The discovery of novel actions is affected by very brief reinforcement delays and reinforcement modality.

WALTON, Tom, THIRKETTLE, Martin <http://orcid.org/0000-0002-6200-3130>, REDGRAVE, Pete, GURNEY, Kevin N and STAFFORD, Tom

Published version

WALTON, Tom, THIRKETTLE, Martin, REDGRAVE, Pete, GURNEY, Kevin N and STAFFORD, Tom (2013). The discovery of novel actions is affected by very brief reinforcement delays and reinforcement modality. Journal of motor behavior, 45 (4), 351-360.

Published as: Walton, T., Thirkettle, M., Redgrave, P., Gurney, K. N., & Stafford, T. (2013). The Discovery of Novel Actions Is Affected by Very Brief Reinforcement Delays and Reinforcement Modality. *Journal of Motor Behavior, 45*(4), 351-360.

The discovery of novel actions is affected by very brief reinforcement delays and reinforcement modality

Tom Walton, Martin Thirkettle, Pete Redgrave,

Kevin N. Gurney and Tom Stafford

Department of Psychology, University of Sheffield

Author Note

The authors thank the ABRG for providing a supportive network in which to conduct research. We thank Ashvin Shah for discussion of reinforcement learning and computational learning theory in general. This work was supported by an EPSRC funded PhD grant and Doctoral Prize Fellowship awarded to Thomas Walton and by the European Community 7th Framework Programme (FP7/2007–2013), “Challenge 2 - Cognitive Systems, Interaction, Robotics,” grant agreement No. ICT-IP-231722, project “IM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots”.

Correspondence concerning this article should be addressed to

Tom Stafford, Department of Psychology, University of Sheffield, Western Bank, Sheffield, S10 2TP; Tel: +44 (0) 114 22 26620; Email: t.stafford@sheffield.ac.uk

Published as: Walton, T., Thirkettle, M., Redgrave, P., Gurney, K. N., & Stafford, T. (2013). The Discovery of Novel Actions Is Affected by Very Brief Reinforcement Delays and Reinforcement Modality. *Journal of Motor Behavior, 45*(4), 351-360.

Abstract

We investigated the ability of human participants to discover novel actions under conditions of delayed reinforcement. Participants used a joystick to search for a target indicated by visual or auditory reinforcement. Reinforcement delays of 75-150 ms were found to significantly impair action acquisition. We also found an effect of modality, with acquisition superior with auditory feedback. The duration at which delay was found to impede action discovery is, to our knowledge, shorter than that previously reported from work with operant and causal learning paradigms. The sensitivity to delay we report, and the difference between modalities, is consistent with accounts of action discovery that emphasise the importance of a 'time stamp' in the motor record for solving the credit assignment problem.

Keywords: Credit assignment problem; delayed reinforcement; reinforcement learning; action acquisition

Published as: Walton, T., Thirkettle, M., Redgrave, P., Gurney, K. N., & Stafford, T. (2013). The Discovery of Novel Actions Is Affected by Very Brief Reinforcement Delays and Reinforcement Modality. *Journal of Motor Behavior, 45*(4), 351-360.

The discovery of novel actions is affected by very brief reinforcement delays and modality

Thorndike (1911) famously distilled the process of action acquisition down to the idea that learning depends on the reinforcement of recent motor output; the ‘law of effect’. It has long been recognised that in order for an instrumental response to be acquired in this way, an animal must solve the computational task of determining which portion of its motor output was necessary for causing an event to occur. Because there is no objective way of deciding how far back into the record the animal should look in order to identify the earliest causally relevant behavioural components, previous research has found that learning is continuously attenuated by increasing delay, but with no absolute cut off. The challenge of teasing out causal components from a stream of behavioural variance is known as the credit assignment problem (Minsky, 1961), something that is particularly easy to appreciate in situations where reinforcement is delayed. One common solution for algorithms used in models of reinforcement learning is to cope with delay by maintaining a trace of the successful pattern of activity for an extended period of time, such that it remains eligible for reinforcement at the moment when the outcome eventually occurs (Barto, Sutton & Brouwer, 1981; Singh & Sutton, 1996; Wickens, 1990). Although these temporal difference algorithms have been shown capable of solving many learning delay problems, it remains an open question as to exactly how the credit assignment problem is solved in the animal brain. Experiments

involving the manipulation of reinforcement delay are, therefore, important to understanding action acquisition and eligibility periods because they provide a means of exacerbating the credit assignment problem whilst leaving other aspects of a task unchanged. A basic question which this paper addresses is the duration at which a delay between action and outcome begins to exert an effect on learning.

The subject of action acquisition with delayed reinforcement has received much experimental attention. Studies have employed a wide range of subject species (e.g, humans, Okouchi, 2009; to fish, Lattal & Metzger, 1994), using both resetting and non-resetting delays (Lattal & Gleeson, 1990; Dickinson, Watt & Griffiths, 1992), different forms of reinforcement (cf Shanks & Dickinson, 1991; Van Haaren, 1992; Lattal & Metzger, 1994; Snycerski et al., 2005) and used a variety of dependent variables (explicit judgments of causality, Shanks, Pearson and Dickinson, 1989; percentage of correct responses, Okouchi, 2009; reaction times following priming Elsner & Hommel, 2004; differences from yoked responses, Dickinson et al., 1992; and simple rate of response, Snycerski, Laraway & Poling, 2005). Finally, the means of reporting the effect of delay varies widely between studies. This includes the threshold at which acquisition is considered to have occurred (Snycerski et al., 2004; Snycerski et al., 2005), whether the result is reported as the point at which delay begins to exert an effect (Black, Belluzzi & Stein, 1985), whether learning is even possible (Critchfield & Lattal, 1993) and also the point at which learning is no longer possible (Dickinson et al., 1992). Taken

as a whole, this research has resulted in extremely wide estimates of the general sensitivity of action acquisition to the effects of delay, with delays as short as 1 s (Black et al., 1985) being shown to severely disrupt learning at one end of the spectrum, whilst other studies show that learning is still possible with delays of as much as 45 s (Snycerski et al., 2005).

However, the effect of delay on action acquisition in much recent research can be difficult to interpret because there has been a departure from techniques designed specifically to record acquisition, towards techniques designed primarily to record response maintenance. The latter are limited in some important respects. Free operant lever-press procedures, for example, are an excellent methodological option if we are interested in measuring choice or the decision to respond, but they do not provide direct performance metrics, such as the time taken to produce an action, the number of errors produced or the overall efficiency of a given movement. When adopting this approach for the study of acquisition, the extent to which an action has been learnt is gauged by recording the frequency of elicitation rather than any direct measure of action performance. Therefore, when employing maintenance procedures, it can be unclear how much of the effect is attributable to differences in the ability to acquire a response and how much down to differences in an animal's inclination to respond.

