Learn Behav (2011) 39:245–258
DOI 10.3758/s13420-011-0025-7

# Pigeon and human performance in a multi-armed bandit task in response to changes in variable interval schedules

**Deborah Racey · Michael E. Young · Dennis Garlick · Jennifer Ngoc-Minh Pham · Aaron P. Blaisdell**

Published online: 6 March 2011


**Abstract** The tension between exploitation of the best options and exploration of alternatives is a ubiquitous problem that all organisms face. To examine this trade-off across species, pigeons and people were trained on an eight-armed bandit task in which the options were rewarded on a variable interval (VI) schedule. At regular intervals, each option's VI changed, thus encouraging dynamic increases in exploration in response to these anticipated changes. Both species showed sensitivity to the payoffs that was often well modeled by Luce's (1963) decision rule. For pigeons, exploration of alternative options was driven by experienced changes in the payoff schedules, not the beginning of a new session, even though each session signaled a new schedule. In contrast, people quickly learned to explore in response to signaled changes in the payoffs.

**Keywords** Pigeon · Human learning · Associative learning · Acquisition

Direct interaction with the environment provides much of the information that informs subsequent actions. Rarely is choice made in the presence of perfect knowledge. In a multitude of domains, organisms begin by choosing almost blindly; what is learned about the environment varies according to which of the possibilities are experienced. The world often fails to reveal information about the utility of options not chosen—the route not taken, the career not selected, the product not purchased (Taleb, 2007). In a complex environment where options are many and/or variable, complete knowledge of prevailing contingencies may require very long-term exploration. Even after long experience with the prevailing contingencies, continued exploration of options with less utility may be necessary in order to adapt to change.

Under similar conditions, what leads some choosers to exploit their knowledge of differential utility and others to explore their options? Continued exploration may be an adaptive behavior learned through experience with changing environments (Rakow & Miler, 2009; Stahlman, Roberts, & Blaisdell 2010; Stahlman, Young, & Blaisdell 2010), or it may be that imperfect knowledge maintains exploration so that responding to changing conditions is a side effect rather than an adaptation. A complete study of the trade-off between exploration and exploitation will require the use of choice environments in which more than two options are available (cf. Rakow & Miler, 2009). We examined this trade-off in the present project by investigating human and pigeon behavior in an eight-option task.

In addition to contending with the real-world complexity related to large numbers of options, most species live in changing environments. Although researchers in foraging behavior have investigated decision-making mainly through familiar, stationary environments, such that the individuals are fully informed about the nature of the options (e.g., Lin & Batzli, 2002; Zach, 1979), there is increasing interest in how such information is acquired (e.g., Mettke-Hofmann, Wink, Winkler, & Leisler 2004; Plowright & Shettleworth, 1990). The introduction of environmental changes has often been used to study how animals gather information about their environment.

D. Racey
Western Carolina University,
Cullowhee, NC, USA

M. E. Young (✉)
Southern Illinois University at Carbondale,
Carbondale, IL, USA
e-mail: meyoung@siu.edu

D. Garlick · J. Ngoc-Minh Pham · A. P. Blaisdell
University of California at Los Angeles,
Los Angeles, CA, USA

We took an approach that was inspired by the study of reinforcement-learning algorithms as applied to machine learning (Koulouriotis & Xanthopoulos, 2008; Sutton & Barto, 1998). In its simplest form, reinforcement-learning analyses often use the multi-armed (or "*n*-armed") bandit task to evaluate various methods of distributing exploration and exploitation (e.g., Dimitrakakis & Lagoudakis, 2008; Sikora, 2008). This task provides an excellent platform to explore choice in stationary (with unchanging payoffs) and nonstationary (with changing payoffs) environments, and it has also been applied to the domains of human learning and cognition (e.g., Burns, Lee, & Vickers 2006; Plowright & Shettleworth, 1990), economics (e.g., Banks, Olson, & Porter 1997), marketing and management (e.g., Azoulay-Schwartz, Kraus, & Wilkenfeld 2004; Valsecchi, 2003), and math and computer science (e.g., Auer, Cesa-Bianchi, Freund, & Schapire 1995; Koulouriotis & Xanthopoulos, 2008).

## Task description

The multi-armed bandit task (MABT) usually involves choosing among multiple possible actions that lead to immediate reward and about which nothing is initially known. The MABT took its name from the "one-armed bandit," another term for the slot machine. Rather than the one arm of a slot machine, however, a MABT has *n* options. It can be thought of as a set of *n* slot machines, each with an independent payoff schedule. After each selection, the reinforcer is awarded based on an underlying schedule of reinforcement. A player must explicitly explore an environment in order to learn the expected payoffs for these *n* options, and then can later exploit this knowledge. In a four-armed bandit task similar to the one used in the present study, Steyvers, Lee, and Wagenmakers (2009) employed a Bayesian optimal-decision model derived from the softmax equation (Luce, 1963) to explore how humans balance exploration with exploitation. In addition, eight-stimulus arrays very similar to the one used in the present study have been used with nonhuman animals (Jensen, Miller, & Neuringer 2006) and humans (Rothstein, Jensen, & Neuringer 2008), and in both cases behavior came under the control of the prevailing contingencies. Thus, this MABT provides a decision task that is potentially both complex and challenging, yet at the same time simple enough that it can be used to study a wide range of decision-making in both humans and other animals.

## Exploration versus exploitation

An arm pull is an action, and at any point an actor is expected to rely on an estimate of action values based on the sampling history with each option. Choosing the action with the highest estimated action value (the "greedy" action) is exploitation, because the actor is exploiting its current knowledge. If the actor chooses a nongreedy action, it is exploring—a behavior that potentially enhances overall knowledge by improving the estimate of a nongreedy option. Greedy actions allow the actor to maximize its chance of immediate reward for the very next action, but nongreedy actions may be preferable, in order to maximize long-term reward or value (i.e., they actually are greedy, but over an extended time horizon).

Reward may be lower in the short term when exploring, but long-term value may be greater, since the actor may discover actions that are better than the current greedy action or that provide viable alternatives if the action with the long-run highest value is currently less profitable (due to molecular aspects of the payoff schedule in which an option's value is temporarily lower; e.g., for VI schedules) or later becomes unprofitable (due to molar changes in the payoff schedule; e.g., changing from a variable ratio 5 to variable ratio 50). Whether exploration or exploitation is best at any given choice point will depend on the expected changes in these payoffs, inter alia. For a nonstationary bandit task, option values change during the task by changing the underlying molar contingencies—as if the room full of slot machines were reprogrammed occasionally during the allotted time of play. Continued exploration is critical if an organism is to track and adapt to these changes.

The machine-learning literature provides some guidance regarding methods for action selection appropriate to the bandit task. The greedy strategy may be used to solve stationary bandit problems, and it requires that every response be made to the option with the highest value (i.e., the richest reinforcement schedule). This strategy results in quick and complete preference for one option, which is precisely what should be avoided in a nonstationary environment.

Alternatively, Luce's (1963) decision rule (often called *softmax*) describes a strategy that uses the expected rewards of the options to choose them probabilistically. In other words, it assigns the highest selection probability to the greedy option, but the rest of the remaining options are chosen according to their value estimates. The probability of choosing action $a$ is

$$P(action_a) = \frac{e^{\theta \cdot value_a}}{\sum_{j=1}^{n} e^{\theta \cdot value_j}}, \quad (1)$$

where $\theta$ is the exploitation parameter, $value_i$ denotes the current estimated value for the $i$th action, and $n$ is the number of possible actions. When $\theta$ is zero, exploitation is absent and exploration of alternatives is predicted to be maximal, such that each action is equiprobable. Higher values of $\theta$ result in

higher levels of exploitation; the option with the highest action value (the greedy response) is selected more frequently as $\theta$ increases. At very high levels of $\theta$, Luce's decision rule becomes indistinguishable from the greedy decision strategy. The inclusion of the $\theta$ parameter allows for adjustment of the levels of exploitation and exploration to describe a particular organism's behavior, depending on variables such as time, satiety, environmental uncertainty, and use by each species and subject and at each stage in learning.

