Running Head: DELAYED REWARDS AND HABITS

Delayed Rewards Facilitate Habit Formation

Gonzalo P. Urcelay[1] & Sietse Jonkman[2]

[1] Department of Neuroscience, Psychology & Behaviour, University of Leicester, UK.

[2] Sarepta Therapeutics, Massachusetts, USA.

Corr. author: Gonzalo P. Urcelay
Department of Neuroscience, Psychology and Behaviour.
University of Leicester
Lancaster Rd. Leicester, LE1 7HA, UK
TEL: +44 (0116) 229-7173
E-mail: gpu1@le.ac.uk

Abstract

The transition from goal-directed to habitual forms of instrumental behaviour is determined by variables such as the amount of training, schedules of reinforcement, the availability of choices, and exposure to drugs of abuse. Less is known about the control of goal-directed behaviour when reinforcement is delayed rather than immediate. In these experiments, we investigated in rats the role of response-outcome contiguity on the control of goal-directed action, assessed through satiety-specific outcome devaluation tests. In Experiment 1 using a within-subjects design we observed goal-directed behaviour after six days of FR1 training when the outcome was presented immediately following the lever press, but not when it was delayed for 20 s, revealing habit formation with delayed outcomes. Experiment 2 revealed that the habitual control observed with 20 s delays of reinforcement can be prevented if, immediately before each instrumental training session, the rats were exposed to the experimental context in the absence of both the lever and reinforcement. In summary, these experiments suggest that response-outcome contiguity plays an important role in the control of goal-directed actions and habits.

Key Words: Instrumental learning; habits; delayed reward; context; goal-directed.

Delayed Rewards Facilitate Habit formation

Learning the relationship between actions and their consequences is of adaptive value because it allows animals to operate in their environment to satisfy their needs. Instrumental learning was described early on in the studies by Thorndike, in which hungry cats learned to press a lever in order to escape a puzzle box and access food (Thorndike, 1911). It was on the basis of these experiments that Thorndike proposed the law of effect, which states that a positive reinforcer (i.e., food) strengthens the connection between environmental stimuli (i.e., the box) and the responses made immediately before the reinforcer was presented, so that in subsequent encounters, the stimulus elicits the response. In the last decades however, it has become increasingly clear that instrumental behaviour can be controlled by the current value of its consequences and not just their value during instrumental training. In an experiment using hungry rats, Adams and Dickinson (1981) trained the animals to press a lever using two types of pellets. One pellet was contingent upon lever press whilst the other was delivered non-contingently (i.e., independent of the animal's behaviour). Following training, they devalued either the contingent or non-contingent pellets (in different groups) by pairing them with an injection of Lithium Chloride (LiCl), which produces nausea and an aversion to the pellets. During a subsequent test of lever press behaviour in extinction, the group that had the contingent pellet associated with LiCl pressed less vigorously than the group that had the non-contingent pellet associated with illness. In other words, instrumental behaviour was goal-directed. This and other findings have led to the suggestion that instrumental behaviour can be controlled by a representation of the current value of the outcome (i.e., goal-directed) or driven by environmental stimuli (i.e., habitual; Dickinson, 1985). Instrumental behaviour is goal-directed when it encodes a representation of the causal relationship between the response and the outcome, and a representation of the current value of the outcome (Dickinson, 1994).

The distinction between goal-directed versus habitual control of instrumental responding (with the use of outcome devaluation technique) has become popular and relevant to both rodent and human behaviour. It has led to research programs in both species that investigate the neural basis of instrumental behaviour (Dolan & Dayan, 2013), and to the proposal that some psychiatric disorders

such as OCD (Robbins et al., 2012; Gillan, et al. 2014) and drug addiction (Everitt & Robbins, 2005; 2015) result from aberrant habit formation. In fact, in rodent models drug exposure per-se has been observed to facilitate habit formation (Nelson & Killcross, 2006). In the psychological literature, multiple variables are known to influence whether instrumental behaviour is goal directed or habitual. Perhaps the most intuitive is the amount of instrumental training. Adams (1982) trained hungry rats to press a lever for sucrose pellets, and allowed different amounts of training in separate groups. Two groups pressed the lever until they obtained 100 sucrose pellets, whereas two additional groups pressed the lever until they obtained 500 sucrose pellets. One group from each training condition then received outcome devaluation (sucrose -> LiCl pairings) whilst the remaining group received unpaired presentations of sucrose and LiCl (a control for outcome devaluation). During test, the only group that was sensitive to devaluation of sucrose pellets was the group that obtained 100 sucrose pellets, suggesting that extended training (i.e., 500 pellets) renders the behaviour habitual.

Two additional variables can determine the observation of goal directed behaviour and habits. Dickinson and colleagues (Dickinson, Nicholas and Adams, 1983) trained two groups of rats on a ratio schedule of reinforcement (that arranges a correlation between response rates and reinforcement rates). They compared them with groups trained on interval schedules, which were yoked to the ratio animals so that the total rate of reinforcement was roughly equivalent during training in all groups. In interval schedules, the reinforcer is set up based on time rather than number of presses and this decreases the correlation between response and outcome rates. Following outcome devaluation, only those animals trained on ratio schedules showed depressed lever pressing, suggesting that for behaviour to be goal-directed, animals need to experience the fixed positive correlation between response and outcome rates, unlike what occurs with interval schedules where the availability of reinforcement is determined by time, and thus relatively independent of rate of responding. Finally, it has been shown that even with extended training, the opportunity of a choice between different levers can prevent the development of habits. This was clearly documented in an experiment by Kosaki and Dickinson (2010), in which they trained for extended periods groups of rats either with two levers and two reinforcers (group Choice), or one lever (group One Lever) plus free reinforcers to match

reinforcer rates with group Choice. Following outcome devaluation, group One Lever (but not group Choice) showed insensitivity to outcome devaluation (i.e., habits), suggesting that the choice procedure, by arranging a situation in which the animal can experience the minute-by-minute response-reinforcer correlations, rendered the behaviour goal-directed despite extended training.

