learning and Motivation, 15, 37-57.

Pearce, J.M., Montgomery, A., & Duckinson, A. (1981). Contralateral transfer of inhibitory and excitatory eyelid conditioning in the rabbit. *Quarterly Journal of Experimental Psychology*, *33*, 45-61.

Peirce, J.W. (2007). Psychopy - Psychophysics software in Python. J*ournal of Neuroscience Methods*, *162*, 8-13.

Rescorla, R. A. (1969). Pavlovian conditioned inhibition. *Psychological Bulletin*, *72*, 77-94.

Rescorla, R. A., & Holland, P. C. (1977). Associations in Pavlovian conditioned inhibition. *Learning and Motivation*, *8*, 429-447.

Robbins, T. W., & Everitt, B. J. (1999). Drug addiction: Bad habits add up. *Nature*, *398*, 567-570.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D. & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychononic Bulletin & Review*, *16*, 225-237.

Sanchis‐Segura, C., & Spanagel, R. (2006). Behavioural assessment of drug reinforcement and addictive features in rodents: An overview. *Addiction Biology*, *11*, 2-38.

Saunders, B.T., & Robinson, T.E. (2013). Individual variation in resisting temptation: Implications for addiction. *Neuroscience and Biobehavioral Reviews*, *37*, 1955-1975.

Stewart, J., de Wit, H., & Eikelboom, R. (1984). Role of unconditioned and conditioned drug effects in the self-administration of opiates and stimulants. *Psychological Review*, *91*, 251-268.

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

Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779-804.

Wagner, A.R., & Brandon, S.E. (1989). Evolution of a structured connectionist model of Pavlovian conditioning (ÆSOP). In S.B. Klein and R.R. Mowrer (Eds.), *Contemporary learning theories: Pavlovian conditioning and the status of traditional learning theories*. (pp. 149-189). Hillsdale, NJ: Erlbaum.

Williams, D.A., Travis, G.M. & Overmier, J.B. (1986). Within-compound associations modulate the relative effectiveness of differential and Pavlovian conditioning inhibition procedures. *Journal of Experimental Psychology: Animal Behavior Processes*, *12*, 351-362.

Table 1. Design of Experiment 1.

Pavlovian stage			Summation test	Instrumental training	PIT test
A->O_1	A->O_1	F->O_1	FX		FX (R_1/R_2)
AB->O_1	AB->O_1	G->O_2	FC	R_1->O_1	FC (R_1/R_2)
	AX-		GX	R_2->O_2	GX (R_1/R_2)
	CH-		GH		GH (R_1/R_2)

Note: A, B, C, H, X, F and G refer to fractal images, R_1 and R_2 to keyboard responses, and O_1 and O_2 to food and drink images; - denotes no outcome. For further details see text.

Table 2. Mean CS-preCS response rates in each block of the PIT tests of Experiment 1 (top), 2 (centre) and 3b (bottom). Scores are presented separately for each test compound, for congruent / incongruent responses.

Experiment/Block	FX	FC	GX	GH
1 / 1	-1.36 / 23.2	27.2 / 5.5	16.4/ 4.8	15.7 / 5.5
1 / 2	15.7 / 10.9	55.9 / -4.8	24.6 / 5.5	39.6 / 4.1
2 / 1	-18.0 / -18.5	8.5 / -41.5	-6.5 / -17.5	26.0 / -44.0
2 / 2	0.0 / -20.0	21.0 / -34.0	5.0 / -22.5	7.5 / -20.0
2 / 3	16.0 / -6.0	16.0 / -17.5	15.5 / -31.0	14.5 / -16.5
3b / 1	24.1 / -14.5	46.1 / 12.9	37.5 / -4.3	33.8 / 4.8
3b / 2	41.3 / 10.2	46.6 / -3.8	53.6 / -3.2	61.1 / 3.8

Table 3. Design of Experiment 2.

Instrumental training	**Pavlovian stage**			**Summation test**	**PIT test**
R1->O_1	F->O_1	A->O_1	A->O_1	FX	FX (R_1/R_2)
R2->O_2	G->O_2	AB->O_1	AB->O_1	FC	FC (R_1/R_2)
			AX- X-	GX	GX (R_1/R_2)
			CH- C-	GC	GC (R_1/R_2)

Note: A, B, C, H, X, F and G refer to fractal images, R_1 and R_2 to keyboard responses, and O_1 and O_2 to food and drink images; - denotes no outcome. For further details see text.