People and pigeons are not Turing machines, and their estimates of action values may be imperfect. Regardless, these action values may be based simply on an overall history with each option, such as the proportion of total responses to that option that have been reinforced, or by some more complex calculation. For example, these estimates may be weighted to more recent experience or sensitive to the changes in reinforcement probability over time that are inherent in VI schedules. For this study, we assumed these action values to be equal to the overall programmed likelihood of reinforcement represented by the VI schedule for each option. Thus, we operationally defined exploration as choosing a response that has a lower molar reinforcement rate.

The present experiments examined both pigeon and human performance using a nonstationary MABT. Each species chose from among eight response options in order to provide a complex set of choices that would constrain the theoretical analysis. We were interested in testing three hypotheses. First, could Luce's decision rule be used to assess the balance between exploitation and exploration for pigeons and humans in our choice task? Second, would both species adaptively and quickly increase their level of exploratory behavior in response to environmental cues that signal a change in choice payoffs? For pigeons, each daily session began with a new set of choice payoffs, and thus an adaptively optimal pigeon would begin each session with maximal exploration and be unaffected by the previous day's programmed schedules. For people, a new session began every few minutes and was signaled by a discriminative cue at the top of the display that should prompt a sudden increase in exploration. Third, would exploration continue throughout a session, or would pigeons and people exhibit a higher level of exploitation later in the session, once differential choice value had been determined?

## Experiment 1

### Method

#### *Subjects*

A total of 6 experimentally naïve adult White Carneaux pigeons (*Columba livia*) participated in the experiment. The pigeons were individually housed in steel home cages with metal wire mesh floors in a vivarium, and a 12-h light:dark cycle was maintained. Testing was conducted 5–7 days/week during the light cycle. The pigeons were maintained at approximately 85% of their free-feeding weights, and were given free access to grit and water while in their home cages.

#### *Apparatus*

Testing was conducted in a flat-black Plexiglas chamber (38 cm wide × 36 cm deep × 38 cm high). All stimuli were presented by computer on a color LCD monitor (NEC MultiSync LCD1550M) visible through a 23.2 × 30.5 cm viewing window in the middle of the front panel of the chamber. Pecks to the monitor were detected by an infrared touch screen (Carroll Touch, Elotouch Systems, Fremont, CA) mounted on the front panel. A 28-V houselight located in the ceiling of the box was used for illumination, except during time outs. A food hopper (Coulbourn Instruments, Allentown, PA) was located below the monitor with an access hole situated flush with the floor. All experimental events were controlled and data recorded by a Pentium III class computer (Dell, Austin, TX). A video card controlled the monitor using the SVGA graphics mode (800 × 600 pixels).

#### *Procedure*

*Preliminary training* The 6 pigeons were first trained to eat from the hopper in the chamber. Next, responses were autoshaped to a white disk that appeared in the center of the screen. Pecking to the disk resulted in the hopper rising for 3 s before lowering again. This was followed by a 60-s intertrial interval (ITI) before the next disk was displayed. Once the pigeon was consistently responding to the disk, training began.

*Bandit training* The pigeons were presented with differently colored disks on the screen, with each disk approximately 2 cm in diameter. The disks were arranged in a circular array starting at the top of the screen, such that disks that were opposite each other were approximately 8 cm apart (see Fig. 1). This display was located so that the bottom of the lowest disk was 3 cm above the bottom edge of the screen. The colors used for the disks, from the left clockwise, were gray, light blue, red, yellow, pink, green, dark blue, and orange. The reward value given to a particular disk was fixed throughout the session, but the reward values were randomly redistributed across disks from one session to the next. Thus, in one session the values assigned to disks clockwise from the top may have been 6, 192, 12, 3, 384, 24, 48, and 96, but the distribution on the following session may have been 12, 192, 48, 6, 96, 384, 3, and 24. This redistribution of values was done at the

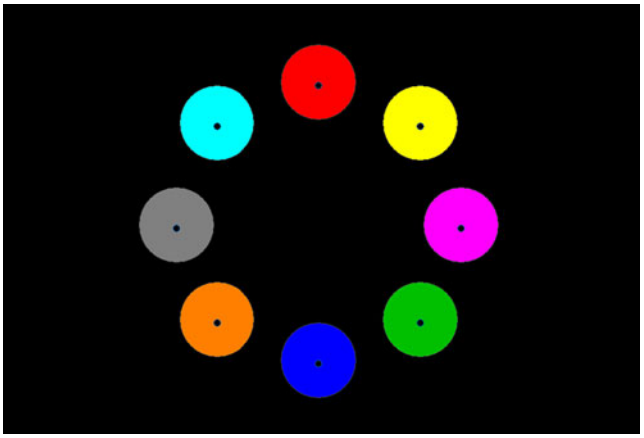


**Fig. 1** The computer screen as it was presented to the pigeons. The disks were identified in the analysis by consecutive numbers in a clockwise direction, with the top disk being 0

beginning of each session. The relative positions of the colors were not changed from session to session. Throughout training and testing, sessions were 60 min long.

Initial training consisted of assigning random ratios (RRs) to the disks, using the following probabilities of each peck being rewarded: .61, .37, .22, .14, .08, .05, .03, and .02. After 60 sessions of training, it became clear that the pigeons were showing strong biases to disks located in particular positions and were not pecking to disks located in other positions, even if they had the highest reward value.

One possibility was that pecking to the disks was relatively cheap to the pigeons, so the difference in reward structure was not very tangible. Another factor was that pigeons tend toward maximization (i.e., high exploitation) on RR schedules by showing nearly exclusive responding for the option with the richest experienced payoff structure (Herrnstein & Loveland, 1975). To increase sensitivity to reward and to encourage exploration by temporarily decreasing the reward value of a disk, the reward structure was changed from a random ratio to a variable-interval schedule. The variable intervals used were 3, 6, 12, 24, 48, 96, 192, and 384 s and varied by up to ±50% of the scheduled interval (e.g., for VI 3, the interval varied between 1.5 and 4.5 s). After another 60 sessions, it was clear that the pigeons were still showing strong biases to disks located in particular positions. Shifting the color assignments revealed that the bias was based on location and not color.

The pigeons completed 5 sessions in which only one disk from the display was shown, and the disk had a .61 probability of reward. In this situation, the pigeons did reliably peck to the disk, regardless of its color or position.

The pigeons then completed 40 sessions with all eight disks present, one of which had a .61 probability of reward and seven of which had no reward. The pigeons still showed a strong bias to particular disk locations, even if the locations were not associated with reward in a given session.

A final attempt to equalize the perceived reward value of the disks and encourage exploration involved presenting the pigeons again with all eight disks for 24 sessions. However, the reward schedule was made more extreme, with VIs of 3, 9, 27, 81, 243, 729, 2,187, and 6,561 s (with experienced intervals again varying up to ±50% of the scheduled interval). In addition, at the end of the 24 sessions, the disk that was most pecked was eliminated. For the subsequent 24 sessions, only the remaining seven disks were present, with the longest reward interval was not assigned to a disk. At the end of this set of 24 sessions, the most pecked disk was again eliminated along with the longest reward interval still being used. This procedure progressed until the pigeons were given 24 sessions with only the three least-pecked (by location) disks remaining. To keep the pigeons at 85% of free-feeding weight, a session was terminated once 300 rewards had been received during the session.

*Testing* For the test sessions, the pigeons were presented with all eight disks for 24 sessions with VIs of 3, 9, 27, 81, 243, 729, 2,187, and 6,561 s. Assignment of VI schedule to the disks varied daily. Only the data from this final set of testing sessions were analyzed.

## Results

To analyze the data, we used two approaches. First, we will describe the frequency with which each disk was chosen as a function of its programmed payoff. This approach will provide a general assessment of the degree of control established by the reward structure. Second, we will provide an analytical assessment of the pigeons' exploratory behavior using Luce's decision rule (Luce, 1963).

From a reinforcement-learning perspective, low $\theta$ values indicate that a chooser either has not learned the differential payoffs or has maintained high exploration despite the differential payoffs. However, a sudden decrease in $\theta$ (when responding is not a function of previous disk value) indicates that a chooser has recognized that the payoffs have changed, thus prompting an increase in exploratory behavior.

The complicating factor in our analysis is that the programmed contingencies may not have been experienced equally by every organism. A pigeon may have undersampled a particular choice and thus obtained a biased estimate of its payoff. Pigeons frequently showed disk biases and failed to fully explore each of the options. Thus, in our second set of choice analyses for pigeons, we used disk location as an independent predictor of the best-fitting

$\theta$ values and predicted lower $\theta$s (i.e., poor response differentiation as a function of payoff value) for less-preferred disks.

