INTERNATIONAL JOURNAL OF COMPARATIVE PSYCHOLOGY

2014, 27 (2), 248-264
A. Blaisdell & W. D. Stahlman
Special Issue
Peer-reviewed

Effects of Response Frequency Constraints on Learning in a Non-Stationary Multi-armed Bandit Task

Michael E. Young
Kansas State University, U.S.A.

Deborah E. Racey
Western Carolina University, U.S.A.

An *n*-armed bandit task was used to investigate the trade-off between exploratory (choosing lesser-known options) and exploitive (choosing options with the greatest known probability of reinforcement) human choice in a trial-and-error learning problem. A different probability of reinforcement was assigned to each of eight response options using random-ratios (RRs), and participants chose by clicking buttons in a circular display on a computer screen using a computer mouse. To differentially increase exploration, relative frequency thresholds were randomly assigned to each participant and acted as task constraints limiting the proportion of total responses that could be attributed to any response option. The potential benefit of increased exploration in non-stationary environments was investigated by changing payoff probabilities so that the leanest options became the richest or the richest options became the leanest. On the average, forcing participants to explore at moderate to high levels always resulted in their earning less reinforcement, even when the payoffs changed. This outcome may be due to humans' natural level of exploration in our task being sufficiently high to create sensitivity to environmental dynamics.

Rarely is any choice made in the presence of perfect knowledge. Much of what informs our choices is learned as we interact with a stochastic environment in which outcomes are not consistent. This experience may accurately represent the true long-run nature of an option, or it may not. Our initial visits to a local restaurant or the first few lottery tickets purchased may produce an overly optimistic judgment if those experiences were unrealistically positive. Thus, the complete preference for one option after a limited amount of sampling may produce suboptimal preference. The problem is further exacerbated when the value of the options is not stationary. A continuing and exclusive preference for the best option today (e.g., a certificate of deposit) may lead to ignorance of improvements in the alternatives (e.g., stock funds). The present study considers whether a chooser should continue to explore under-sampled or sub-optimal options in the short run thus sacrificing near term outcomes in order to ensure sensitivity to possible changes in the quality of available options. If so, how can we encourage continued exploration and under what conditions?

Importantly, behavioral variability plays a role in learning and in choice under uncertainty (Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998). Completely repetitive behavior restricts exploration of alternatives, and purely random behavior usually disallows exploitation of what may have been learned. Increasing the level of variability (i.e., exploration) can benefit learning, perhaps contributing to discovery of better options through enhanced exploration (Neuringer, 2004; Stokes, 2001). The machine learning literature has provided a useful framework for understanding the acquisition of information about the relative utility of choices: reinforcement learning (Koulouriotis & Xanthopoulos, 2008; Sutton & Barto, 1998). Although the framework has broader applicability, this study focuses on the analysis of the multi-armed bandit task. The bandit task provides an excellent platform to explore choice in stationary (with unchanging payoffs) and non-stationary (with changing payoffs) environments and the possible effects of increased exploration on learning in these environments.

Please send correspondence to Dr. Michael Young (Email: michaelyoung@ksu.edu).

The Multi-Armed Bandit Task

In the multi-armed (or *n*-armed) bandit task, people choose among *n* options each of which has a probability of payoff. Each payoff is independent, and the probabilities are initially unknown. A learner must explicitly explore the environment to learn the expected payoffs and then can later choose to exploit this knowledge. In the reinforcement learning framework, choosing the option with the highest estimated value based on current knowledge (a *greedy* choice) is described as exploitation. Choosing a nongreedy option is described as exploratory because the choice potentially sacrifices short-term value by enhancing overall knowledge of the nongreedy options.

Whether exploration or exploitation is optimal at any given choice point will depend on current payoff estimates, the number of remaining plays, and the stability of the payoff structure. For example, although the simplest method of estimating value is to average all received rewards for each option, in a non-stationary bandit problem where the true values are changing, it might be more appropriate to use a recency-weighted average in which more recent rewards are weighted more heavily. Similarly, if there are only a few choices left before the task terminates, exploitation is likely optimal because there is little remaining time to take advantage of the new knowledge gained by additional exploration.