Table 4: Mean preCS response rates in each block of the PIT test of Experiment 2. Scores are presented separately for each trial type, for congruent / incongruent responses.

Block	FX	FC	GX	GC
1	70.0 / 59.0	57.5 / 77.0	64.0 / 41.0	24.5 / 90.0
2	62.5 / 62.0	52.5 / 79.0	76.5 / 57.0	75.0 / 55.5
3	58.5 / 41.0	44.5 / 49.5	57.5 / 73.0	78.0 / 44.0

Table 5: Design of Experiment 3b

Pavlovian stage			Instrumental training	PIT test
A->O_1	A->O_1	F->O_1		FX (R_1/R_2)
AB->O_1	AB->O_1	G->O_2	R_1->O_1	FC (R_1/R_2)
	DX-> 'Nothing'		R_2->O_2	GX (R_1/R_2)
	X-> 'Nothing' CH- C-			GC (R_1/R_2)

Note: A, B, C, D, H, X, F and G refer to fractal images, R_1 and R_2 to keyboard responses, and O_1 and O_2 to food and drink images; - denotes no outcome, and "Nothing" denotes presentation of the word 'Nothing' on the screen. For further details see text.

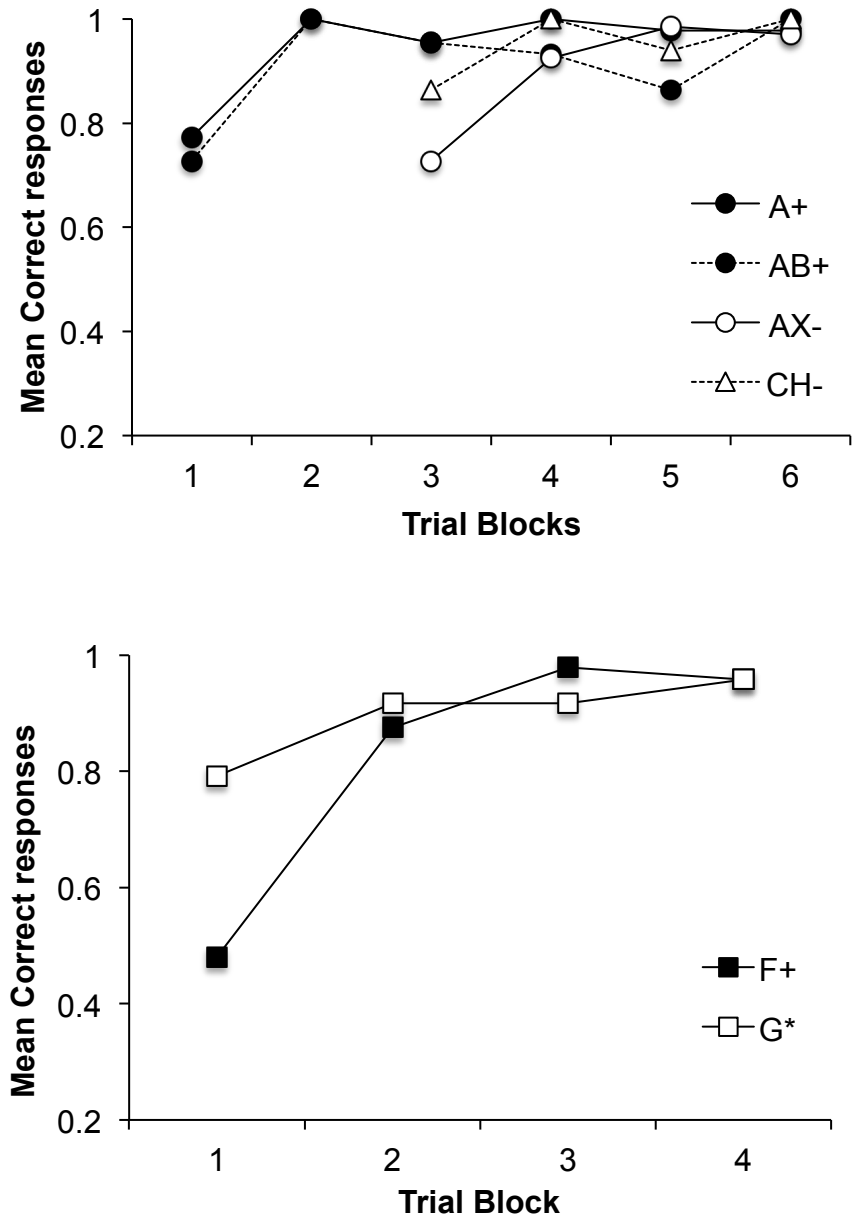

Figure 1. Mean number of correct responses per trial block for each trial type in the pretraining and inhibitory training phases (upper panel) and test excitor training phases (lower panel) of Experiment 1.

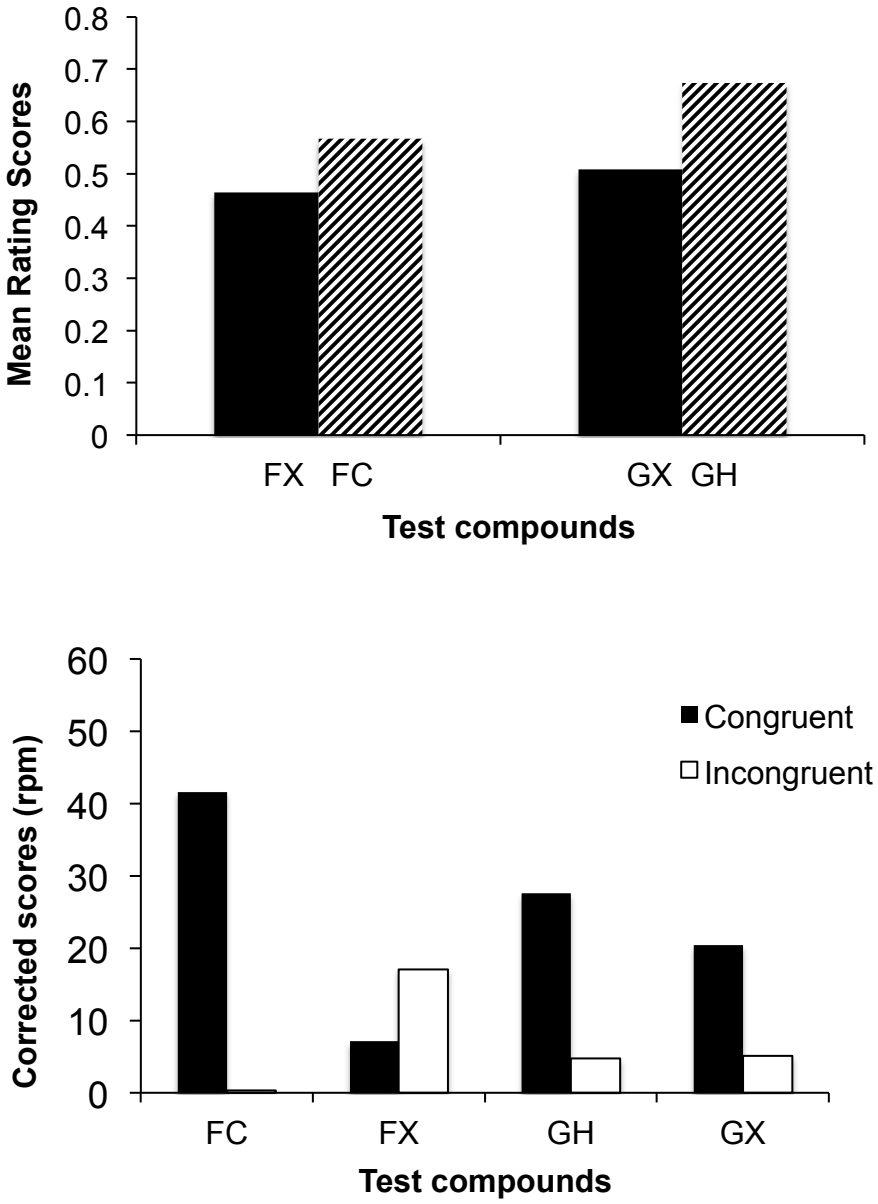

Figure 2. Upper panel: Mean ratings of the likelihood of O_1 occurrence during FX and FC, and O_2 occurrence during GX and GH, in the summation test of Experiment 1. Lower panel: Mean rates of congruent and incongruent responding for F and G in compound with the inhibitor X or the control stimulus, C or H, averaged over all blocks of the PIT test of Experiment 1.