To estimate behavioral differentiation, we used the following instantiation of Luce's decision rule:

$$P(key_i) = \frac{e^{\theta \cdot payoff_i}}{\sum_{j=1}^{8} e^{\theta \cdot payoff_j}}, \quad (2)$$

in which $payoff_i$ is the logarithm of the inverse of the programmed VI. The equation generates eight probabilities, one for each of the eight disks, that sum to 1.0.

To fit Luce's decision rule to behavior, we used nonlinear mixed-effects modeling and identified the maximum likelihood best-fitting parameter values (Cudeck & Harring, 2007; Davidian & Giltinan, 2003). Mixed-effects modeling is used to simultaneously generate estimates of parameter estimates for each subject and as a function of the independent variables (e.g., Laird & Ware, 1982; Pinheiro & Bates, 2004). This approach is superior to the two-stage approach, in which parameter estimates are derived independently for each subject and the estimates are used in a subsequent analysis, because the results of the first stage do not include information about uncertainty in the parameter estimates that are used in the second stage (Shkedy, Straetemans, & Molenberghs 2005).

We examined changes in the maximum likelihood for $\theta$ in Eq. 1 across birds (random effect) as a function of our predictors (fixed effects). To apply Luce's decision rule, we needed to identify the best proxy for disk value (i.e., payoff). Preliminary analyses identified that an appropriate function mapping VI to value was the logarithm of the reinforcement rate (1/VI). The inverse translates the VI into an expected rate, so that higher values are associated with better schedules (Fig. 2 reveals that this transformed variable is a good proxy for the relative long-run probability that the pigeon was rewarded for choosing that disk). The log transformation produced a stronger fit than the untransformed reinforcement rates.

## Choice differentiation as a function of programmed payoffs

When we examined the proportion of trials on which each disk was chosen by each pigeon, the pigeons showed a marked preference for disks with the richest programmed VI schedules (see Fig. 3, solid lines). One pigeon, Cosmo, showed a strong preference for the disk with the second-best payoff schedule. A closer examination of the pigeons' disk choices, however, revealed that despite our attempts to train out disk biases, the pigeons still showed general preferences for disks in the lower part of the display (see the peck location density plots shown in Fig. 4; these plots were produced using JMP's nonparametric bivariate density function; SAS Institute Inc., Cary, NC). For some pigeons, certain disks were so rarely sampled that these choices are not visible in our density plots. When these less-preferred disks were associated with high payoffs for a particular session, the pigeon rarely experienced the high value of these disks.

As a baseline of comparison, we initially ignored these disk biases and identified the best-fitting $\theta$ for Eq. 1 [using log (1/VI) as a proxy for payoff rate] as a function of 5-min trial block (1–12). The analysis revealed that the degree of response differentiation, $\theta$, varied as a function of block, $F(11, 8975) = 5.71$, $p < .0001$, BIC = −4711, $R^2 = .40$. The maximum likelihood value of $\theta$ was .10 in Block 1, reached .32 by Block 3, peaked at .34 in Block 6, and steadily decreased toward .22 in Block 12. Thus, the pigeons tended

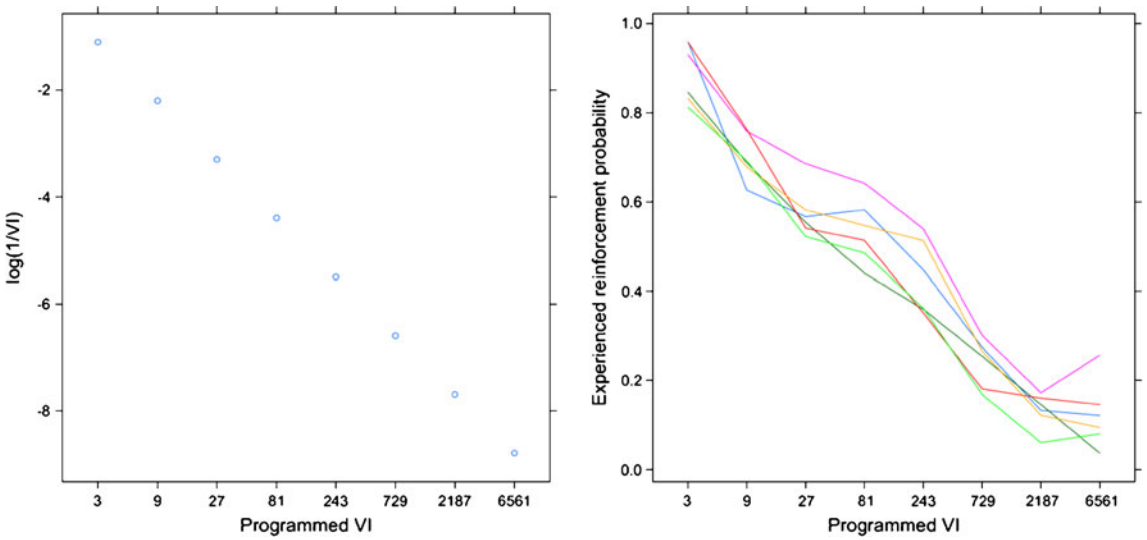



**Fig. 2** Log (1/VI) and experienced reinforcement probability as a function of the programmed VI for the 6 pigeons

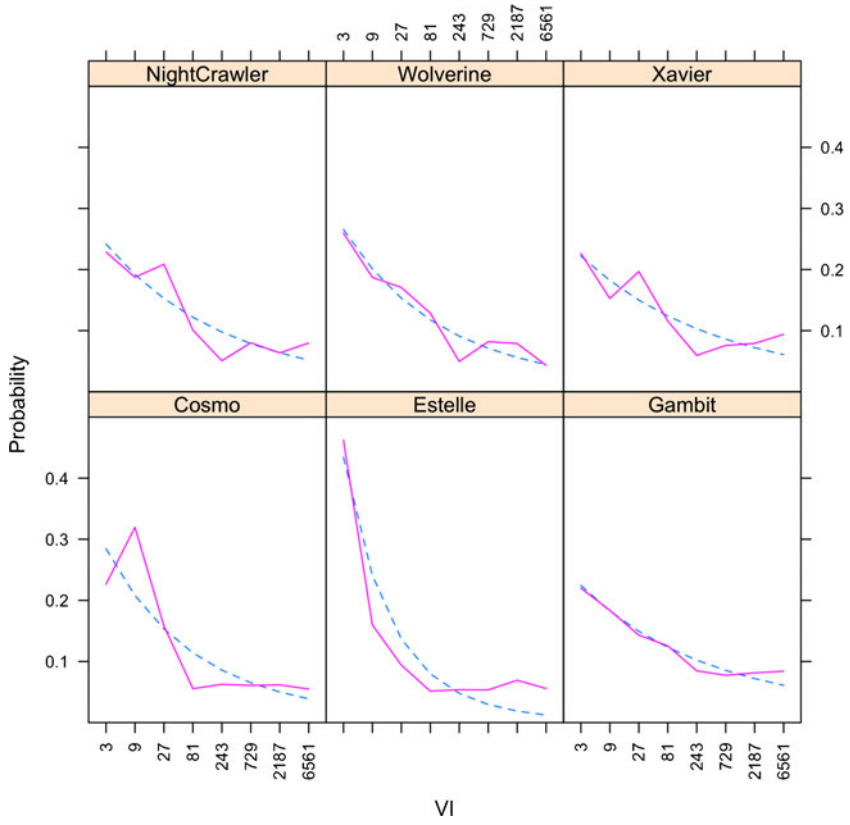



**Fig. 3** Probability of choosing a disk associated with each programmed VI for Experiment 1 (pigeons), with the best-fitting Luce function superimposed (dashed lines)

to quickly differentiate the better disks among the choice alternatives, but as the session progressed, their behavior became increasingly undifferentiated. Interestingly, this behavior was highly correlated with the number of pecks produced throughout the session: Pecking was highest during Blocks 2–4 and then gradually fell throughout the session. By Block 12, responding averaged 28% of the peak rate of responding. It appears that as the pigeons' level of satiety increased, the motivation to differentiate among the payoff disks decreased, or the motivation to exploit abated.