Alternatively, discrete-trial procedures are aimed specifically at deriving learning curves that directly describe performance across trials containing single attempts at a task. There is a history of using this approach to investigate the acquisition of novel actions in various forms including puzzle-boxes (Thorndike, 1911), maze tasks (Morris, 1981; Tolman, 1948) and runway tasks (Hill, 1939). Whilst a technical definition of discrete-trial procedures can be offered by drawing a contrast with free-operant techniques (e.g. Hachiya & Ito, 1991), in practice the most relevant characteristic of this approach is the opportunity it provides for intervening with the timing and the spatial properties of performance. Experimenters can treat each new trial as a clean example of an animal's ability to perform the action under investigation, and the timing and body position of the animal can be manipulated across trials and across individuals.

Despite the early methodological focus on response acquisition heralded by Thorndike's work, the majority of theoretical and physiological research of recent years has focused on the decision to respond and other economic aspects of reinforcement learning (much of this work referencing the highly influential study by Schultz, Dayan & Montague, 1997). More recently, however we have raised questions regarding the neural underpinnings of acquisition (as opposed to maintenance or selection) with the idea that a reinforcement learning system centred on the subcortical basal ganglia circuit might be specialised to perform the kind of trial and error action acquisition

originally described by Thorndike (Redgrave & Gurney, 2006; Redgrave, Gurney & Reynolds, 2008). In particular, it is suggested that timing information carried by dopamine neurons might serve to “stamp in” recent motor output before the organism has had time to produce contaminating – non-contingent – motor output in response to the stimulus. This dopamine signal arrives in the striatum at a very short latency, typically 70-100 ms after a stimulus (Schultz, 1998). This is an incredibly rapid transmission time and one which suggests functional significance. Specifically, such a low latency signal would be ideally suited, we have proposed, for allowing the identification of components of ongoing action which are responsible for triggering unexpected outcomes. In this way the dopaminergic signal could be instrumental in solving the credit assignment problem and thus laying the foundation for the discovery of novel action-outcome pairings.

This idea places particular emphasis on the very earliest phase of acquisition, what we might call ‘action discovery’: the period before an action has been added to the behavioural repertoire, during which the animal begins to discover its ability to causally interact with a stimulus. Such learning is thought to rely, in part, on the uninsightful and immediate reselection of behaviours that brought about novel but non-noxious outcomes. As such, acquisition is a type of learning that ought to be especially sensitive to delays of reinforcement as a result of the contaminating motor output that must be discounted during the learning process.

In order to better focus on the very early period of acquisition, we have developed a new procedure to investigate action learning. For a detailed introduction to the current and related procedures, see Stafford et al. (2012). Our task is designed to allow experimental investigation of those processes involved in the initial acquisition of an action, as movements which cause reliable outcomes are identified and bound together. As such, although our task is informed by studies of rate of responding it is conceptually separate.

Obviously a behavioural experiment, such as reported in the current paper, must have minimal bearing on underlying biological mechanisms such as are the subject of the Redgrave and Gurney (2006) theory. This theoretical position does however offer a biological solution to the credit assignment problem. This must be solved in the process of action acquisition. Analysis, such as presented by Redgrave and Gurney (2006), suggests that delays in the arrival of sensory information at a point where it can be combined with a record of motor output will have a strong effect on the efficacy of action learning. The suggestion is that delays for any reason - external or internal to the nervous system - will impair action learning. In other words, even delays which arise due to differences in afferent transmission times may have an effect on action learning.

Perhaps surprisingly, delays of less than half a second have, to our knowledge, received no attention in the specific area of trial and error action discovery and acquisition. We therefore sought to investigate action discovery at a range of delays all beneath 0.5 s in duration, the aim being to determine the shortest duration at which an effect of delay on action discovery is detectable and reveal whether the underlying learning mechanism is genuinely as robust to delayed outcomes as previous research seems to indicate. Furthermore, if learning is particularly sensitive to delay, such sensitivity could also be revealed through differences in the efficacy of reinforcing signals presented to different sensory modalities. There are known differences in sensory transmission times between the sense, which can be detected using direct recording from multisensory areas (Wallace, Wilkinson & Stein, 1996) or behavioural measures such as manual reaction time (Sanders, 1998). In particular auditory transmission times are briefer than visual, suggesting that reinforcement delivered via audition will be subject to a smaller relative delay. For this reason we extended the study to manipulate the modality of the reinforcing signal.

Experiment 1

Method

Participants

51 people (47 female) participated in the experiment. Ages ranged from 18 to 24 years with a mean age of 19 (SD = 1.4 years). Participants were all undergraduate students at the University of Sheffield who took part in return for credits in the department's research participation scheme. All participants reported normal or corrected to normal vision and hearing and were naive to the purpose of the experiment and the independent variable. Ethical approval was granted by the department's ethics committee. In both this experiment and experiment 2 we found no indication of gender differences in task performance.

Apparatus & Task specification

The essence of the task is for the participant to find a target area within the range of movements possible with a joystick. The location of the target area is signalled by what we term 'reinforcing stimuli'. Successfully completing the task requires the participant to combine information from the reinforcing stimuli, memory of how they have recently moved the joystick (this does not necessarily have to be explicit memory) and decisions about how to move next. The criterion which defines successful identification of the target area was termed the 'escape criterion'.

The experimental program was written in Matlab (Version 2007), using the

Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). A commercial joystick (Logitech extreme 3D pro joystick, P/N: 863225-1000) was used as the input device.

The search space was defined as a square with a side length of 1024 units. Movements of the joystick were physically restricted by a square aperture at the base of the stick and mapped onto movements within the search space in a one to one fashion, with the joystick starting in the centre of the search space at the beginning of each trial. Once released from the grip of a participant, the joystick's internal spring returned it to the centre of the search space within a tolerance of 10 units.

The size of the target area (the 'hotspot') that participants were required to find was determined through pilot tests and ultimately set to occupy 0.91% of the search space. At the beginning of each new trial, the centre of the hotspot was positioned randomly on an annulus shaped region of the search space (Figure 1). The inner edge of the annulus was placed at a distance equal to the diameter of the hotspot from the centre of the search space. The outer edge of the annulus was a distance equal to the radius of the hotspot from the outside edge of the search space at its closest point. These dimensions ensured that the hotspot never overlapped the central starting point or the outer edge of the search space.