The three findings described above (the amount of training, the schedule of reinforcement and the choice procedure) can be explained following the suggestion that instrumental behaviour is goal-directed when, at the time of test the animal represents both the correlation between its actions and outcomes, and the current value of the outcome. In addition to the goal-directed system, stimuli-response (S-R) associations can control instrumental behaviour, so that at the time of test environmental stimuli elicit instrumental responses (that do not encode the identity or value of the outcome). According to this view, instrumental behaviour results from the sum of these two systems, which develop in parallel as animals undergo instrumental training (Dickinson, 1985; 1994; Dickinson & Perez, 2018). This analysis suggests that another variable, that is the contiguity between response and outcome, may well influence whether behaviour is goal-directed or habitual, because when outcomes are delayed the correlation between responses and outcomes is low (Baum, 1973; see Perez 2017 for recent simulations). DeRusso and colleagues (DeRusso et al., 2010) suggested this possibility, although they did not directly manipulate the contiguity between response and outcome. Numerous experiments have shown that instrumental behaviour is sensitive to response-outcome delays. For example, Dickinson, Watt and Griffiths (1992) trained different groups of hungry rats to press a lever to obtain a sucrose pellet as outcome. For one group there was no delay in reinforcement, but other groups experienced the outcome 2, 4, or 16 seconds after they pressed the lever. They observed that rates of lever presses showed a systematic decline with longer intervals between responses and outcomes. Similarly, Shanks and Dickinson (1991) observed a decline in rates of space-bar pressing (in a computer keyboard) when they introduced delays between responses and outcomes in a task with human participants. In addition to the observation in rates of responses (behaviour), they saw a similar decline in causal ratings when human participants were required to judge the causal relationship. In other words, humans beliefs about response-outcome

causal relationship decreased with the increasing delays. Based on these observations, we hypothesized that the introduction of a delay between response and outcome will affect the experienced correlation between action and outcomes, and hence render instrumental behaviour habitual. In other words, with delayed outcomes, animals may not experience a strong correlation between responses and outcomes, and this will lead to habits as assessed by satiety-specific outcome devaluation.

Experiment 1

We tested the prediction that a delay in outcome presentation will render instrumental behaviour insensitive to outcome devaluation using rats in a free-operant procedure similar to that used by Dickinson and colleagues (1992; 1996). In this procedure, each lever-press sets up a pellet (FR1), which is presented immediately or sometime later. In Experiment 1 each rat experienced two sessions of training a day, in each of two distinctive contexts, with two different levers and pellets. Thus, each rat learned in one context (A) to press with one lever (for example, right) for an immediate outcome, and in a second context (B) to press an alternative lever (i.e., left) for an alternative outcome, which came 20 seconds after each lever press. Contexts (A vs B), levers and outcomes were all counterbalanced. Following 6 days of instrumental acquisition, we conducted satiety-specific outcome devaluation. In this procedure, animals received ad libitum access to either the pellets that they were receiving during instrumental sessions, or an alternative. In other words, animals are sated with the same or a different pellet, followed by a test on extinction. The benefit of this procedure is that each animal can be pre-fed more than once, and hence outcome devaluation is implemented within-subjects.

Method

Subjects

The subjects were male Lister Hooded rats (n = 14), experimentally naïve, purchased from Charles River (Margate, UK). The rats weighted 292-341 g at the start of the experiment. All rats were maintained at 85% of their free-feeding weight throughout the experiment. Subjects were housed in

groups of 4 with controlled temperature and humidity conditions under an alternating light/dark cycle (red lights on from 7.30 a.m. to 7.30 p.m.).

Apparatus

In Experiment 1, we used 4 Med Associates chambers, in addition to 4 distinctively different chambers (Paul Fray, Cambridge, UK). Each set of 4 chambers were located in different rooms. The Med Associates chambers (29.5×32.5×23.5 cm; Med Associates, Georgia, VT) were equipped with two 4-cm wide retractable levers (standard flat-plate) that were mounted in the intelligence panel 12 cm apart and 8 cm from the grid floor. For each rat, only one lever was present throughout the experiment (right or left, counterbalanced). A white house light (2.5 W, 24 V) was located on the opposite wall. A pellet dispenser delivered individual 45 mg food pellets into a recessed magazine (3.8 cm side and 5.5 cm from the grid floor) situated between the levers. The floor of the chamber was covered with a metal grid with bars separated by 1 cm. The testing chamber was placed within a sound- and light-attenuating housing equipped with a ventilation fan that also screened external noise. Each Paul Fray chamber was housed in an individual sound-attenuating box, and equipped with two retractable levers (4.8 width x 1.6 thick x 2.0 cm protrusion), a recessed food magazine located centrally between two levers, and a 2.8-W house light that illuminated the chamber. A pellet dispenser delivered 45-mg precision pellets into the magazine. A flap door attached to the opening of food magazine detected animal's head entry to the magazine. For each rat, there were two different chambers (Paul Frey or Med Associates), in which only one of two different levers (right vs left) could lead one of two different types of pellets (grain based [5TUM; Catalog # 1811156] vs sucrose based with chocolate flavour [5TUT; Catalog # 1811256]; all Test Diet dustless pellets). These variables were counterbalanced between subjects. As an example, in one setting, right lever presses were followed by immediate grain pellets, whereas in the alternative setting left lever presses they were followed by a chocolate pellet 20 s after each lever press. The operant conditioning chambers were controlled by software using the Whisker control system (Cardinal & Aitken, 2010).

Procedure

In order to decrease the neophobic response to food, and acclimate to the pre-feeding procedure rats were exposed to both kind of pellets in both chambers. The design of Experiment 1 is shown in Figure 1A.

Acclimation to the satiety-specific outcome devaluation. We pre-exposed rats to the pellets by giving them 22 gr of each pellet type in each type of chamber. That is, on days 1-4, each rat was exposed to each setting (Paul Fray or Med Associates) twice, and each time allowed to eat up to 22 gr of grain based or chocolate flavoured pellets. Each rat was exposed to both settings, and experienced both types of pellet in each setting. During these sessions the house lights were off, and no levers were presented. A second bowl containing water was presented in these sessions.