We have defined *exploitation* for this experiment as a response to the option with the richest VI schedule; thus, the decreases in $\theta$ later in the session indicate increased exploration/decreased exploitation. A molecular definition of exploitation would involve the choice of the response with the highest momentary probability of payoff. When payoffs are delivered by VI schedule, the longer it has been since a particular option has been chosen, the greater that probability is. The response option with the leanest overall VI schedule may be the richest at the moment, if enough time has passed since it was last chosen. If the increasing exploration of options later in a session were the result of pigeons learning to choose other options due to an increase in their momentary reinforcement rate, we would expect an increase in payoff rate to accompany it. This outcome did not occur. Figure 5 shows the proportion of responses reinforced for each trial block within a session and indicates that—with the exception of Estelle—decreases in differentiation were associated with decreases, not increases, in reinforcement.

The consequences of our using a VI schedule are revealed in the likelihood of continuing to respond on a disk that has just been rewarded. Figure 6 (left column) shows a smoothed spline of the likelihood of returning to a disk as a function of time elapsed since it was last rewarded, for the disks assigned the three richest schedules. The figure reveals a temporary decrease immediately following reward for some pigeons, at least for the VI 3-s and 9-s disks. During this dip, the pigeons were more likely to choose another disk (an exploratory response) as a function of its relative payoff likelihood, as shown in Figure 3.

The predicted disk choices for each pigeon are shown superimposed on Figure 3. Luce's decision rule predicts that responding is a monotonic function of disk value, and thus the rule cannot account for the unusual data patterns observed in Cosmo when disk value was solely a function of programmed (not experienced) payoff. However, the other birds' behavior was well approximated by Eq. 2.

Finally, we examined the degree to which disk value on a previous session lingered into the next session. In the first 5-min part of a session (Block 1), response likelihood was as much a function of a disk's value on the previous session [$t(6) = 4.33$, $p < .01$] as of its value for the current session [$t(6) = 4.07$, $p < .01$]. Over the next five blocks, the effect of a disk's previous value steadily decreased (*t*s of 2.52,

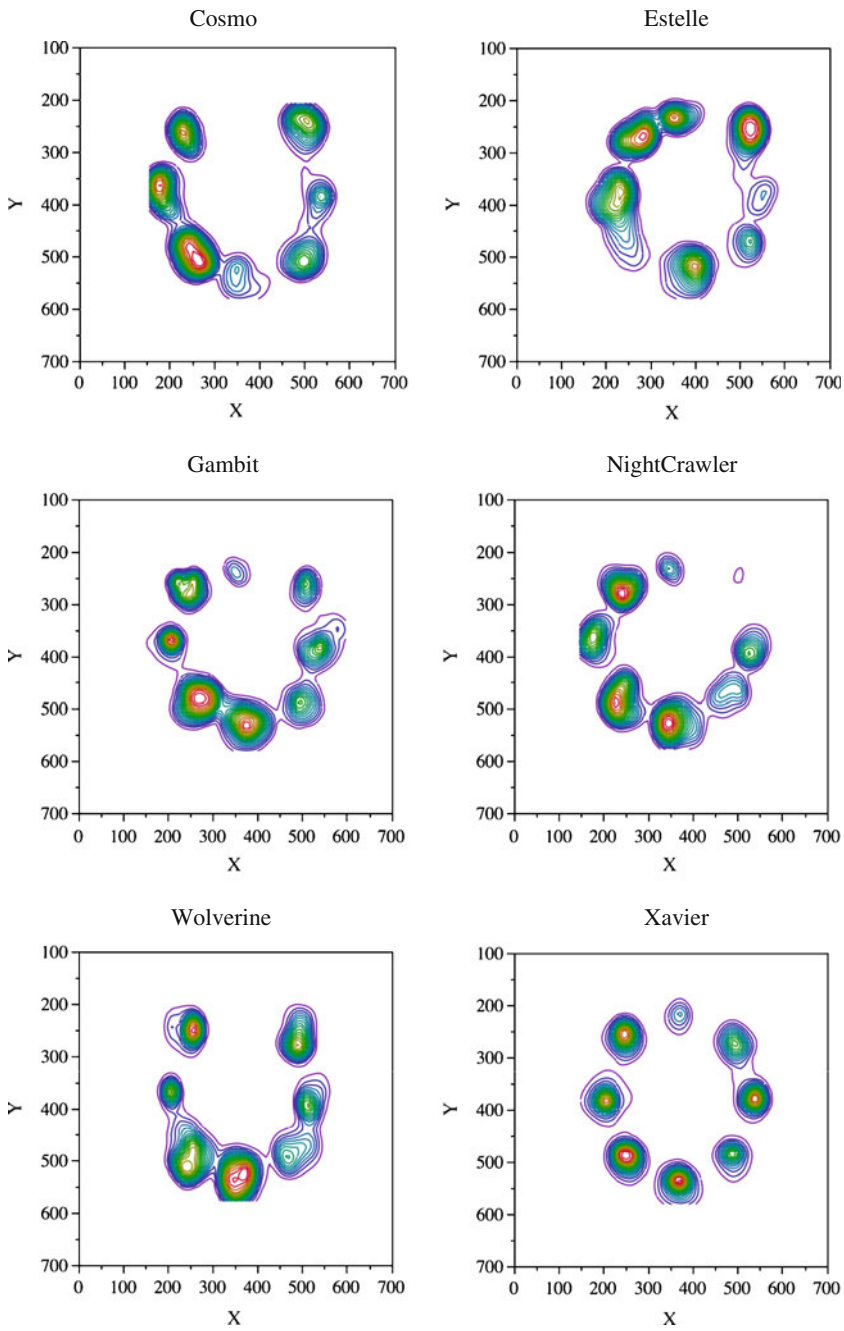



**Fig. 4** Peck density plots for each pigeon in Experiment 1. Only pecks on the disks are shown

1.66, 1.28, and 0.51), whereas the effect of a disk's current value was maintained (*t*s = 4.54, 3.74, 3.88, and 4.55).

## Choice differentiation as a function of programmed payoffs and disk location

Because some pigeons were not showing sufficient exploration of all eight response disks, using the programmed payoff in fitting Luce's decision rule is problematic. To incorporate the effect of disk location for individual birds, we assessed $\theta$ as a function of both trial block and disk location. The analysis revealed that the degree of response differentiation, $\theta$, varied as a function of both block, $F(11, 8968) = 3.74$, $p < .0001$, and disk location, $F(7, 8968) = 7.94$, $p < .0001$, BIC = −5,048, $R^2 = .46$. A model that included an interaction produced a poorer fit, BIC = −4,510, indicating that it was overparameterized, and thus the interaction was not included in our analysis.

The best-fitting $\theta$ values as a function of trial block and disk are shown in Figure 7 which shows the main effects of both block (line graph) and disk location (star plot). It is readily apparent that exploitation (i.e., behavioral differen-