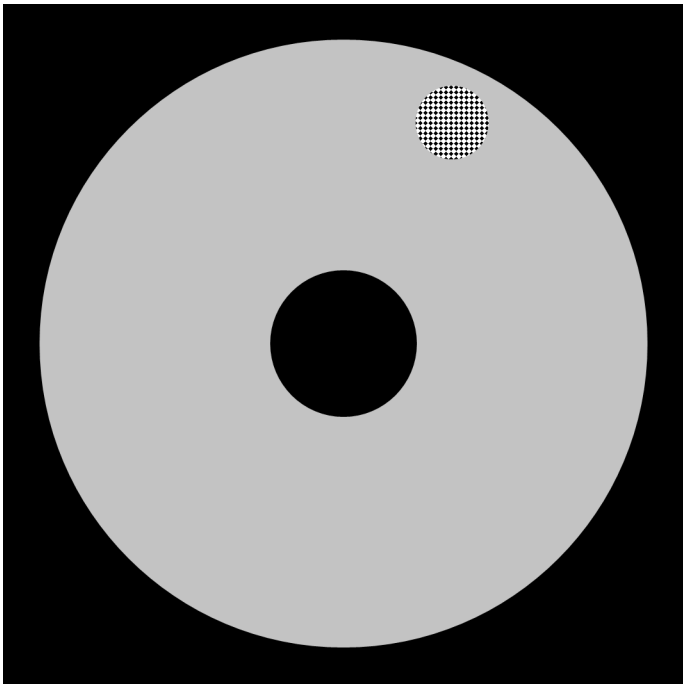

Figure 1. Experimental search space and hotspot positioning. The black area represents the search space and the overlying grey annulus represents the area of the search space in which the centre of the hotspot (patterned circle) could be randomly positioned at the start of each trial. The diagram is drawn to scale.

Any movement of the joystick into the hotspot region of the search space was defined as a 'hit'. The occasion of each hit was signalled by a stimulus, which we term the reinforcing signal or reinforcer. In the auditory reinforcement condition this was a

600 Hz pure tone of 10 ms duration. In the visual reinforcement condition this was a screen flash in which the display monitor changed from black to white and back to black, taking 17 ms.

The reinforcement delay followed the occasion of a hit with a delay which was consistent within each trial but could take one of six values between trials; 0, 75, 150, 225, 300 and 375 ms. In order that reinforcing signals should be clearly and discretely presented a refractory period of 25 ms was introduced after each instance of reinforcement during which another stimulus could not occur. The relation of participant actions to hits and reinforcement is shown schematically in Figure 2.

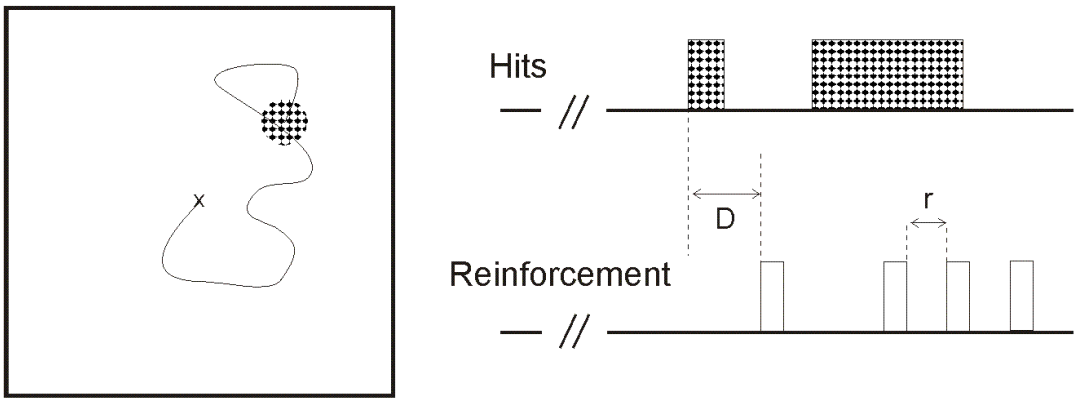

Figure 2. Schematic of path on representative trial (left) and corresponding timeline of registration of hits and display of reinforcement (right). Path of participant's movement begins at X and moves through and then back into the target area (shaded). Reinforcement is offset from hits by a delay, D. Instances of reinforcement separated by refractory period, r.

Generating a single hit was not sufficient to bring the current trial to an end. Instead, the participant was required to hold the joystick in a stable position to terminate the trial. This escape criterion was defined by the number of hits required within 1 s to bring an end to a trial. Just as with the hotspot size parameter, this was set through pilot testing at 15 hits within a second. From an individual participant's perspective this meant that the aim on any given trial was to find the hotspot and try to maintain the

position of the joystick over this region until having achieved 15 hits within 1 second.

Procedure

Participants sat at a desk in front of a joystick, a keyboard, and a 19-inch computer monitor. Before starting the experimental program, the task was briefly described verbally with the goal being phrased in terms of “finding the correct position to place the joystick in”. They were encouraged to move through the full range of joystick positions because pilot testing revealed a tendency of some people to explore only the very edges of the joystick’s travel and not central regions of the search space. Participants were told that the experiment involved no deception and that the correct position could always be found. Finally, participants were informed that if they were having especial difficulty completing a trial, they could press the space bar to abandon the current trial and move on to the next one. Following the brief verbal guidance, the program was started and the participants were asked to follow the onscreen instructions (see appendix 1). After reading the instructions, 3 practice trials commenced automatically. The practice trials involved no reinforcement delay and, as with all trials in the experiment, no feedback or screen graphics were provided during the trial (the monitor display was entirely black until the end of a trial, except for the screen flashes in the visual reinforcement condition). Once the practice trials were completed the

experimental trials began and participants were left to complete all 18 trials.

Design

Participants received either audio or visual reinforcement signals, and each reinforcement was presented after one of the 6 delay durations had elapsed from the moment of encountering the hotspot. So modality was a between-subjects factor (auditory, n = 27; visual, n = 24) and delay a within-subjects factor. Each experimental session was made up of 21 trials: 3 of which were practice trials (involving no delay); and 18 of which were experimental trials - 3 at each delay level. The delay conditions were ordered in 3 randomised batches of 6, such that all 6 levels of delay were experienced in each of the three batches. This was done to ensure that the 3 attempts at a particular delay condition were spread over the full testing session.

Results

Because each participant had 3 attempts at each delay condition, it was the mean of their successful attempts that was submitted to analysis. Due to the open-ended nature of trials, it was anticipated prior to testing that the data distributions would be non-normal with positive skew. Analysis of the distributions and inspection of the

frequencies confirmed this and all data were corrected using log-transformation prior to analysis (Keene, 1995). Analysis of variance (ANOVA) was used for the main analysis, with Bonferroni corrected, paired-sample t-tests being employed for post hoc comparisons. An alpha level of .05 was used to determine significance.