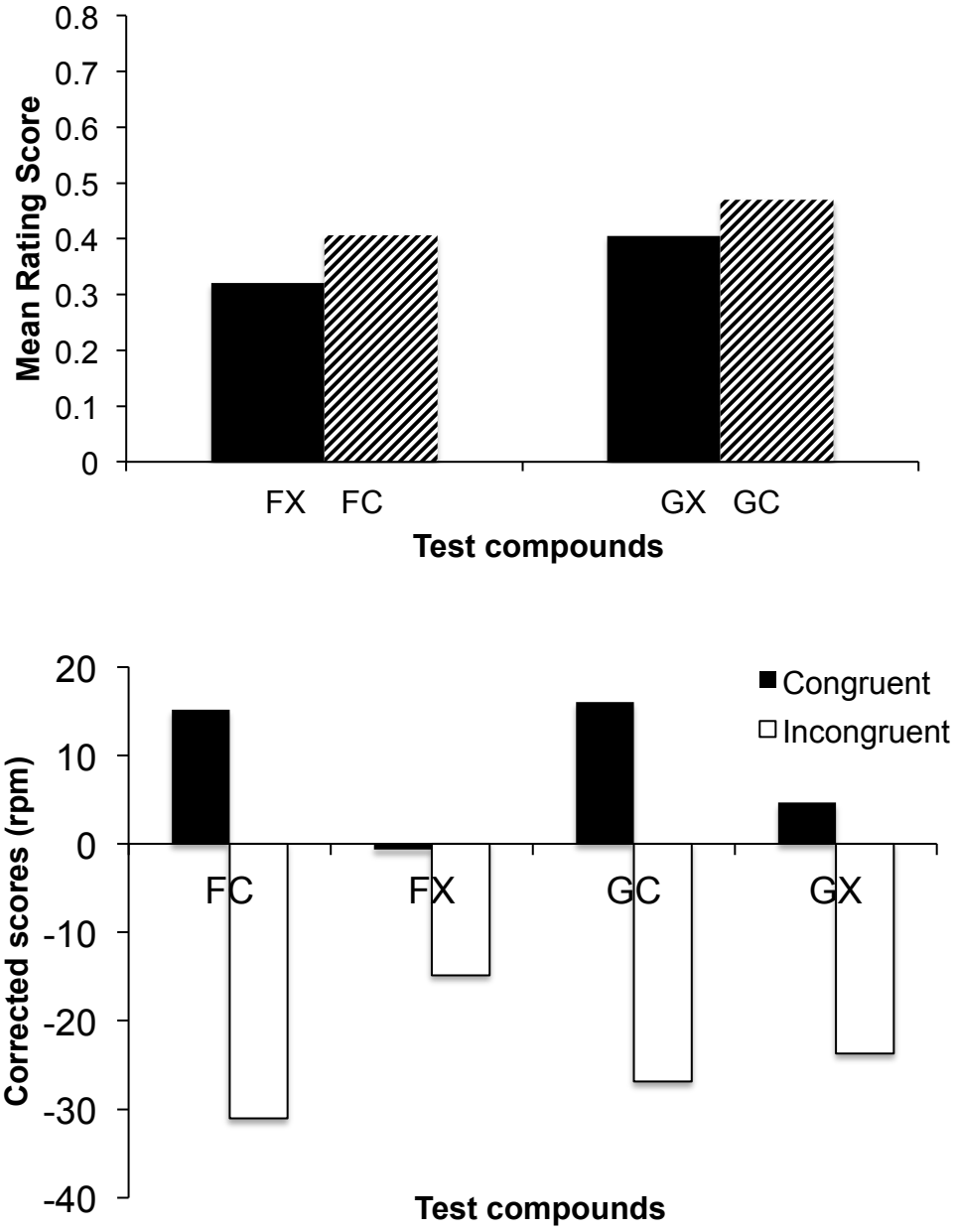

Figure 3. Upper panel: Mean ratings of the likelihood of O_1 occurrence during FX and FC, and O_2 occurrence during GX and GH, in the summation test of Experiment 2. Lower panel: Mean rates of congruent and incongruent responding for F and G in compound with the inhibitor X or the control stimulus, C, averaged over all blocks of the PIT test of Experiment 2.