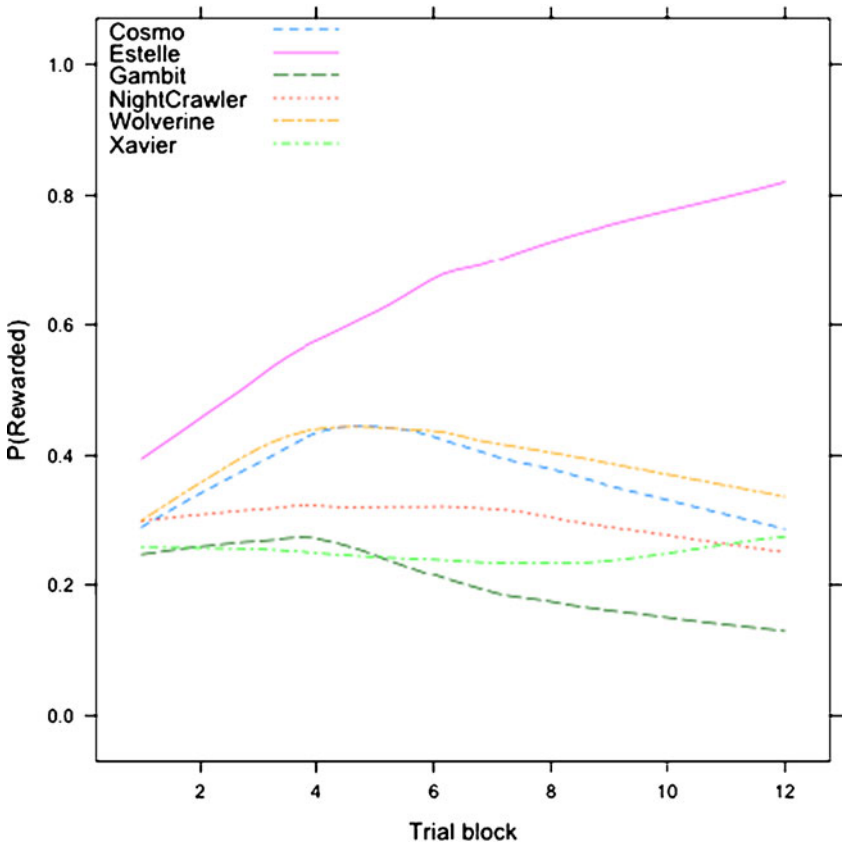



**Fig. 5** Proportions of responses that were rewarded as a function of trial block for each pigeon in Experiment 1

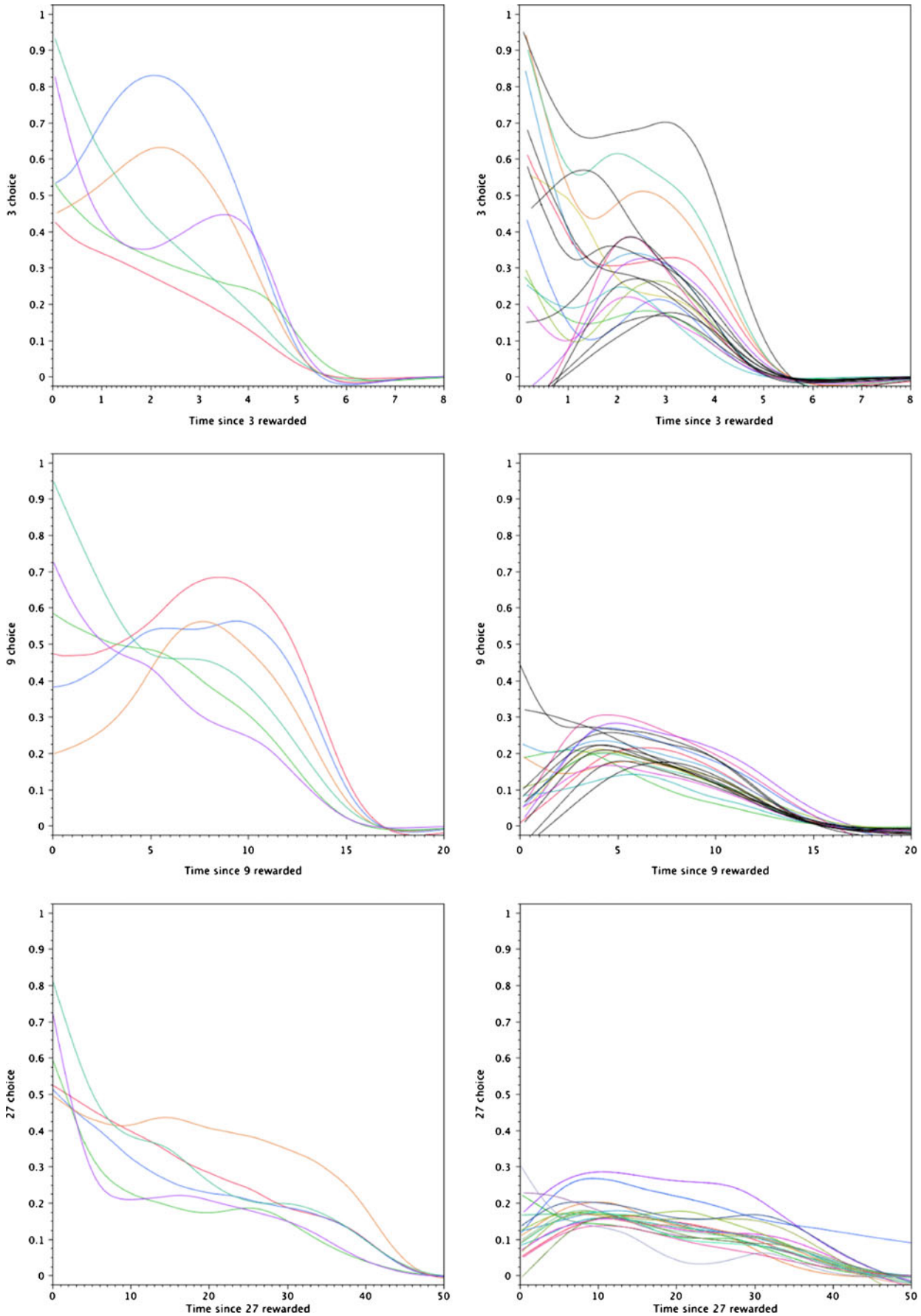
tiation as a function of disk payoff) peaks relatively early in a session and steadily decreases, paralleling our earlier analysis that did not include disk location as a predictor. It is also apparent that responses on disks in the upper right part of the display (disks 0, 1, and 2) produce weaker behavioral differentiation as a function of disk payoff (i.e., lower $\theta$s), confirming the behavioral patterns documented in the peck density plots of Figure 4. Although the fit was better for this analysis, the improvements were relatively minor.

## Discussion

In an eight-armed bandit task, pigeons' disk choice was largely a function of the VI schedule associated with each disk. For 4 of the pigeons, their behavior was broadly consistent with that predicted by Luce's decision rule as applied to the programmed reinforcement rate [log (1/VI)], thus suggesting that the derived $\theta$ values are good estimates of the degree of exploitation exhibited by the pigeons. Pigeons did not demonstrate high degrees of exploration at the beginning of a session that was cued by session onset, but rather their low $\theta$ values were a result of behavior being heavily influenced by carryover from the prior session's disk values. Within 10 min, however, their responding was largely driven by the new reinforcement contingencies. Thus, increases in exploration were likely produced by adversity—only when preferred disks were no longer paying off at a high rate did the pigeons begin to explore other choices (see Gallistel, Mark, King, & Latham 2001, for an alternative interpretation of matching in nonstationary environments).

Our pigeons, which were working for primary reinforcers, showed less exploitation as a session progressed. This change could have been due to an anticipated change in disk payoffs, but the evidence suggests that exploitation decreased due to an increase in satiety. Regardless of this pattern, we did not see high degrees of exploitation at any point in a session. Averaged across every session and trial block, no pigeon chose its preferred disk more than 45% of the time (see Fig. 3). When these results were averaged across sessions but broken down by trial blocks, no pigeon chose its preferred disk more than 55% of the time (not shown). The pigeons were not adopting greedy strategies in

**Fig. 6** Smoothed response likelihood plots for each pigeon or human participant, showing the relative likelihoods of choosing the VI 3-s disk (top row), the VI 9-s disk (second row), and the VI 27-s disk (bottom row) as a function of time since the disk's previous reinforcement. Pigeons are on the left, and humans are on the right. The *x*-axis scales differ for each figure; the scales range from 0 to 8 in the top graphs, 0 to 20 in the middle graphs, and 0 to 50 in the bottom graphs


3 choice
Time since 3 rewarded
3 choice
Time since 3 rewarded
9 choice
Time since 9 rewarded
9 choice
Time since 9 rewarded
27 choice
Time since 27 rewarded
27 choice
Time since 27 rewarded

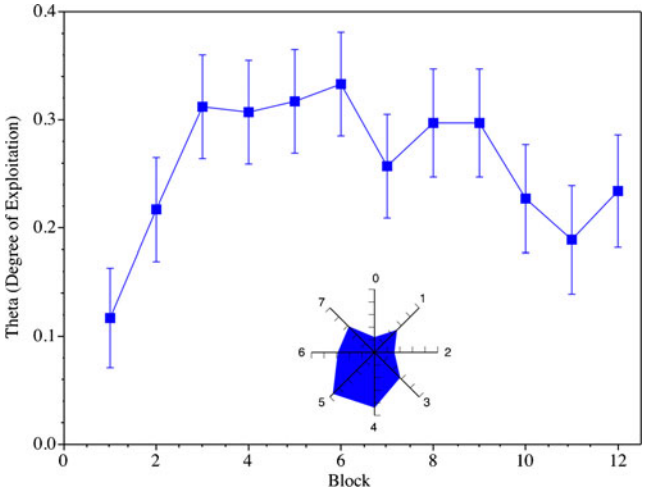



**Fig. 7** Best-fitting values of $\theta$ as a function of 5-min trial block (line graph) and disk location (star plot; the axis range is 0 to .5). Error bars represent ±1 standard error

our nonstationary environment. Despite our attempts to eliminate disk biases, the birds continued to show location preferences that were independent of a disk's programmed reinforcement schedule. We attempted to incorporate these biases into our analysis as an independent factor that allowed less behavioral differentiation (lower $\theta$ values) for certain disk locations, but the fit was only marginally better. An alternative formulation that would retain Luce's decision rule would be to incorporate disk location into our estimates of value, thus making a disk's value a function of both its scheduled payoff and its location. Unfortunately, this approach would require a post hoc assessment of disk preferences for each bird.