Magazine training. On Days 5 and 6, all rats received two 15-min magazine training sessions per day, in each of the two settings. The lever was not present and pellets were delivered non-contingently on a variable time schedule (VT30s; range 1-59). The houselight was on in this and all subsequent sessions (except during satiety specific pre-feeding).

Instrumental acquisition and re-baseline sessions. On Days 7 to 12, all rats experienced two instrumental acquisition sessions per day, one with immediate outcomes and the other with delayed outcomes (setting [Med Associates vs Paul Frey], lever [right vs left] and type of pellet [grain based vs chocolate] were all counterbalanced with delay [immediate vs delay]). During these sessions, the lever was always reinforced on an FR1 schedule. Sessions lasted until each rat received 30 pellets, or a maximum of 60 minutes, whichever came first. The order of sessions (immediate or delayed) was also counterbalanced for each rat, so that the opposite order was given on each day. The same training was administered during re-baseline sessions following each of the outcome devaluation + extinction tests. These re-baseline sessions were given on days 14, 16, and 18.

Outcome Devaluation + Extinction Test. On Days 13, 15, 17 and 19, rats were placed in the chamber with the houselight off, and allowed to eat from a bowl containing 15g of the pellets that they received in that chamber during instrumental sessions (Same) or the alternative (Different). A second bowl contained water. Because rats preferred chocolate flavoured pellets over grain based pellets, we limited the amount that they could eat during pre-feeding to 15 grams, in order to reduce variance.

Pre-feeding lasted 50 min and after the bowl was removed, the houselights were turned on, and the lever presented for 5 min during which rats could press the lever on extinction. Four devaluation followed by extinction sessions were given, so that in each setting (Immediate vs Delayed) rats were pre-fed with the same or a different outcome. These 4 sessions were interspersed with retraining sessions (see above).

Outcome Devaluation + Reinforced Test. On Days 20-23, the rats received a second round of satiety-specific devaluation tests, but test sessions were as during training, in which each lever press set up a pellet (Immediate or Delayed) as during training. The rats were pre-fed with the same or different pellets as they earned during training in each setup, and tested in both setups.

Data Analysis

The main dependent measure in these experiments was the rate of lever pressing (during acquisition and extinction test sessions). During satiety-specific outcome devaluation, we measured the amount of food eaten during pre-feeding, and compared groups and consumption during pre-feeding with Same vs. Different pellets. Because repeated satiety-specific devaluation tests reduce responding, the data during reinforced tests were analysed as a ratio of lever press relative to the average of the last two retraining sessions. We used within-subjects ANOVAS with Condition (Immediate vs Delayed) as a within-subjects variable (Contiguity). Session (during training) or minute (during test) were within-subject variables, as was Devaluation. During tests, omnibus ANOVAS were followed up with analyses in each condition assessing whether devaluation was successful or not. In all cases we report partial eta squared as a measure of the unbiased, effect size (Cohen, 1992), and the 95% confidence intervals (CI) for the effect size (Nelson, 2016). In addition, we assessed evidence for the null and alternative hypotheses with Bayes Factors using JASP (Wagenmakers et al., 2018).

Results and Discussion

Figure 1B shows the acquisition of instrumental behaviour with immediate and delayed rewards, for all rats in Experiment 1. As can be seen in the Figure, acquisition was faster for the lever followed by immediate outcomes, relative to that followed by a pellet 20 seconds later. These

impressions were supported by the following statistical analyses. A 2 (Contiguity; Immediate vs Delayed) by 6 (Day; 1-6) mixed ANOVA revealed a main effect of Day, F (5, 65) = 12.62, $p < .01$, ηp^2 = 0.49, 95% CIs = 0.27, 0.58, an effect of Contiguity, F (1, 13) = 45.26, $p < .01$, ηp^2 = 0.77, 95% CIs = 0.44, 0.86, and a Day x Contiguity interaction shown in Figure 2B, F (5, 65) = 7.21, $p < .01$, ηp^2 = 0.36, 95% CIs = 0.13, 0.46. Thus, acquisition was faster in the Immediate condition relative to Delayed.

Figure 1C shows the result of the outcome devaluation tests on extinction, averaged over the 5-min extinction session. As it can be appreciated in the figure, all rats showed a clear satiety-specific devaluation effect when tested in the Immediate setting, but no devaluation (i.e., habits) in the Delayed setting. These impressions were supported by the following statistical analyses. A 2 (Contiguity: Immediate vs Delayed) x 2 (Devaluation: Same vs Different) by 5 (Minute: 1-5) mixed ANOVA revealed a main effect of Devaluation, F (1, 13) = 5.21, $p < .05$, ηp^2 = 0.29, 95% CIs = 0.00, 0.56, a main effect of Min, F (4, 52) = 24.89, $p < .01$, ηp^2 = 0.65, 95% CIs = 0.46, 0.73, a Contiguity x Devaluation interaction, F (1, 13) = 5.31, $p < .05$, ηp^2 = 0.29, 95% CIs = 0.00, 0.56, and a contiguity x Min interaction, F (4, 52) = 11.14, $p < .01$, ηp^2 = 0.46, 95% CIs = 0.21, 0.57. To follow up the Contiguity x Devaluation interaction, we compared Same vs Different in each Contiguity condition. These analyses revealed a clear devaluation effect in condition Immediate, F (1, 13) = 9.27, $p < .01$, ηp^2 = 0.41, 95% CIs = 0.03, 0.65, but not in condition Delayed, F (1, 13) = 0.19, $p = .89$, ηp^2 = 0.01, 95% CIs = 0.00, 0.26. Bayes Factors supported the alternative hypothesis in the Immediate condition (BF_{10} = 2.18), and the null hypothesis for the Delay condition (BF_{01} = 5.67). Clearly, there was a devaluation effect when the immediate outcome was devalued, but not when the delayed outcome was devalued. A within-subjects ANOVA comparing consumption across the 4 pre-feeding tests revealed no differences in consumption, F <1 (Immed-Same, M = 12.94, SD = 0.61; Immed-Diff, M = 13.28, SD = 0.52; Delay-Same, M = 14.02, SD = .32; Delay-Diff, M = 13.30, SD = 0.58). During pre-feeding sessions, all rats ate all 15 grams of chocolate pellets, and an average of 11.7 gr of grain based pellets.