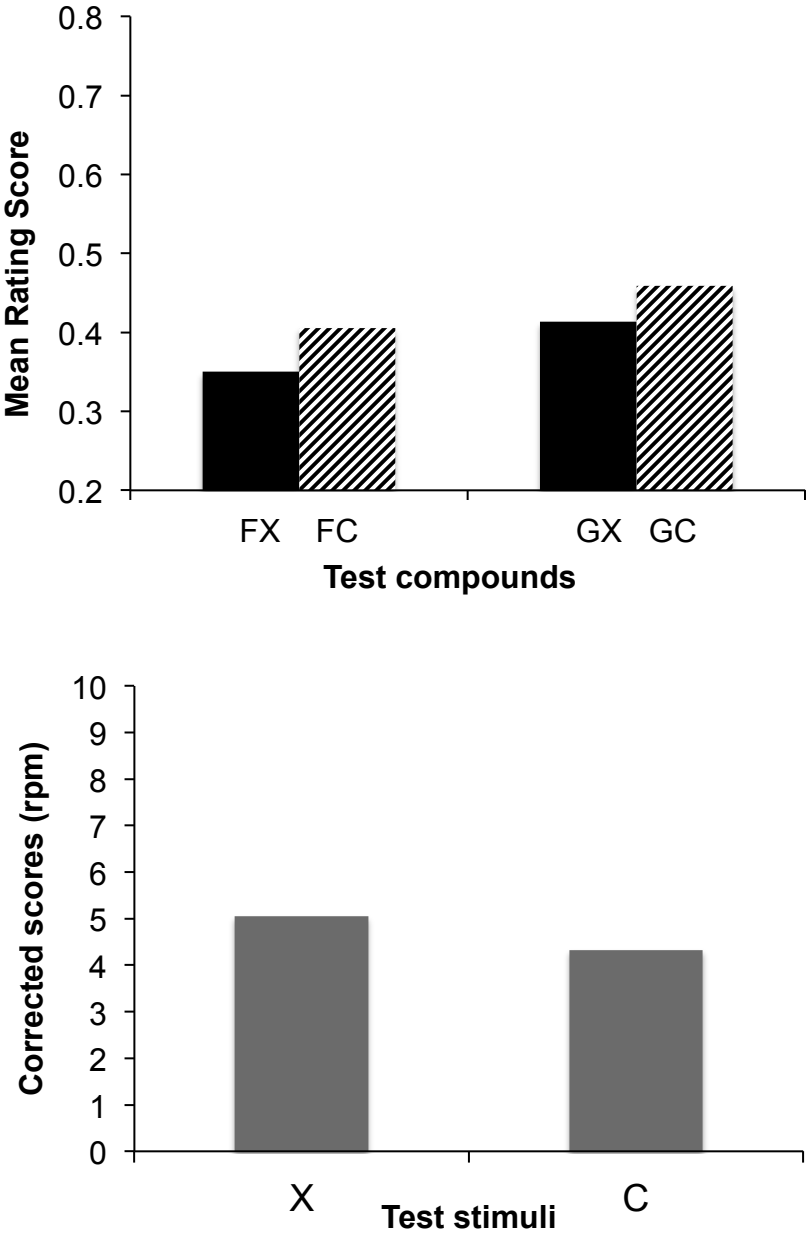

Figure 4. Upper panel: Mean ratings of the likelihood of O_1 occurrence during FX and FC, and O_2 occurrence during GX and GH, in the summation test of Experiment 3a. Lower panel: Mean corrected rates of responding during C and X averaged over all test blocks of Experiment 3a.