## Experiment 2

In our second experiment, we used a similar design to examine exploration versus exploitation in humans. We anticipated rapid changes in $\theta$ and fewer location preferences that were independent of payoffs. The literature on risky choice and risk perception suggests that people might be well adapted to identifying and responding to changes in payoffs for decisions under uncertainty (for a discussion of various examples, see Rakow & Miler, 2009).

### Method

#### *Participants*

A total of 20 undergraduates (16 female, 4 male) at the University of California, Los Angeles (UCLA), received course credit for participating in the experiment.

#### *Apparatus*

Testing was conducted on a notebook computer with a 38-cm (diagonal) color monitor set at 1,152 × 864 pixels. Participants used a mouse to guide a cursor around a screen, and a response was recorded every time the left mouse button was clicked. A built-in speaker was used to give auditory feedback when a reward was given.

#### *Procedure*

Before the experiment commenced, participants recorded their gender, age, ethnicity, and grade point average at UCLA. We told participants that they would be doing a test of intelligence, that they would be presented with eight differently colored disks on the screen, and that they would be required to click on the disks using the cursor (see Fig. 8). The instructions indicated that sometimes when they did this, a box at the bottom of the screen would light up with the text "Click for a point," at which point they should click the box to receive a point; their objective was to earn as many points as possible. Participants then completed a sample trial where the disks were present in an identical arrangement to that used for the pigeons (however, the particular disk color assignments differed from those used in Experiment 1). Clicking on any of the disks resulted in the box at the bottom lighting up. When participants clicked on the box, they heard the sound of a penny dropping, the box went dark, and they were awarded a point.

Following the sample trial, participants completed six sessions. Each session was 6 min long. We used the same reward schedule that had been used with the pigeons: VIs of 3, 9, 27, 81, 243, 729, 2,187, and 6,561 s, with ±50% variation. The assignment of variable intervals to disks was constant within a session but was rearranged from session to session.

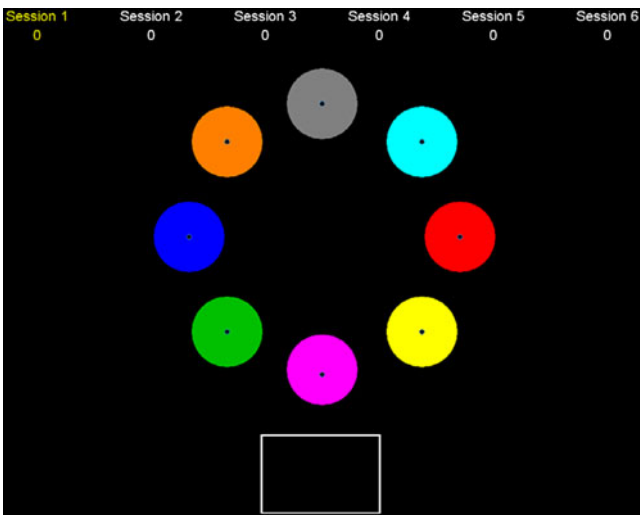



**Fig. 8** The computer screen as it was presented to the human participants in Experiment 2. For the purpose of analysis, the disks were numbered consecutively in a clockwise direction, with the top disk being disk 0

The same rearrangement from session to session was used for each participant. Counters were provided at the top of the screen giving an indication of how many points had been collected in each session, and the appropriate counter was updated every time a point was collected.

At the conclusion of each session, the participants needed to click on a button (not shown in Fig. 8) to start the next session. At the end of the fifth session, they were asked to type into the computer answers to the questions "What do you think was happening during the task?" "What strategy did you use to earn points?" "Within (not between) a given session, how did the colored discs differ from each other?" and "Was there a difference from one session to another? If so, what was the difference?" Following this, they were asked to do the final, sixth session.

## Results

When we examined the proportion of trials on which each disk was chosen by each person, all of the participants showed a systematic relationship between the scheduled payoff rate and the likelihood of choosing the corresponding disk (see Fig. 9, solid lines). A closer examination of the participants' disk choices revealed no strong disk biases, unlike the strong biases observed for the pigeons (not shown due to the large number of participants). Every participant showed sufficient sampling of each disk; the participant showing the strongest bias still chose the least preferred disk 62 times (3% of total choices) across the six sessions.

We initially analyzed the degree of response differentiation, $\theta$, as a function of session (1–6) and blocked time within session (30-s Blocks 1–12) using Eq. 1 (Fig. 9 shows the data fits as dashed lines; Fig. 10 shows changes in $\theta$ as a function of session and block). The analysis revealed that the degree of response differentiation, $\theta$, varied as a function of block, $F(11, 11389) = 56.48$, $p < .0001$, and session, $F(11, 11389) = 4.48$, $p < .0001$, with a Block × Session interaction, $F(55, 11389) = 2.62$, $p < .0001$, BIC = −27,010, $R^2 = .89$. The maximum likelihood value of $\theta$ increased steadily throughout a session, and did so more rapidly in the later sessions (unfilled symbols) than in the earlier sessions (filled symbols), with only the first session showing a prolonged and gradual increase in $\theta$.

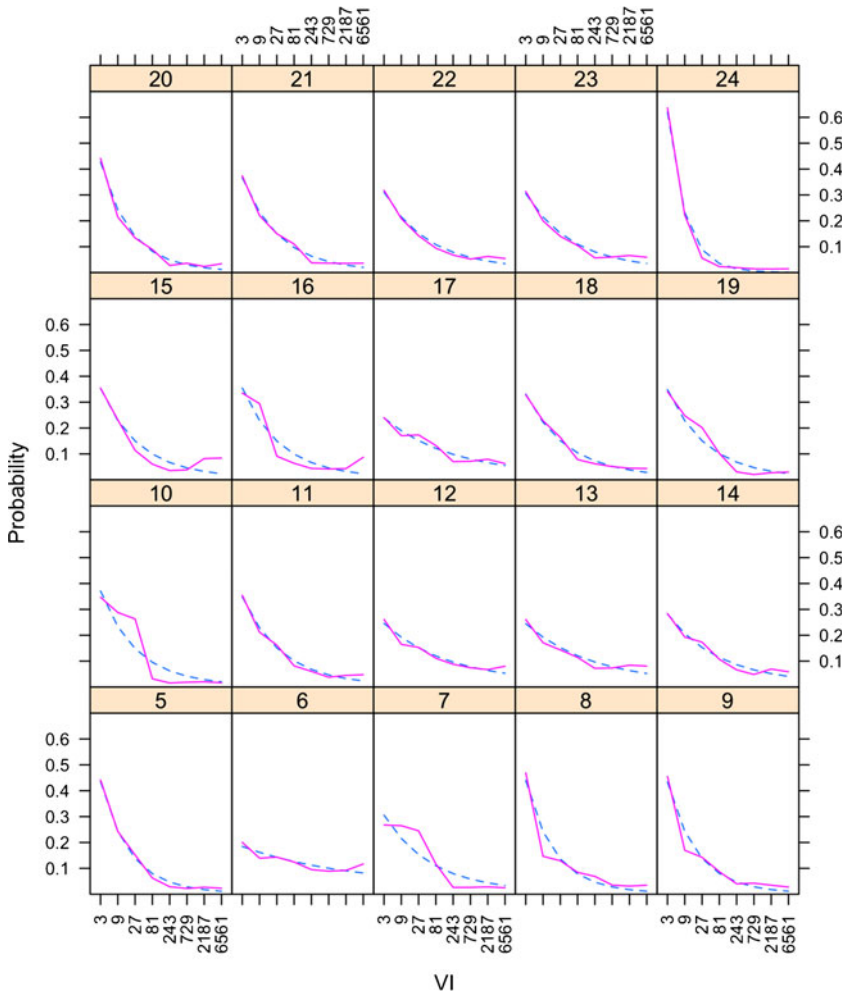



**Fig. 9** Actual probability of choosing each disk with the associated VI payoff rate for each participant in Experiment 2, with the best-fitting Luce fit superimposed (dashed lines)