Finally, we conducted satiety-specific outcome devaluation tests, but followed by sessions in which the reinforcer was presented, as in training (baseline and ratio data shown in Table 1). The ratio of responses was analysed with a 2 (Contiguity; Immediate vs Delayed) x 2 (Devaluation; Same vs Different) within-subjects ANOVA. This analysis only revealed an effect of Devaluation, F (1, 13) = 12.28, $p < .01$, $\eta p^2 = 0.48$, 95% CIs = 0.07, 0.69, but no other main effects or interactions (largest $F < 1$).

The main conclusion from this experiment is that, when outcomes are delayed, rats are insensitive to devaluation. Basic theories of learning have in general adopted one of two approaches to explain the effects of delays in reinforcement. One explanation assumes that the longer the temporal separation between response and outcome, the weaker the association between these events. In other words, delaying the outcome makes it a less effective reinforcer, and hence more difficult for the rat to detect the relationship between response and outcome (Rescorla & Wagner, 1972). According to the correlational account described in the introduction (Baum, 1973; Dickinson & Perez, 2018), the correlation between response and reinforcement rates will be lower when action-outcome contiguity is disrupted, because the delayed outcomes are experienced in different sampling periods than the responses that produced them. Thus, a prediction of this account is that exposure to the context in the absence of both responses and outcomes will increase the experienced correlation between responses and outcomes. In other words, by making animals experience samples of time in which response and outcome rates are zero, the range of rates across which the delayed instrumental contingency sampled is increased, so that the experienced correlation between response and outcome rates will be higher. For example, in the experiments by Dickinson and colleagues (Dickinson, Watt & Griffiths, 1992) they did not observe acquisition of free-operant instrumental behaviour (using a similar procedure as used in these experiments) when a 64s delay was interposed between response and outcome. However, if rats were given daily 30 min of exposure to the context in which the instrumental behaviour took place (in the absence of the lever and food pellets), rats did show evidence of instrumental acquisition, even with a 64s delay. Similarly, Reed and Reilly (1990) trained rats with a 6-s delay of reinforcement in a free operant procedure, and observed more

responding during a test session when they interposed context extinction sessions between training and test. What neither of these reports assessed is whether this manipulation (context exposure) would restore sensitivity to outcome devaluation. In Experiment 2, we sought to test whether extinction of context conditioning by exposure to the context alone would restore sensitivity to outcome devaluation in rats trained with a 20-s delay in outcome presentation. Based on the account proposed by Dickinson and Perez (2018), and the results of Experiment 1, we predicted that context exposure would restore sensitivity to outcome devaluation in rats trained with 20-s delay.

Experiment 2

Experiment 2 was designed to test an explanation of the insensitivity to devaluation observed when the outcome is presented with a delay following a response. If habits observed with a delay in outcome presentation are due to subjects failing to experience the correlation between response and outcome rates, exposure to the context alone should increase the experienced correlation, thus restoring sensitivity to outcome devaluation. We tested this prediction in two groups of rats that learned to press a lever for a pellet that was presented 20 s after each lever press. One group received additional 30min sessions to the training context in which neither the lever nor the pellets were presented (similar to Dickinson, Watt & Griffiths, 1992; also see Dickinson, Watt & Varga, 1996). We expected to replicate the habit effect in the group that did not receive such exposure (Delay 20) but to observe goal-directed behaviour in the group that received context exposure (Exp-Delay 20).

Method

Subjects and apparatus

The subjects were 24 male Lister Hooded rats, experimentally naïve, purchased from Charles River (Margate, UK). The rats weighed 227-267 g at the start of the experiment, and were maintained at 85% of their free-feeding weight throughout the experiment. Subjects were housed in groups of 4 with controlled temperature and humidity conditions under an alternating light/dark cycle (red lights on from 7.30 a.m. to 7.30 p.m.). The apparatus was 8 Med Associates boxes like those used in

Experiment 1. Similarly, we used sucrose [5TUM; Catalog # 1811251] and grain based [5TUM; Catalog # 1811156] pellets made by Test Diets, counterbalanced.

Procedure

Subjects were randomly assigned to one of two groups, Delay 20 and Exp-Delay 20 (*ns*= 12; see Figure 2A). In order to decrease the neophobic response to food novelty, the rats were exposed to the pellets used during the experiment in their home cages, a few days before the experiment started.

Magazine training. On Days 1 and 2, all rats received two 15-min magazine training sessions in which the lever was not present and pellets were delivered non-contingently on a variable time schedule (VT30s; range 1-59). The session ended after 15 minutes, or after 30 pellets had been delivered, whichever came first. The house light was on in this and all subsequent sessions, except during pre-feeding during outcome devaluation.

Instrumental acquisition. On Days 3 to 8, all rats experienced six instrumental acquisition sessions in which lever pressing was reinforced on an FR1 schedule. Rats in Group Delay 20 received the pellet (sucrose or grain based, counterbalanced) 20 s after a lever press, as did rats in Group Exp-Delay 20. The difference is that on each training day, rats in the latter group received an additional 30-minute session of exposure to the context, immediately before each instrumental acquisition session. During context exposure, the houselights were on but there was no lever present in the chamber. Nor did rats in this group receive any pellets. Instrumental acquisition sessions lasted until each rat received 30 pellets, or a maximum of 120 minutes, whichever came first. Similar training was given on Day 10, the retraining day in between the Outcome Devaluation + Extinction Tests.

Outcome Devaluation + Extinction Test. On Days 9 and 11, rats were placed in the chamber with the house light off, and allowed to eat from a bowl containing 50 g of the pellets that they received during instrumental sessions (Same) or the alternative (Different). A second bowl contained water. Pre-feeding lasted 50 min and after the bowls were removed, the houselights were turned on, and the lever presented for 10 min during which rats could press the lever on extinction. These two sessions were interspersed with a retraining session (on Day 10) similar to Day 8.