Two metrics of performance were identified for analysis. The first is distance travelled between the first entry into the target area and the successful achievement of the escape criterion. The second is the number of hits required to complete a trial. As Figure 3 shows, there was a significant effect of delay on distance travelled, $F_{GG}(4.24, 186.44) = 13.29$, $p < 0.001$. However, we found no effect of modality, $F(1, 44) = 0.051$, $p = 0.82$, and no interaction $F_{GG}(4.24, 186.44) = 0.79$, $p = .54$. Post hoc tests revealed that the 0 ms condition differed significantly from the 75 ms condition in both the audio, $t(26) = 3.11$, $p < 0.05$), and visual, $t(23) = 3.79$, $p < 0.05$ reinforcement conditions.

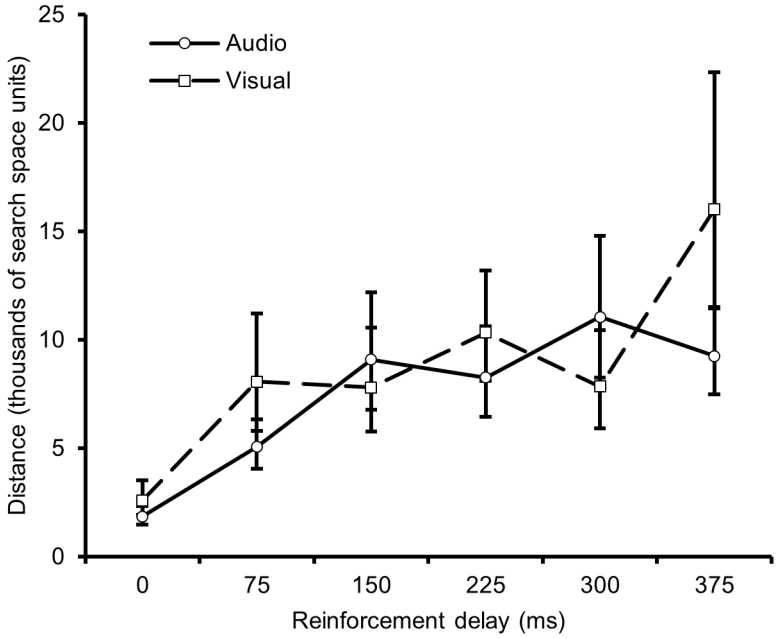

Figure 3. Mean distance (and standard error) for the 6 levels of delayed audio and visual reinforcement. Values are back-transformed from a log transformation.

There are some similarities between the metric of hits and the rate of response used in traditional instrumental learning paradigms; however, they differ in the important respect that trial completion in the current task was contingent on achieving a particular hit rate and thus hits does not represent a metric of choice, but rather a true metric of performance. The same pattern as found in the distance measure was found with hits. The effect of delay on performance was clearer, $F_{GG}(3.83, 168.68) = 27.71$, $p < 0.001$, and there was no effect of modality, $F(1, 44) = 0.021$, $p = 0.89$, or interaction, $F_{GG}(3.83, 168.68) = 0.55$, $p = 0.69$. Again, post hoc tests revealed that the 0 ms condition differed significantly from the 75 ms condition in both the audio, $t(26) = 4.13$, $p < 0.05$, and visual, $t(23) = 4.18$, $p < .005$, reinforcement conditions. Figure 4 shows that the effect of delay was to increase the number of hits during the post-discovery period. These results were reflected at the level of individuals, with 41 of 51 participants requiring more instances of reinforcement in the 75 ms condition than the 0 ms condition.

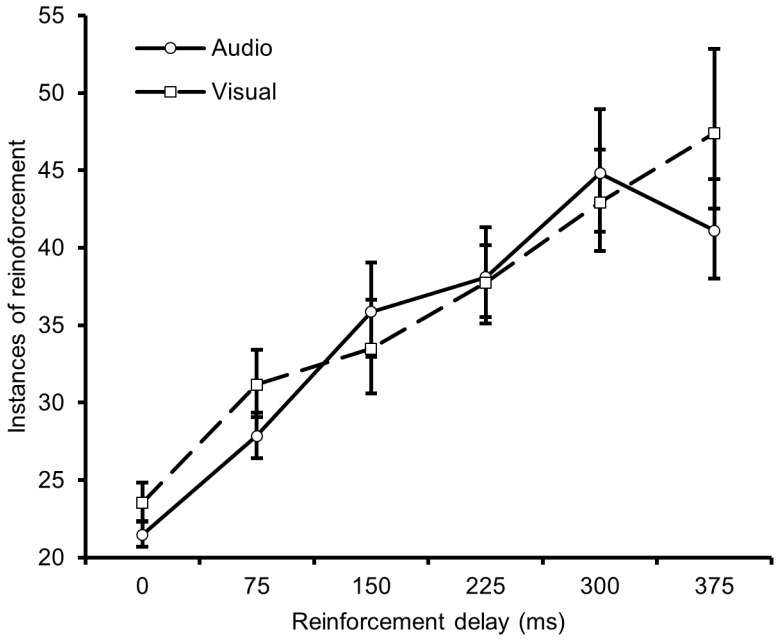

Figure 4. Mean instances of reinforcement (and standard error) for the 6 levels of delayed audio and visual reinforcement. Values are back-transformed from a log transformation.

Experiment 1 shows that action acquisition is sensitive to reinforcement delay of 75ms, a delay an order of magnitude below that previously reported. Two measures of efficiency of action discovery - distance travelled during search and number of reinforcements - show similar patterns across delay conditions. Together they suggest that at higher delays participants were more likely to move through the target area and return to it (perhaps multiple times) before being able to successfully identify its location.

Published as: Walton, T., Thirkettle, M., Redgrave, P., Gurney, K. N., & Stafford, T. (2013). The Discovery of Novel Actions Is Affected by Very Brief Reinforcement Delays and Reinforcement Modality. *Journal of Motor Behavior, 45*(4), 351-360.