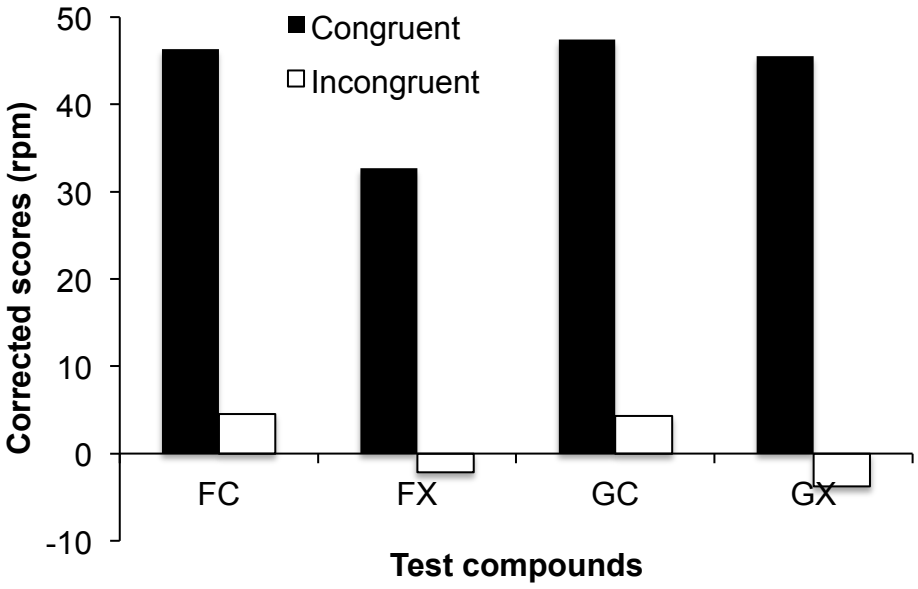

Figure 5. Group mean corrected scores on congruent and incongruent trials for F and G in compound with X, paired with the word 'Nothing', or the control stimulus C, pooled over all test blocks of the PIT test of Experiment 3b.

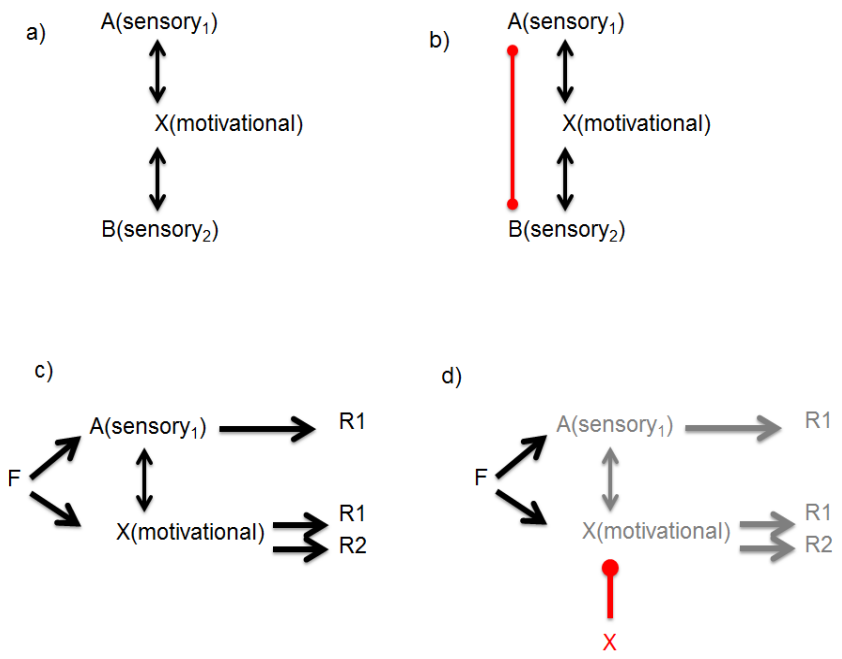

Figure 6. Panel a: Proposed associative structure of two outcomes, O_1 and O_2, with unique sensory properties (A and B respectively) and common motivational properties (cf. Balleine & Ostlund, 2007). Panel b: Proposed inhibitory link resulting from intermixed preexposure to O_1 and O_2. Panel c: Resultant pattern of activation when signal of O_1 is presented Panel d: Effect of a conditioned inhibitor on activation produced by signal for O_1.