Outcome Devaluation + Reinforced Test. On Days 12 and 13, the rats received a second round of satiety-specific devaluation, but test sessions were as during training, in which each lever press set up a pellet (Delay 20) with the same duration as during training. These reinforced tests were conducted to make sure the rats were sensitive at a behavioural level to the satiety-specific devaluation manipulations.

Results and Discussion

Figure 2B shows the data during instrumental acquisition. Although there is a tendency towards more responding in Group Exp-Delay 20 (as expected), this effect in small, which is not surprising given that effects of context exposure have been observed when instrumental responses were followed by an outcome 64 s later (Dickinson et al., 1992). A 2 (Group: Delay 20 vs Exp-Delay 20) by 6 (day: 1-6) mixed ANOVA revealed a main effect of Day, F (5, 110) = 48.75, $p < .01$, ηp^2 = 0.69, 95% CIs = 0.57, 0.74. The effect of Group was not significant, and small sized, F (1, 22) = 2.27, p = .14, ηp^2 = 0.09. These two factors did not interact. In brief, the analysis suggests acquisition of free operant responding in both groups, and a small tendency towards better performance in the group that received context exposure. Figures 2C and D show the rate of lever pressing during the extinction tests (10 minutes) following pre-feeding with the same or different pellets for Group Delay 20 and Exp-Delay 20, respectively. The figure suggests that there was no devaluation effect in Group Delay 20 (Fig 2C), thus replicating the findings from previous experiments. In the Group that received context exposure (Exp-Delay 20; Fig 2D) show restored sensitivity to outcome devaluation. These impressions were supported by the following analyses. Three rats were excluded due to extreme responding (more than 2 SD above the group mean) during extinction tests; 1 rat in Group Delay 20 and 2 in Group Exp-Delay 20. A 2 (Group; Delay 20 vs Exp-Delay 20) by 2 (Devaluation: Same vs Different) by 10 (Minutes: 1-10) mixed ANOVA revealed an effect of Devaluation, F (1, 19) = 6.01, $p < .05$, ηp^2 = 0.24, 95% CIs = 0.00, 0.49, a main effect of Minutes, F (9, 171) = 15.89, $p < .01$, ηp^2 = 0.45, 95% CIs = 0.31, 0.51, a main effect of Group, F (1, 19) = 8.24, $p < .05$, ηp^2 = 0.30, 95% CIs = 0.02, 0.54, a Minute x Group interaction, F (9, 171) = 2.12, $p < .05$, ηp^2 = 0.10, 95% CIs = 0.00, 0.14, and a Devaluation x Minute x Group interaction, F (9, 171) = 2.10, $p < .05$, ηp^2 = 0.09, 95% CIs = 0.00, 0.14.

We followed up the interaction with separate analysis in each group. A 2 (Devaluation: Same vs Different) by 10 (Minutes: 1-10) within-subjects ANOVA in Group Delay 20 revealed an effect of Minute F (9, 90) = 8.87, $p < .01$, $\eta p^2 = 0.47$, 95% CIs = 0.26, 0.54, but no effect of devaluation F (1, 10) = 1.34, $p = .27$, $\eta p^2 = 0.11$, 95% CIs = 0.00, 0.45, nor an interaction, $F < 1$ (see Fig 3C). The Bayes Factor suggested that these data are 1.3 times more likely under the null hypothesis (BF_{01} = 1.31). A similar analysis in Group Exp-Delay 20 revealed a marginal effect of Devaluation, F (1, 9) = 4.41, $p = .06$, $\eta p^2 = 0.33$, 95% CIs = 0.00, 0.61 (with a large effect size), an effect of Minute F (9, 81) = 8.61, $p < .01$, $\eta p^2 = 0.49$, 95% CIs = 0.27, 0.56, and an interaction, F (9, 81) = 2.47, $p < .05$, $\eta p^2 =$ 0.21, 95% CIs = 0.00, 0.28. The Bayes Factor suggested that these data are 64 times more likely under the alternative hypothesis (BF_{10} = 64). A comparison of the devaluation effect in the first 8 minutes, when extinction was not as pervasive, showed an effect of devaluation F (1, 9) = 5.76, $p <$ .05, $\eta p^2 = 0.39$, 95% CIs = 0.00, 0.65 (see Fig 3D). Thus, consistent with the prediction based on the correlational account, extinction of the training context restored goal-directedness even when the outcome was presented 20 s after responding. A 2 (Group; Delay 20 vs Exp-Delay 20) by 2 (Devaluation; Same vs Different) mixed ANOVA on the amount of food eaten during pre-feeding revealed no differences between groups and/or devaluation conditions, all F's <1, (Delay 20-Same, M = 11.83, SE = 1.37; Delay 20-Diff, M = 10.60, SE = 1.26; Exp-Delay 20-Same, M = 11.06, SE = 1.44; Exp-Delay 20-Diff, M = 10.54, SE = 1.33). Thus there were no differences between Groups or Conditions in terms of amount of food eaten.

Finally, we conducted satiety-specific outcome devaluation tests, but followed by sessions in which the outcome was presented, as in training (see Fig 2E). Due to equipment problems (pellet dispensers blocked) 4 rats were excluded because they did not receive pellets throughout the session, 3 in Group Delay 20, and 1 in Group Exp-Delay 20. The ratio of responses was analysed with a 2 (Group: Delay 20 vs Exp-Delay 20) x 2 (Devaluation; Same vs Different) mixed ANOVA. This analysis only revealed an effect of Devaluation, F (1, 18) = 4.49, $p < .05$, $\eta p^2 = 0.20$, 95% CIs = 0.00, 0.46, but no other main effects or interactions (largest $F < 1$). Average (Days 8 and 10) baseline rates

of responding used to calculate the ratios were 5.05 LP/Min (*SD* = 1.82) for Group Delay 20, and 5.85 LP/Min (*SD* = 1.95) for Group Exp-Delay 20.