Experiment 2

Despite finding a clear effect of delay in experiment 1, we were unable to detect an effect of reinforcement modality. Clearly not all delays experienced by animals are external and previous work investigating sensory integration and reaction times has shown that audio stimuli are processed somewhere between 40 and 100 ms faster than visual stimuli (Jaśkowski, Jaroszyk & Hojan-Jezierska, 1990; Lewald & Guski, 2003; Senkowski, Talsma, Grigutsch, Herrmann & Woldorff, 2007; Wallace, Wilkinson & Stein, 1996), due to differences in sensory processing latency. Should an internal sensory delay be equivalent to an external, artificial, delay then we would expect to see better performance with an auditory reinforcing signal. Furthermore, our hypothesis about the subcortical machinery responsible for action-outcome learning (Redgrave & Gurney, 2006), gives us cause to think that the action-learning process should be exquisitely sensitive to delays in reinforcing signal, regardless of their origin.

Whilst the results from experiment 1 might simply have reflected a lack of any underlying modality effect, the pattern of data shown in Figures 3 and 4 suggested to us that there might be an effect at the higher delay durations tested, with the audio condition levelling off and the visual condition continuing to decline in performance. We therefore repeated the experiment, this time varying modality within subjects and making changes to the stimulus properties and experimental constraints (see methods)

to increase the control over stimulus presentation in an attempt to increase the sensitivity of the task to differences in stimulus modality. Finally, we also introduced a time limit for the search behaviour in each trial to prevent anomalous failures to discover the target affecting the results and aggravating the participants.

Method

Participants

27 Sheffield psychology undergraduate students (7 male, mean age 19.2 years, *SD* = 2.8 years) participated in the study for course credit. All participants reported normal or corrected to normal vision and hearing and were naive to the purposes of the experiment.

Apparatus

The experimental program was again conducted within Matlab, and all displays were generated using a Cambridge Research Systems Visage graphics board which was in turn driving a calibrated Mitsubishi Diamond Pro 2070sb 22” monitor screen at 100Hz. This apparatus was used for stimulus calibration and presentation in the experimental tasks. A chin rest ensured the participants remained seated 57cm from the

screen throughout. Depending on the trial condition, encounters with the target resulted in either a 10 ms audio tone (600Hz) or a 10 ms onscreen visual signal. An additional instruction was included, asking participants to fixate on a 0.5º white cross in the centre of an otherwise blank screen during the task to ensure the signal would be visible. A fixation requirement was added to ensure consistent visual stimulation with each reinforcement, in an attempt to lessen variability across subjects and conditions. The visual stimulus was a 2.5º thick white annulus centred on fixation at 9.75º. This shape was chosen to ensure that it didn't give any unwanted and misleading location information. The fixation, instructions and reinforcing signal were all presented at maximum contrast on the screen (white on black), and the audio tone was presented loudly in an otherwise silent room to ensure both signals were presented far above the relevant sensory thresholds. The search space was defined as consisting of 1000 by 1000 units. The hotspot size and its random placement remained the same as in experiment 1. The joystick was also the same as that used in experiment 1.

Procedure

The learning criterion remained the same as in experiment 1, but a time constraint was imposed allowing participants 60 s to complete a trial. This rule was included to remove extremely long trials and possible unwanted effects caused by frustration, a particularly important factor due to the repeated measures design and

longer testing session. If a trial was not completed within the time limit it was stopped and the participant was required to repeat the trial. Participants were not given the opportunity to abandon trials.

Design

Participants completed two experimental blocks in a single session, one with audio reinforcement and the other with visual; the order of these blocks was counterbalanced, to avoid order effects, with a break in-between. Within each block participants performed 3 repetitions of trials where reinforcement presentation was delayed from successful hotspot encounter at one of 6 delay levels (0, 75, 150, 225, 300 and 375 ms). So both modality and delay were within subjects factors. In total each participant completed 36 trials in a single, 45 minute experimental session.

Results

As Figures 5 and 6 illustrate, we replicate our finding that delay between performance and reinforcement has a significant impact on action discovery, this is true for both post-discovery distance, $F(5,130) = 10.4$, $p < 0.001$, and hits, $F(5,130) = 28.2$, $p < 0.001$, although the shortest duration at which the effect was detectable was longer with post-hoc tests showing only 150 ms and above being significantly different from 0

ms when hits is used as the metric of performance (p <0 .05). We also found a significant effect of stimulus type on both measures of performance: hits, $F(5,130) = 28.2$, $p < 0.001$, and post-discovery distance, $F(1,26) = 6.77$, $p < 0.05$. In neither case was there a significant interaction between modality of reinforcement and delay (hits, $F(5,130) = 0.637$, $p > 0.6$; post-discovery distance, $F(5,130) = 0.293$, $p > 0.9$).

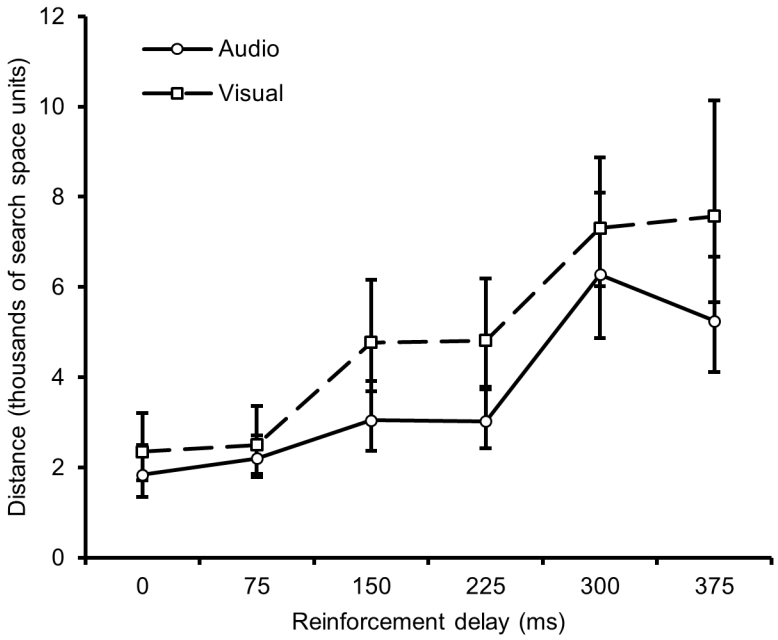

Figure 5. Mean distance (and standard error) for the 6 levels of delayed audio and visual reinforcement. Values are back-transformed from a log transformation. Task performance is better with an auditory rather than visual reinforcement signal. Delaying the reinforcing signal has a similar effect on both modalities.