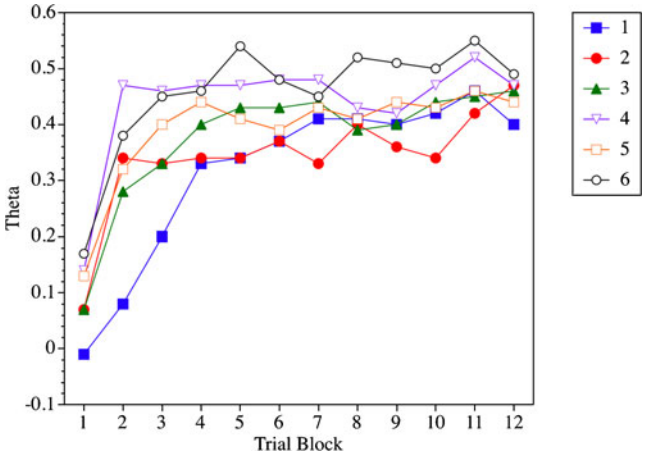



**Fig. 10** Best-fitting values of $\theta$ as a function of 30-s trial block (1–12) and session (1–6, as indexed in the legend). The standard errors were approximately .035

Unlike the pigeons, there was no indication of a loss of control as each session progressed.

As a consequence of the use of a VI schedule, most people showed a temporary decrease in the likelihood of choosing a disk after it was rewarded. Figure 6 (right column) shows the individual smoothed likelihood splines for each participant for the three richest schedules, and the vast majority of participants developed an aversion to returning to a disk that was just rewarded; the likelihood of returning to it was a function of its VI schedule. Thus, due to the temporary decrease in the efficacy of a recently rewarded response, participants were being encouraged to explore by sampling other disks.

Finally, we examined the degree to which disk value on a previous session lingered into the next session. In the first 30 s of a session (Block 1), response likelihood was largely a function of a disk's value for the current session [$t(19) = 7.75$, $p < .01$], but there was a small, nonsignificant effect of the disk's value from the previous session [$t(19) = 1.68$, $p = .11$]. Over the next four blocks, the effect of a disk's previous value remained small ($t$s of 2.63, 0.78, 1.23, and 1.13) and was only significant in Block 2, whereas the effect of a disk's current value increased and leveled off ($t$s = 10.47, 10.87, 11.95, and 11.89). By the final block, performance was entirely a function of a disk's value for the current session [$t(19) = 14.01$, $p < .01$], with little effect of the disk's value for the previous session [$t(19) = 0.99$, $p = .32$].

The strategy reports were largely uninformative. Six of the participants reported that points earned was somehow a function of time or delay (the correct controlling variable), 2 reported that points were a function of the number of times chosen, 1 reported a complex geometrical relationship, and the remaining participants' reports were either vague or equivalent to reporting that they did not know. Sex, self-reported GPA, and self-reported strategy did not significantly predict the best-fitting value of $\theta$, but our sample size was too small to identify all but the largest individual-difference effects (a prior study had found a weak negative correlation, $r = -.09$, between intelligence and exploratory behavior; Steyvers et al., 2009).

## Discussion

In our eight-armed bandit task, human disk choice was largely a function of the VI schedule associated with each disk. Behavior was generally consistent with that predicted by Luce's decision rule as applied to the programmed reinforcement rate [log (1/VI)]. Exploration was high early in a session and was only weakly a function of a disk's previous value. This lack of carryover, accompanied by a high degree of exploration in the first block of a session (see Fig. 10), likely occurred because the transition from session to session was clearly demarcated for the participants (Fig. 8 shows the highlighting of the current session at the top of the screen). Thus, our human participants showed an adaptive increase in exploration in the presence of a signal that indicated a change in disk payoffs, unlike the pigeons in Experiment 1. Finally, like the pigeons, our human participants did not demonstrate a greedy strategy (see Fig. 9). Instead, they continued to explore other alternatives late in a session.

## General discussion

Both pigeons and people produced response patterns that were often well modeled by Luce's (1963) decision rule. Although there were some exceptions (most notably the pigeon Cosmo in Exp. 1), these deviations may have been driven by differences in the programmed and experienced disk payoffs or by idiosyncratic strategies that we have not assessed. Additionally, neither species demonstrated greedy strategies in the nonstationary environments used in the present study. Whereas exclusive choice of the highest value disk would seem adaptive once a chooser has learned that disks only change their value across sessions, the use of a VI schedule likely contributed to higher exploration by producing a temporary decrease in the value of a disk (see Fig. 6). Given the clocked nature of a VI, a disk with a leaner schedule is more likely to be rewarded than a disk with a richer schedule if the lean disk has not been chosen in a long time. For example, consider the choice between a VI 3-s and a VI 9-s disk. If the VI 3-s disk was chosen 8 s into a session, it would have an average delay of 3 s until its next reward was available (i.e., 11 s into the session). By contrast, the VI 9-s disk would have an average delay of 1 s until its next reward was available (i.e., 9 s into the session). Thus, the adoption of an optimal fully informed

strategy would cause a chooser to occasionally sample the leaner schedules as a function of the elapsed time since their last reinforcement. Both the pigeons' and people's behavior often demonstrated a temporary decrease in the likelihood of choosing a disk that was recently rewarded, along with a rapid increase soon after (Fig. 6). After a peak in likelihood, responding gradually fell, which is largely a result of responses to a disk eventually being rewarded, thus truncating the distribution.

The greatest species differences involved (a) strong disk biases in the pigeons but not in people and (b) the weak carryover of disk value across sessions for people but the strong carryover for pigeons. The strong disk biases were quite intransigent in our pigeons. Even after extensive attempts to train out these biases, the pigeons still underexplored certain responses (see Fig. 4). We believe that there are two significant contributors to these biases. First, the upper disks may have required substantial effort to reach, thus reducing their value due to a high response cost (cf. Jensen et al., 2006). Second, the pigeons may have been content to satisfice, such that there was insufficient motivation to maximize their reward rate. Given that the response rate gradually abated later in the session, satiation may have reduced the incentive to identify the disk with the highest value.

The second large species difference involved the fact that the pigeons' behavior early in a session was heavily influenced by the disk values from the previous session, whereas people showed little session-to-session carryover of value. This result is even more remarkable given the extensive experience that the pigeons had with daily changes during training (309 sessions) and testing (24 sessions), an ample opportunity to learn that disk value did not (except in rare instances) carry over across sessions. In contrast, our human participants received only 6 min of training before disk payoffs changed and yet showed little value carryover. Thus, the pigeons increased exploration largely in response to an experienced change in payoff rates, whereas people increased exploration when a discriminative cue dictated.

The control over performance exerted by disk values from the prior session is striking when one considers that nonstationary procedures reveal strong constraints on the duration of working memory in the pigeon. Pigeon working memory has been found to last from tens of seconds, in delayed matching-to-sample procedures (e.g., Grant, 1976; White, Ruske, & Colombo 1996), to no more than 1 or 2 h, on open-field spatial search tasks (Spetch, 1990; Spetch & Honig, 1988). This stands in stark contrast to retention of correct responses in stationary procedures, which have been shown to last for months or years (e.g., Cook, Levison, Gillett, & Blaisdell 2005; Vaughan & Greene, 1984). Above-chance retention of disk values over a 24-h interval after only a single session of exposure has previously been reported in two-choice situations (e.g., Grace & McLean, 2006; Kyonka & Grace, 2008; Schofield & Davison, 1997). These studies involving between-session changes in reinforcement schedules reveal some lasting influence of the prior session's reinforcement contingencies at the beginning of the next session. To our knowledge, however, ours are the first results showing similar carryover effects on schedules involving more than two choice options. This suggests that pigeons acquired some memory for the distribution of values across multiple choice options from a single session, the influence of which persisted in the following session. We can only speculate that our task contained features that better tap into processes of long-term memory than have previous working memory procedures.