General Discussion

The purpose of this study was to test the hypothesis that delayed rewards promote habit formation, as assessed by outcome devaluation tests following free-operant instrumental acquisition. An additional objective was to ascertain the mechanism underlying the habit with delayed outcomes. In two experiments, we observed no effect of outcome devaluation with delayed outcomes. Finally, in Experiment 2 we tested the hypothesis that the habit we observed with delayed rewards was due to subjects experiencing at best a weak correlation between response and outcome rates. We tested this possibility by administering context alone exposure which should increase the experienced correlation, and this restored goal-directedness despite the rats having received the outcome 20 seconds following a response.

The current findings can be accommodated by the suggestion that goal-directed instrumental behaviours are controlled by two representations 1) one of the experienced correlation between responses and outcomes, and 2) a representation of the value of the outcomes (Dickinson, 1994). One determinant of the relation between responses and outcomes is the currently experienced correlation between rates of responding and rates of reinforcement (Dickinson, 1985). This can be implemented using Baum's correlational-based law of effect (1973), which states that the rate of responding is determined by the correlation between rates of responding and rates of reward assessed across a series of time samples. This account explains the findings outlined in the Introduction concerning extended training, schedules of reinforcement and the use of choice procedures in determining whether goal-directed or habits are observed. With extended training, rates of responding are stable and thus animals will experience less of a correlation between responses and outcomes (Adams, 1982). In other words, in the absence of response variation the rat no longer experiences a rate relationship. Similarly, interval schedules of reinforcement prevent such correlation, because reward rates are only weakly related to response rates (Dickinson et al., 1983). This explains why both procedures lead to the expression of habits. In addition, this explanation can

also explain why the availability of a choice procedure can attenuate the expression of habits following extended training. With the use of a choice procedure, animals will experience the correlation between each response and each outcome when they alternate between responses even with extended training, and hence remain goal-directed (Kosaki & Dickinson, 2010). Recent simulations of a formal version of the *dual-system theory* of instrumental behaviour show how it can successfully handle some of these phenomena (Dickinson & Perez, 2018).

According to the *dual-system theory* of instrumental behaviour, delayed outcomes weaken the experienced relationship between responses and outcomes, rendering the behaviour less sensitive to outcome devaluation. Perez (2017; Chapter 4) recently simulated Baum's correlational approach (which instantiates the response-outcome causal representation) and found that it models well the effects of outcome delays observed by Dickinson and colleagues (1992). The reason why the correlational account explains the delay of reinforcement effect is that the delayed outcome is more likely to fall on a different time-sample from the response, and hence degrade the experienced response-outcome correlation (Baum, 1973). Dickinson et al., (1992) provided evidence that rats are sensitive to these response-outcome delays, and human participants have been observed to reduce both their rate of responding, and causal judgements between a response and the outcome when the latter is delayed by a few seconds (Shanks & Dickinson, 1991). Therefore, unlike other models aimed at explaining goal-directed and habitual forms of instrumental behaviour (Daw, Niv & Dayan, 2005; Dezfouli & Balleine, 2012), the model advanced by Dickinson & Perez (2018; also see Perez, 2017) can uniquely explain the present observations that instrumental behaviour followed by delayed consequences is less sensitive to outcome devaluation. The alternative models do not readily explain the current findings, simply because they do not account for the effects of delayed rewards, and hence are silent about the effects reported here. Notably, the findings of Experiment 2 are also consistent with the *dual-system theory* of instrumental behaviour. That is, exposure to the context will force rats to experience time samples in the absence of responses and outcomes, and this exposure should enhance the response range across which the animal assesses the response-outcome correlation. This should increase the experienced correlation between responses and outcomes, in

particular when the correlation between responses and outcomes is weak due to the delay in outcome presentation, and ultimately restore goal-directed behaviour.

To our knowledge, this is the first time that habit formation has been tested with delayed consequences, and the current findings suggest that this is an important candidate condition favouring habit formation. This is relevant to human behaviour because many of the day-to-day instrumental activities and decisions performed by humans are followed by delayed (rather than immediate) consequences. For example, humans make plans for retirement, or they decide to save money for their children's education. All these activities require bridging long periods of time between the response and its consequences. Besides these long delays, there are some activities such as gambling that also involve delayed consequences, and may result in habit development. In slot machines, the delay between pressing the button and seeing the result is between 3 to 6 s (Cho et al., 2017; see Footnote 1). Fully electronic roulettes in the UK allow roughly 4 spins per minute[1], which implies a delay of over 10 s between the moment participants place their bets, and when they see the outcome. Given these delays between response and outcome presentation, and the current findings, it is not surprising that participants may be able to play so vigorously and unable to stop. There are obvious differences between the present controlled experiments in rodents and real-life scenarios, but these findings may offer some information about the underlying causes of habitual behaviour in humans.

In summary, our experiments show for the first time in rodents that delayed consequences may lead to habit formation, as assessed by outcome devaluation technique. Furthermore, we made a first step in uncovering the underlying mechanisms for habit formation with delayed consequences, by showing that context exposure (extinction) can restore goal-directed behaviour. Overall, these findings are consistent with a *dual-system theory* of instrumental behaviour.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 34B*, 77-98.

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 33B*(2), 109–121.

Baum, W. (1973). The Correlation-Based Law of Effect. *Journal of the Experimental Analysis of Behavior, 20*: 137–153.

Cardinal, R. N. and Aitken, M. (2010). Whisker: a client-server high-performance multimedia research control system. *Behavior research methods, 42*: 1059–1071.

Chu SWM, Limbrick-Oldfield E, Murch WS, Clark L. (2018) Why do slot machine gamblers use stopping devices? Findings from a 'Casino Lab' experiment. *International Gambling Studies, 18*: 310-326. doi: 10.1080/14459795.2017.1413125

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*, 1704–1711.

DeRusso, A. L., Fan, D., Gupta, J., Shelest, O., Costa, R. M., & Yin, H. H. (2010). Instrumental uncertainty as a determinant of behavior under interval schedules of reinforcement. *Frontiers in Integrative Neuroscience, 4*. https://doi-org.ezproxy3.lib.le.ac.uk/10.3389/fnint.2010.00017

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. European Journal of Neuroscience, 35, 1036–1051.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences, 308*: 67–78.