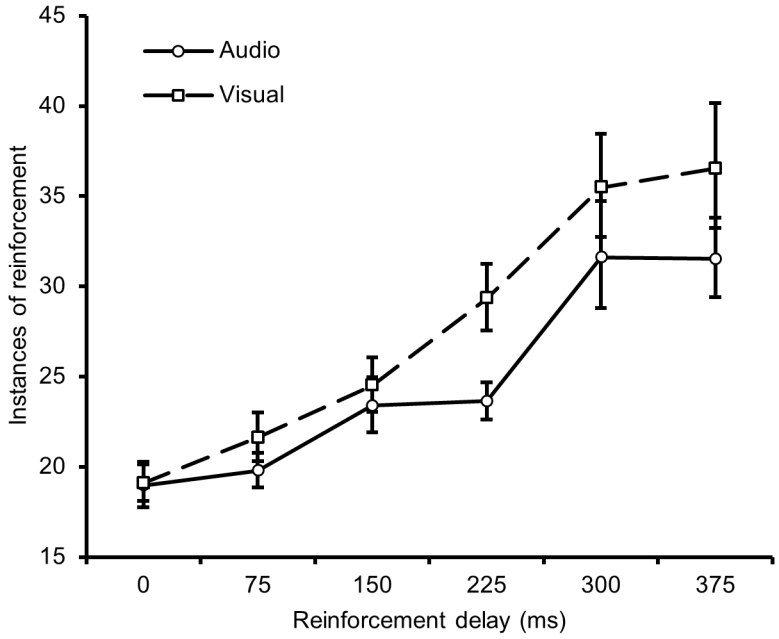

Figure 6. Mean instances of reinforcement (and standard error) for the 6 levels of delayed audio and visual reinforcement. Values are back-transformed from a log transformation. Task performance is better with an auditory rather than visual reinforcement signal. Delaying the reinforcing signal has a similar effect on both modalities.

Discussion

As predicted, enhanced stimulus control allowed a difference between

reinforcement modalities to show. Auditory reinforcement signals were consistently more effective for action discovery. One interpretation for this difference is the well known superiority of sensory transmission times for auditory compared to visual stimuli (Jaśkowski, Jaroszyk & Hojan-Jezierska, 1990; Lewald & Guski, 2003; Senkowski, Talsma, Grigutsch, Herrmann & Woldorff, 2007; Wallace, Wilkinson & Stein, 1996).

If auditory information passes to the neural circuitry necessary to encode the coincidence of motor output with stimuli information faster then it should promote binding of motor output to effects. If this was the case, it would further validate the basic sensitivity of the putative action discovery mechanism to small differences in delay.

Despite revealing an effect of modality, experiment 2 was slightly less sensitive to the effects of delay than experiment 1; delays of 150ms were significantly different from delays of 0ms but delays of 75ms did not reach significance. Possible reasons for this difference include a reduction in power, due to the testing of approximately half the number of participants in experiment 2 compared to experiment 1. Another possibility is that the introduction of a search time limit in experiment 2, which would have had the effect of preventing exceptionally long search periods, attenuated the effects of delay. In experiment 1 trials with very long search times (which did however result in eventual success) overwhelmingly occurred in trials with longer delay periods. However that the basic effect of delay persists in experiment 2 suggests that this effect is pervasive on all

aspects of action learning - not merely restricted to those trials for which learning the target takes an abnormally long time. The lower power of experiment 2 to detect an effect of delay could suggest, however, that a non-trivial portion of the effect of delay is manifest in the higher end of the search time distribution (a distribution which is curtailed by the cut-off used in experiment 2).

General Discussion

The current findings demonstrate an exquisite sensitivity to delays, both external and internal, of the mechanisms underpinning the learning of action-effect associations. Theoretical and neurobiological analysis (Redgrave & Gurney, 2006) suggests that such delay sensitivity is to be expected. Previous work on delay in reinforcement found less sensitivity to delay because it focussed on response rate as the variable of interest. Adjusting the rate of a particular, pre-learnt, response is more robust to temporal delay as the relevant motor output has already been determined and so should be robust to contamination within an eligibility period. Colloquially, if you already have representations of the action and the outcomes, the association between them is less sensitive to delays.

Evidence that delay affects learning has a long history in the study of behaviour (Hull, 1943; Grice, 1948). Although we don't see our task as fitting the mold of previous experimental studies of response rate learning, there is an analogue on the issue of contamination of the motor record. Within studies of delayed conditioning, this was addressed by interference theory (Revusky, 1971, but see Lieberman, McIntosh and Thomas, 1979). Debate has revolved around the issue of whether learning with long delays is impaired by the mere fact of delay (i.e. that memory decays inexorably with the passage of time), or whether it is the presence of intervening actions or stimuli

which makes learning temporally separated associations difficult. The latter option is what we could call contamination. The work reported here does not address this issue, although the paradigm is certainly amenable to its study.

Contrast with existing paradigms of motor learning

Previous studies of the effect of delay have focussed on response rate as an index of response acquisition. We have argued (above, and in Stafford et al, 2012), response rate is not an appropriate index for the process(es) we are interested in - those of the initial stages of action acquisition. Those studies that there are which study the effect of delay on response acquisition report a relative insensitivity to delay, with estimates of the delay at which acquisition is impaired ranging up from 1 second (Black et al., 1985) to tens of seconds (Snycerski et al., 2005; Lattal & Gleeson, 1990).

In contrast to this literature, our experiments show a sensitivity to delay of action acquisition which is on the order of tens or hundreds of milliseconds. This level of sensitivity aligns with that reported for some motor learning paradigms. For example, in visual tracking tasks, feedback delays as short as 300 ms can have large effects on

performance (Foulkes & Miall, 2000; Miall & Jackson, 2006; Miall, Weir, & Stein, 1985) and tasks involving adaptation to visually displaced targets have been shown to be equally sensitive (Held, Efstathiou, & Greene, 1966), with some demonstrating an effect at just 50 ms (Kitazawa, Kohno & Uka, 1995).

A defining feature of these motor learning tasks is the demand they place on motor accuracy. These are unlike response rate tasks, which commonly use button-pressing actions, where no emphasis is placed upon accuracy. There is no special reason why this should be the case. All behavioural tasks rely on some element of motor control. Studies which focus on response rate neglect the problem of motor control in favour of the problem of learning the value of actions. Many studies of motor control, such as tracking tasks, neglect the problem of action valuation in favour of the problem of motor control.

Our task has a motor control demand, but ultimately is more akin to response acquisition tasks than motor control tasks. Feedback on performance is contingent on finding the target, making reinforcement learning the most appropriate computational framework for understanding the task (Barto & Dietterich, 2004). In contrast, it is common for some motor learning paradigms to have feedback is provided regardless of the success of the movement (e.g., as cited above, Foulkes & Miall, 2000; Miall & Jackson, 2006; Miall, Weir, & Stein, 1985; Held, Efstathiou, & Greene, 1966; Kitazawa,

Kohno & Uka, 1995). In addition studies of the role of feedback in motor learning, such as in “knowledge of results” paradigms (Swinnen, Schmidt, Nicholson, & Shapiro, 1990) also indicates an effect of delay on learning, albeit strikingly different from the one presented here. Leukel & Jenson (2013) provide a recent review of the area, which for reasons of space we cannot explore further here. These paradigms, while related, are best understood using a supervised learning framework (Wolpert, Ghahramani & Flanagan, 2001).