Although our human participants showed adaptive increases in exploratory behavior at the beginning of a session, session onset was clearly signaled. It is not known how quickly people would increase their exploratory behavior if change was not signaled. Without a signaled change in schedule, any increase in exploration would likely be a function of the magnitude of the change in disk value and of which disks (e.g., those of previously high or low value) changed their value. If a low-value, and thus undersampled, disk suddenly became the richest option, a high exploiter would be slow to discover this change. In contrast, if a high-value, and thus heavily sampled, disk suddenly decreased in value (which was typically the case in the present experiments), this change would be apparent to both high and low exploiters.

People's sudden increase in exploratory behavior at the onset of each session suggests a level of operant control that goes beyond merely responding to changes in the payoffs of the operanda. One possibility is that this result provides further evidence of behavioral variability as an operant (Neuringer, 2002; Page & Neuringer, 1985), but the rapidity with which our human participants responded suggests insufficient time for variability to have been reinforced during the confines of our experiment. Thus, people previously must have learned the utility of exploration in the face of a rapidly changing environment. Pigeons, on the other hand, may be better adapted to more stable environments that reward perseveration over flexibility.

Although an actor always faces uncertainty about the utility of future actions, the randomness of events underlying this uncertainty extends from that conforming to well-understood linear-based Gaussian distributions to those best described by poorly understood nonlinear power laws (Taleb, 2007). It would be very interesting to understand how actors as diverse as humans and pigeons face action-making decisions in these vastly different types of stochastic contexts that characterize real-world situations. Given the importance of understanding choice and the common desire to optimize choice strategies in stationary

and nonstationary environments, we hope that more researchers will consider spending less time exploiting the study of simple choice tasks with stationary payoffs, and instead allocate more effort toward exploring many-choice tasks in nonstationary environments (e.g., Davison & Baum, 2000; Ward & Odum, 2008).

## References


Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science* (pp. 322–331). Piscataway, NJ: IEEE Press.

Azoulay-Schwartz, R., Kraus, S., & Wilkenfeld, J. (2004). Exploitation versus exploration: Choosing a supplier in an environment of incomplete information. *Decision Support Systems, 38*, 1–18.

Banks, J., Olson, M., & Porter, D. (1997). An experimental analysis of the bandit problem. *Economic Theory, 10*, 55–77.

Burns, N. R., Lee, M. D., & Vickers, D. (2006). Individual differences in problem solving and intelligence. *Journal of Problem Solving, 1*, 20–32.

Cook, R. G., Levison, D. G., Gillett, S. R., & Blaisdell, A. P. (2005). Capacity and limits of associative memory in pigeons. *Psychonomic Bulletin & Review, 12*, 350–358.

Cudeck, R., & Harring, J. R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology, 58*, 615–637.

Davidian, M., & Giltinan, D. M. (2003). Nonlinear models for repeated measurements: An overview and update. *Journal of Agricultural, Biological, and Environmental Statstics, 8*, 387–419.

Davison, M., & Baum, W. M. (2000). Choice in a variable environment: Every reinforcer counts. *Journal of the Experimental Analysis of Behavior, 74*, 1–24.

Dimitrakakis, C., & Lagoudakis, M. G. (2008). Rollout sampling approximate policy iteration. *Machine Learning, 72*, 157–171.

Gallistel, C. R., Mark, T. A., King, A. P., & Latham, P. E. (2001). A rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. *Journal of Experimental Psychology: Animal Behavior Processes, 27*, 354–372.

Grace, R. C., & McLean, A. P. (2006). Rapid acquisition in concurrent chains: Evidence for a decision model. *Journal of the Experimental Analysis of Behavior, 85*, 181–202.

Grant, D. S. (1976). Effect of sample presentation time on long-delay matching in pigeons. *Learning and Motivation, 7*, 580–590.

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior, 24*, 107–116.

Jensen, G., Miller, C., & Neuringer, A. (2006). Truly random operant responding: Results and reasons. In E. A. Wasserman & T. R. Zentall (Eds.), *Comparative cognition: Experimental explorations of animal intelligence* (pp. 459–480). New York: Oxford University Press.

Koulouriotis, D. E., & Xanthopoulos, A. (2008). Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation, 196*, 913–922.

Kyonka, E. G. E., & Grace, R. C. (2008). Rapid acquisition of preference in concurrent chains when alternatives differ on multiple dimensions of reinforcement. *Journal of the Experimental Analysis of Behavior, 89*, 49–69.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963–974.

Lin, Y. K., & Batzli, G. O. (2002). The cost of habitat selection in prairie voles: An empirical assessment using isodar analysis. *Evolutionary Ecology, 16*, 387–397.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). New York: Wiley.

Mettke-Hofmann, C., Wink, M., Winkler, H., & Leisler, B. (2004). Exploration of environmental changes relates to lifestyle. *Behavioral Ecology, 10*, 2004.

Neuringer, A. (2002). Operant variability: Evidence, functions, and theory. *Psychonomic Bulletin & Review, 9*, 672–705.

Page, S., & Neuringer, A. (1985). Variability is an operant. *Journal of Experimental Psychology: Animal Behavior Processes, 11*, 429–452. doi:10.1037/0097-7403.11.3.429.

Pinheiro, J. C., & Bates, D. M. (2004). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Plowright, C. M., & Shettleworth, S. J. (1990). The role of shifting in choice behavior of pigeons on a two-armed bandit. *Behavioural Processes, 21*, 157–178. doi:10.1016/0376-6357(90)90022-8.

Rakow, T., & Miler, K. (2009). Doomed to repeat the successes of the past: History is best forgotten for repeated choices with nonstationary payoffs. *Memory & Cognition, 37*, 985–1000.

Rothstein, J. B., Jensen, G., & Neuringer, A. (2008). Human choice among five alternatives when reinforcers decay. *Behavioural Processes, 78*, 231–239. doi:10.1016/j.beproc.2008.02.016.

Schofield, G., & Davison, M. (1997). Nonstable concurrent choice in pigeons. *Journal of the Experimental Analysis of Behavior, 68*, 219–232.

Shkedy, Z., Straetemans, R., & Molenberghs, G. (2005). Modeling anti-KLH ELISA data using two-stage and mixed effects models in support of immunotoxicological studies. *Journal of Biopharmaceutical Statistics, 15*, 205–223.

Sikora, R. T. (2008). Meta-learning optimal parameter values in non-stationary environments. *Knowledge Based Systems, 2*(8), 800–806.

Spetch, M. L. (1990). Further studies of pigeons' spatial working memory in the open-field task. *Animal Learning & Behavior, 18*, 332–340.

Spetch, M. L., & Honig, W. K. (1988). Characteristics of pigeons' spatial working memory in an open-field task. *Animal Learning & Behavior, 16*, 123–131.

Stahlman, W. D., Roberts, S., & Blaisdell, A. P. (2010). Effect of reward probability on spatial and temporal variation. *Journal of Experimental Psychology: Animal Behavior Processes, 36*, 77–91.

Stahlman, W. D., Young, M. E., & Blaisdell, A. P. (2010). Response variability in pigeons in a Pavlovian task. *Learning & Behavior, 38*, 111–118.

Steyvers, M., Lee, M. D., & Wagenmakers, E. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology, 53*, 168–179.

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

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Valsecchi, I. (2003). Job assignment and bandit problems. *International Journal of Manpower, 24*(7), 844–866.

Vaughan, W., & Greene, S. L. (1984). Pigeon visual memory capacity. *Journal of Experimental Psychology: Animal Behavior Processes, 10*, 256–271. doi:10.1037/0097-7403.10.2.256.

Ward, R. D., & Odum, A. L. (2008). Sensitivity of conditional-discrimination performance to within-session variation of reinforcer frequency. *Journal of the Experimental Analysis of Behavior, 90*, 301–311.

White, K. G., Ruske, A. C., & Colombo, M. (1996). Memory procedures, performance and processes in pigeons. *Cognitive Brain Research, 3*, 309–317. doi:10.1016/0926-6410(96)00016-X.

Zach, R. (1979). Shell dropping: Decision-making and optimal foraging in northwestern crows. *Behaviour, 68*, 106–117.