Dickinson, A. (1994). Instrumental conditioning. *Animal Cognition and Learning*. Ed. by N. J. Mackintosh. London: Academic Press. Chap. 3: 45–78.

Dickinson, A., Nicholas, D. J., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *Quarterly Journal of Experimental Psychology, 35B*, 35-51.

Dickinson, A., & Pérez, O. D. (2018). Actions and Habits: Psychological Issues in Dual-system Theory. In: *Understanding Goal-Directed Decision Making: Computations and Circuits*. R. Morris, A. Bornstein & A. Shenhav (Eds.). London: Academic Press. Ch 1: 1–25.

Dickinson, A., Watt, a. and Griffiths, W. J. H. (1992). Free-operant acquisition with delayed reinforcement. *The Quarterly Journal of Experimental Psychology Section B, 45*: 241–258.

Dickinson, A., Watt, A., & Varga, Z. I. (1996). Context conditioning and free-operant acquisition under delayed reinforcement. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology, 49B*, 97–110.

Dolan, R. and Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80*: 312–325.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience, 8*, 1481–1489. https://doi-org.ezproxy4.lib.le.ac.uk/10.1038/nn1579

Gillan, C. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M., … Robbins, T. W. (2014). Enhanced avoidance habits in obsessive-compulsive disorder. *Biological Psychiatry, 75*, 631–638.

Kosaki, Y., & Dickinson, A. (2010). Choice and contingency in the development of behavioral autonomy during instrumental conditioning. *Journal of Experimental Psychology: Animal Behavior Processes, 36*, 334-342. doi:10.1037/a0016887

Nelson, A., & Killcross, S. (2006). Amphetamine exposure enhances habit formation. *The Journal of Neuroscience, 26*, 3805–3812.

Pérez, O. D. (2017). A cooperative dual-system model of instrumental conditioning. Unpublished Doctoral Dissertation. University of Cambridge. UK.

Reed, P., & Reilly, S. (1990). Context extinction following conditioning with delayed reward enhances subsequent instrumental responding. *Journal of Experimental Psychology: Animal Behavior Processes, 16*, 48–55. https://doi-org.ezproxy4.lib.le.ac.uk/10.1037/0097-7403.16.1.48

Rescorla, R. and 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.

Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S., & Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. *Trends in Cognitive Sciences, 16*, 81–91.

Shanks, D.R., & Dickinson, A. (1991). Instrumental judgement and performance under variations in action-outcome contingency and contiguity. *Memory and Cognition, 19*, 353-360.

Thorndike, E. L. (1911). *Animal intelligence: experimental studies*. New York: Macmillan.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*, 58–76.

Author Notes

This study was funded by a Marie Curie Intra-European Fellowship (PIEF-GA-2009-237608) awarded by the European Commission to Gonzalo Urcelay and Jeff Dalley. All or part of these findings were presented by Urcelay and Jonkman at the Associative Learning Symposium at Gregynog (Wales, UK), in 2011, and by Urcelay et al., at the International Meeting of the Spanish Society for Comparative Psychology (Meeting in Avila, 2018). The authors wish to thank Prof Anthony Dickinson for invaluable discussions during the conception and running of these experiments. All data analysed and reported in this manuscript are available at:

https://doi.org/10.25392/leicester.data.8299577

Footnotes

[1] Dr Luke Clark. Personal communication on 05 June, 2018.

.

Table 1

Contiguity	**Baseline**	**Devaluation**	***Mean***	***SD***	**N**
Immediate	*M* = 18.84	Same	0.51	0.31	14
	SD = 6.08	Different	0.72	0.17	14
Delayed	*M* = 6.41	Same	0.63	0.38	14
	SD = 3.31	Different	0.75	0.44	14

Table 1: Descriptive results of reinforced tests in Experiment 1. The Baseline column shows the mean LP/min during the three retraining sessions (Days 14, 16 and 18). These values were used to calculate ratios during the reinforced test sessions, which are shown in the columns labelled Mean and SD for the different conditions in the rows.