Learning with delay

Thorndike spoke of task-relevant actions being gradually stamped in as an animal practiced a particular set of behaviours and this idea has survived with the concept of neural time stamps. If time stamping is used in solving the credit assignment problem it is appropriate to ask how it is implemented in the mammalian brain. Dopamine neurons in the ventral midbrain of vertebrates are known to respond to novel and rewarding stimuli with a latency of approximately 70-100 ms (Bayer & Glimcher, 2005; Guarraci & Kapp, 1999; Horvitz, Stewart & Jacobs, 1997; Ravel & Richmond 2006; Schultz, 1998; Takikawa, Kawagoe & Hikosaka, 2004) and it is widely thought that this neural activity plays a key role in valuation and economic decision-making (Schultz, 2007). However, Redgrave and colleagues (Redgrave & Gurney, 2006;

Redgrave et al., 2008, Redgrave, Prescott & Gurney, 1999) have argued that the speed of this response affords insufficient time for the brain to process identifying characteristics of the stimuli, including those indicating relative reward value. Consequently, they suggest that, instead, this neural signal is an indiscriminate time-stamp which indicates the last segment of the animal's motor record that could have played a role in eliciting a novel stimulus, irrespective of what that stimulus might be. In this way, they propose that the dopamine response is central to the tasks of agency detection, action discovery and the learning of action-effect contingencies.

The proposed advantage of time-stamping so quickly, even before the event has been fully processed and identified, is that it reduces the opportunity for further non-contingent behaviour to be expressed, reducing the credit assignment problem that the animal has to solve. Indeed, Redgrave and Gurney (2006) and Redgrave et al. (2008) have speculated that this is why artificial delays are so detrimental to reinforcement learning. The current results demonstrate that delays of similar duration to the latency of the dopamine timestamp itself can have a significant detrimental impact on action discovery. The importance of time-stamping explains existing results which demonstrate that informative signals which do not provide additional reward can promote response acquisition in rats (Reed, Schactman and Hall, 1991; Reid, Chadwick, Dunham and Miller, 2001; Reid, Nill and Getz, 2010). Schaal and Branch (1988) show that marking signalling can restore decrements in response rate produced by reinforcement delay.

Such marking stimuli can also increase judgements of causality in humans under conditions of delay (Reed, 1992; Reed 1999).

Theoretical analyses have suggested for a long time that the credit assignment problem is a major issue for mechanisms of action learning. Our paradigm which is designed specifically to look at the issue of action acquisition rather than response rate valuation has shown action acquisition to be impaired by delays smaller than those previously reported. This behavKohno & Uka, 1995). In addition studies of the role of feedback in motor learning, such as in “knowledge of results” paradigms (Swinnen, Schmidt, Nicholson, & Shapiro, 1990) also indicates an effect of delay on learning, albeit strikingly different from the one presented here. Leukel & Jenson (2013) provide a recent review of the area, which for reasons of space we cannot explore further here. These paradigms, while related, are best understood using a supervised learning framework (Wolpert, Ghahramani & Flanagan, 2001).ioural evidence aligns with, but does not confirm the function of the short-latency dopaminergic signal proposed by Redgrave and Gurney (2006). What it does suggest is that time-stamping of some form, as indicated by a crucial sensitivity to delay, lies at the heart of action discovery.

References

Barto, A.G. and Dietterich, T.G. (2004). Reinforcement Learning and Its Relationship to Supervised Learning In Si, J., Barto, A.G., Powell, W.B., and Wunsch, D., (eds.), *Handbook of Learning and Approximate Dynamic Programming*, Chapter 2, pp 47 - 64. Wiley-IEEE Press, Piscataway, NJ.

Barto, A. G., Sutton, R. S., & Brouwer, P. S. (1981). Associative search network: a reinforcement learning associative memory. *Biological Cybernetics, 40*(3), 201-211.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47*(1), 129-141.

Black, J., Belluzzi, J. D., & Stein, L. (1985). Reinforcement delay of one second severely impairs acquisition of brain self-stimulation. *Brain Research, 359*(1-2), 113–119.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial vision, 10*(4), 433-436.

Dickinson, A., Watt, A., & Griffiths, W. (1992). Free-operant acquisition with delayed reinforcement. *The Quarterly Journal of Experimental Psychology Section B, 45*(3), 241-258.

Elsner, B., & Hommel, B. (2004). Contiguity and contingency in action-effect learning.

Psychological Research, 68(2-3), 138-154.

Foulkes, A. J., & Miall, R. C. (2000). Adaptation to visual feedback delays in a human manual tracking task. *Experimental Brain Research, 131*(1), 101-110.

Ghahramani, Z. (2004). Unsupervised learning. In O. Bousquet, U. Luxburg, & G. Rätsch (Eds.), *Advanced lectures on machine learning* (Vol. 3176, pp. 72-112). Berlin, Heidelberg: Springer Berlin Heidelberg.

Grice, G. R. (1948). The relation of secondary reinforcement to delayed reward in visual discrimination learning. *Journal of Experimental Psychology, 38*(1), 1-16.

Guarraci, F. A., & Kapp, B. S. (1999). An electrophysiological characterization of ventral tegmental area dopaminergic neurons during differential pavlovian fear conditioning in the awake rabbit. *Behavioural Brain Research, 99*(2), 169-179.

Hachiya, S., & Ito, M. (1991). Effects of discrete-trial and free-operant procedures on the acquisition and maintenance of successive discrimination in rats. *Journal of the Experimental Analysis of Behavior, 55*(1), 3-10.

Held, R., Efstathiou, A., & Greene, M. (1966). Adaptation to displaced and delayed visual feedback from the hand. *Journal of Experimental Psychology, 72*, 887-891.

Hill, C. J. (1939). Goal gradient, anticipation, and perseveration in compound trial-and-error learning. *Journal of Experimental Psychology, 25*(6), 566-585.

Horvitz, J. C., Stewart, T. & Jacobs, B. L. (1997). Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Research, 759*(2), 251-258.

Hull, C.L. (1943). *Principles of behavior*. New York: Appleton-Century.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signalling. *Cerebral Cortex, 17*(10), 2443-2452.