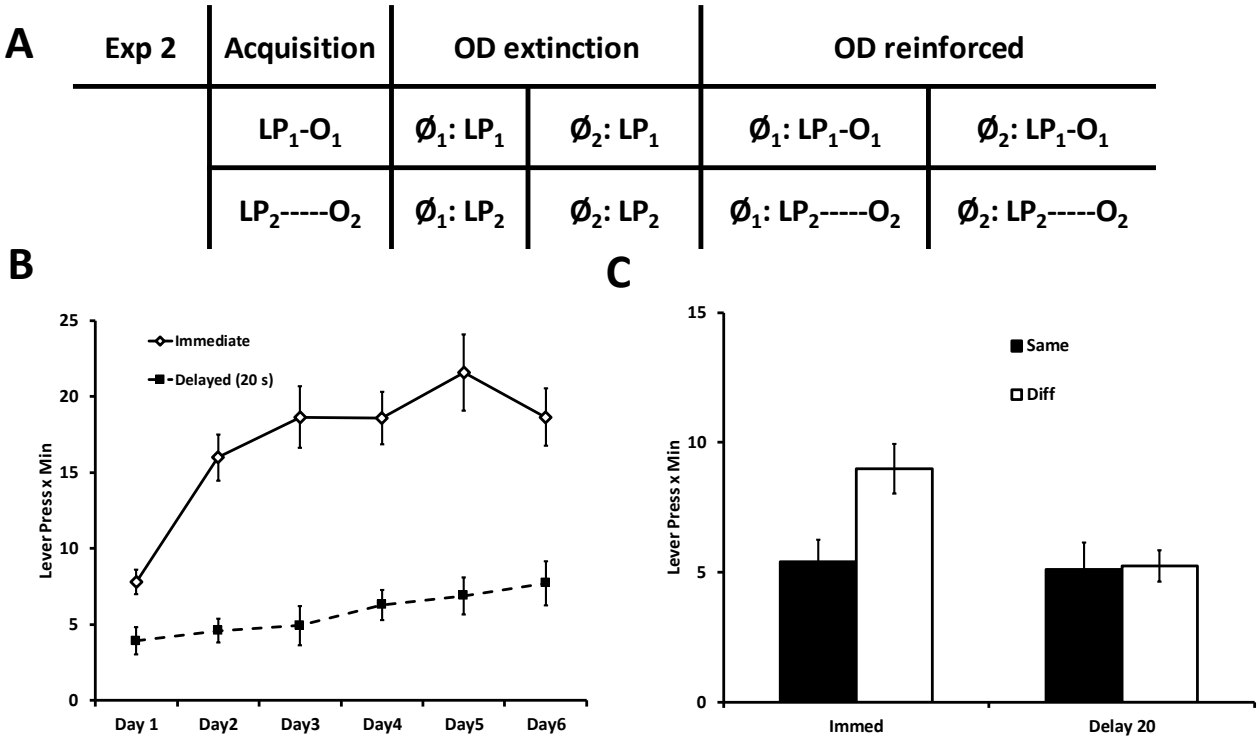

Figure 1. Design and Results of Experiment 1. A) Experimental Design. One group received 6 days of instrumental acquisition (FR1) with either immediate or delayed (20s) outcomes in different environments. Training was followed by satiety-specific outcome devaluation test (first on extinction, followed by a reinforced test). B) Acquisition of instrumental behaviour with Immediate or Delayed (20s) outcomes. C) Results of satiety-specific outcome devaluation tests for conditions Immediate and Delayed 20. Clear outcome devaluation was observed in the context were lever-presses were followed by immediate outcomes, but not in the context where the outcome was delayed. Error bars represent within-subjects SE.

A

Exp 2	Acquisition		OD extinction		OD reinforced	
Delay 20	Home	LP-----O_1	$Ø_1$: LP	$Ø_2$: LP	$Ø_1$: LP-O_1	$Ø_2$: LP---O_1
Exp-Delay 20	CTX	LP-----O_1	$Ø_1$: LP	$Ø_2$: LP	$Ø_1$: LP---O_1	$Ø_2$: LP---O_1

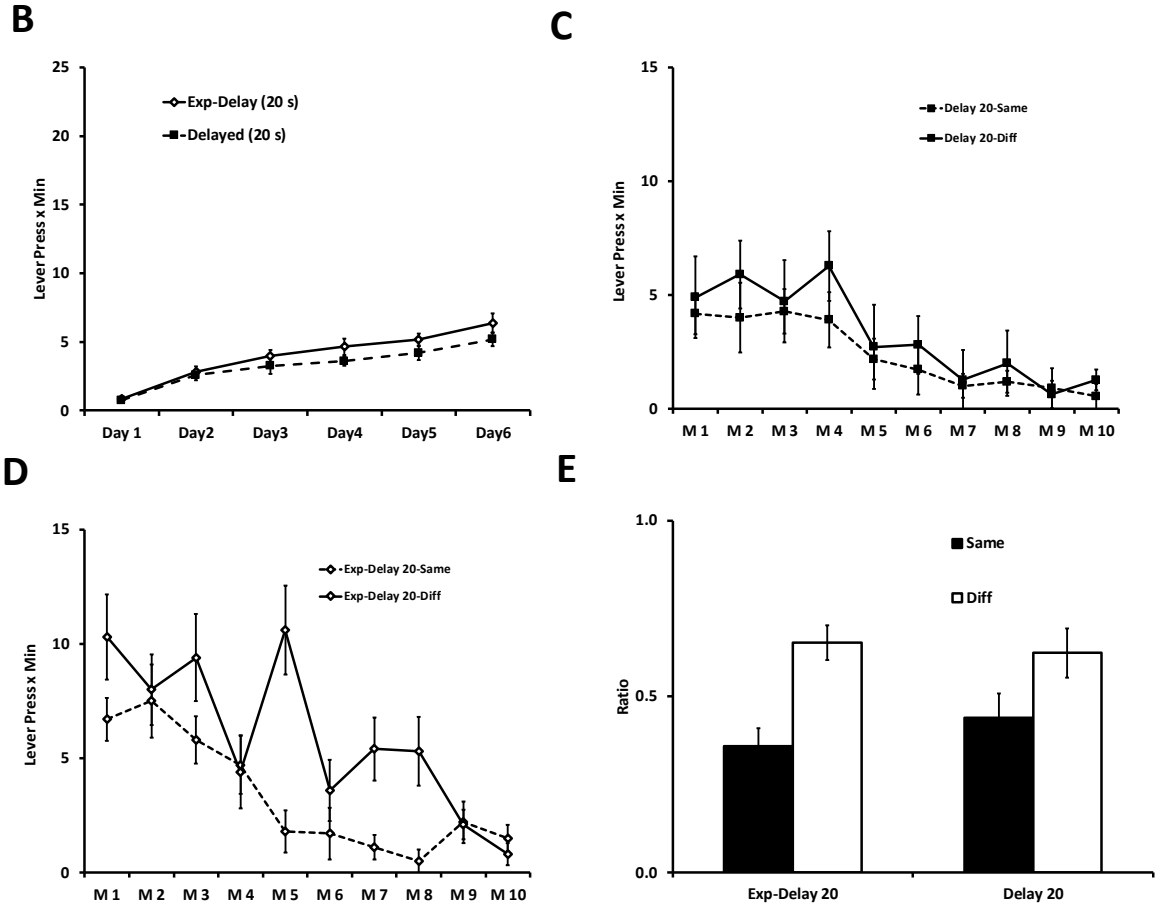

Figure 2. Design and Results of Experiment 2. A) Experimental Design. Two Groups received 6 days of instrumental acquisition (FR1) with delayed (20s) outcomes, but one Group (Exp-Delay 20) received additional 30 min sessions of context exposure (i.e., extinction). Training was followed by satiety-specific outcome devaluation test (first on extinction, followed by a reinforced test). B) Acquisition of instrumental behaviour in the two groups. C) Results of satiety-specific outcome devaluation tests for Group Delay 20. D) Results of satiety-specific outcome devaluation tests for Group Exp-Delay 20. C) and D) are tests on extinction. E) Results of the reinforced test. Error bars represent within-subjects SE.