Jaśkowski, P., Jaroszyk, F., & Hojan-Jezierska, D. (1990). Temporal-order judgments and reaction time for stimuli of different modalities. *Psychological Research, 52*(1), 35–8.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: supervised learning with a distal teacher. *Cognitive Science, 16*(3), 307-354.

Keene, O. N. (1995). The log transformation is special. *Statistics in Medicine, 14*(8), 811-819.

Kitazawa, S., Kohno, T., & Uka, T. (1995). Effects of delayed visual information on the rate and amount of prism adaptation in the human. *The Journal of Neuroscience, 15*(11), 7644-7652.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3? *Perception, 36*(ECVP Abstract Supplement).

Lattal, K. A., & Gleeson, S. (1990). Response acquisition with delayed reinforcement. *Journal of experimental Psychology: Animal Behavior Processes, 16*(1), 27-39.

Lattal, K. A., & Metzger, B. (1994). Response acquisition by Siamese fighting fish (Betta splendens) with delayed visual reinforcement. *Journal of the Experimental Analysis of Behavior, 61*(1), 35-44.

Leukel, C., & Jensen, J. L. (2013). The role of augmented feedback in human motor

learning. In Gollhofer, A., Taube, W., & Nielsen, J. B. (Eds.). *Routledge handbook of motor control and motor learning*. Routledge.

Lewald, J., & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research, 16*(3), 468–78.

Lieberman, D. A., McIntosh, D. C., & Thomas, G. V. (1979). Learning when reward is delayed: A marking hypothesis. *Journal of Experimental psychology: Animal behavior processes, 5*(3), 224-242.

Miall, R. C., & Jackson, J. K. (2006). Adaptation to visual feedback delays in manual tracking: evidence against the Smith Predictor model of human visually guided action. *Experimental Brain Research, 172*(1), 77-84.

Miall, R. C., Weir, D. J., & Stein, J. F. (1985). Visuomotor tracking with delayed visual feedback. *Neuroscience, 16*(3), 511-520.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), 8-30.

Morris, R. G. M. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation, 12*(2), 239–260.

Okouchi, H. (2009). Response acquisition by humans with delayed reinforcement. Journal *of the Experimental Analysis of Behavior, 91*(3), 377-390.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*(4), 437-442.

Ravel, S., & Richmond, B. J. (2006). Dopamine neuronal responses in monkeys performing visually cued reward schedules. *The European Journal of Neuroscience, 24*(1), 277-290.

Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience, 7*(12), 967-975.

Redgrave, P., Gurney, K., & Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Research Reviews, 58*(2), 322-339.

Redgrave, P., Prescott, T.J. & Gurney, K. (1999), Is the short latency dopamine burst too short to signal reinforcement error? *Trends in Neurosciences, 22*, 146-151.

Reed, P., Schachtman, T. R., & Hall, G. (1991). Effect of signaled reinforcement on the formation of behavioral units. *Journal of Experimental Psychology: Animal Behavior Processes, 17*(4), 475-485. doi:10.1037/0097-7403.17.4.475

Reed, P. (1992). Effect of a signalled delay between an action and outcome on human judgement of causality. *The Quarterly Journal of Experimental Psychology Section B, 44*(2), 81-100. doi:10.1080/02724999208250604

Reed, P. (1999). Role of a stimulus filling an action-outcome delay in human judgments of causal effectiveness. *Journal of Experimental Psychology: Animal Behavior Processes, 25*(1), 92--102. doi:10.1037/0097-7403.25.1.92

Reid, A. K., Chadwick, C. Z., Dunham, M., & Miller, A. (2001). The development of functional response units: the role of demarcating stimuli. *Journal of the Experimental Analysis of Behavior, 76*(3), 303–320. doi:10.1901/jeab.2001.76-

303

Reid, A. K., Nill, C. A., & Getz, B. R. (2010). Changes in stimulus control during guided skill learning in rats. *Behavioural Processes, 84*(1), 511–515. doi:10.1016/j.beproc.2010.01.001

Revusky S. (1971) The role of interference in association over a delay. In *Animal Memory* (eds Honig W. K. and James P. H. R.), pp. 155–213. Academic, New York

Sanders, A. F. (1998). *Elements of human performance: reaction processes and attention in human skill*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Schaal, D. W., & Branch, M. N. (1988). Responding of pigeons under variable-interval schedules of unsignaled, briefly signaled, and completely signaled delays to reinforcement. *Journal of the Experimental Analysis of Behavior, 50*(1), 33-54. doi:10.1901/jeab.1988.50-33

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80*(1), 1-27.

Schultz, W. (2007). Behavioral dopamine signals. *Trends in Neurosciences, 30*(5), 203-210.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593-1599.

Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., & Woldorff, M. G. (2007). Good times for multisensory integration: Effects of the precision of temporal

synchrony as revealed by gamma-band oscillations. *Neuropsychologia, 45*(3), 561–71.

Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory and Cognition, 19*(4), 353-360.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology. B., 41*(2), 139-159.

Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning, 22*(1-3), 123–158.

Snycerski, S., Laraway, S., Huitema, B. E., & Poling, A. (2004). The effects of behavioural history on response acquisition with immediate and delayed reinforcement. *Journal of the Experimental Analysis of Behavior, 81*(1), 51-64.

Snycerski, S., Laraway, S., & Poling, A. (2005). Response acquisition with immediate and delayed conditioned reinforcement. *Behavioural Processes, 68*(1), 1-11.

Stafford, T., Thirkettle, M., Walton, T., Vautrelle, N., Hetherington, L., Port, M., Gurney, K., et al. (2012). A novel task for the investigation of action acquisition. *PloS One, 7*(6), e37749.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge: MIT Press.

Swinnen, S. P., Schmidt, R. A., Nicholson, D. E., & Shapiro, D. C. (1990). Information

feedback for skill acquisition: Instantaneous knowledge of results degrades learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 706-716.

Takikawa, Y., Kawagoe, R., & Hikosaka, O. (2004). A possible role of midbrain dopamine neurons in short- and long-term adaptation of saccades to position-reward mapping. *Journal of Neurophysiology, 92*(4), 2520-2529.

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

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*(4), 189-208.

van Haaren, F. (1992). Response acquisition with fixed and variable resetting delays of reinforcement in male and female Wistar rats. *Physiology and Behavior, 52*(4), 767-772.

Wallace, M. T., Wilkinson, L. K., & Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. *Journal of Neurophysiology, 76*(2), 1246–1266.

Wickens, J. (1990). Striatal dopamine in motor activation and reward-mediated learning: steps towards a unifying model. *Journal of Neural Transmission, 80*(1), 9-31.

Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences, 5*(11), 487-494.