Modulating Exploration in Non-Stationary Environments

In non-stationary environments, obvious changes may adaptively increase exploration. If, however, the changes are more subtle but with a profound impact on future reinforcement density (e.g., a previously very low payoff response now has a very high payoff), then maintaining a particular level of exploration may be necessary to detect those changes. The ideal degree of exploration thus depends on the rapidity of environmental change and the nature of that change.

We do not yet know whether people adaptively shift between exploitation and exploration in a way that maximizes outcomes. For machine learning applications, a parameter such as theta in the softmax model can be manipulated in order to control the level of exploration and thus maximize reinforcement rate under various environment dynamics (Sutton & Barto, 1998). Because there is no theta parameter that can be directly manipulated in people, indirect control of the level of exploration in people is required. Previous research has identified four factors that influence behavioral variability and thus exploration: (1) the complexity of the task (Stokes & Balsam, 2003), (2) adversity (Balsam, Deich, Ohyama, & Stokes, 1998; Jensen, Stokes, Paterniti, & Balsam, 2013) or extinction (Antonitis, 1951), (3) reinforcement of variability (Page & Neuringer, 1985), and (4) task constraints (Stokes & Balsam, 2001; Stokes & Harrison, 2002).

These methods are not equally viable for a multi-armed bandit problem and often have behavioral consequences that confound the learner's ability to discern the payoff structure. For example, overt reinforcement of variability changes the experienced payoffs of the options such that it is unclear whether an option was reinforced due to its inherent value or due to the variability reinforcement schedule. Adversity or extinction occurs when reinforcement is withheld which tends to increase exploratory behavior, but by definition this manipulation changes the experienced reinforcement for the available options. Altering task complexity often produces a fundamental change in the task being performed. In contrast to these other approaches, introducing a task constraint is a method of directly and cleanly manipulating the level of exploration that allows the reinforcement probability for each response option to be affected only by the assigned payoff schedule and the sampling history of that option – we can simply prevent the chooser from overexploiting particular alternatives.

The present study was thus designed to address three questions: how would constraining the frequency of responses affect exploration, would particular constraint levels have a benefit on reinforcement rates after a payoff change that would compensate for an ongoing inability to exploit, and would the results depend on the relative discriminability of the response payoffs (i.e., large vs. small differences in payoffs) and the nature of the change of payoffs (readily identified because the change was to preferred responses vs. easily missed because the change was to less-preferred responses).

Participants performed an 8-armed bandit task to ensure that the task was complex enough to produce gradual learning of option values (cf. Jensen, Miller, & Neuringer, 2006; Racey, Young, Garlick, Pham, & Blaisdell, 2011). Each option was assigned a different RR with points used as reinforcers. A running total of points was continuously displayed to motivate participants. Because an optimal choice is in part determined by how many responses remain (the last response would not be well spent on exploration, for instance), a response counter was also displayed. A finite number of responses were available and the counter began at that number and was decremented by one with each response, whether or not reinforcement was earned by that response.

Four sets of eight RR schedules were tested in a pilot study, each set with a different *width* or magnitude of the difference between the option with the highest probability of payoff (richest RR) and lowest probability option (leanest RR). The differences between bandit-arm payoffs should not be so easily discriminable that the options with the highest probabilities are quickly identified and exclusively exploited, thus preventing the examination of learning. Likewise, they should not be so difficult to discriminate that exploration remains very high throughout the task, and preference for the best option(s) is developed very slowly or not at all. The goal was to select payoff ratios that, when assigned to the eight options, allowed participants to demonstrate moderately paced learning of the prevailing contingencies.

The Task Constraint

Our task constraint involved making a response option temporarily unavailable if it was used too frequently. We established a relative frequency threshold with a discounting factor applied so that recent responses were more heavily weighted. When the proportion of recency-weighted responses on any button exceeded the assigned threshold, that button became temporarily unavailable by removing it from the display. When the proportion later dropped below the threshold, the option became available again.

The appropriate constraint values (relative frequency thresholds) for this task were not obvious nor was there precedent to inform selection. Thus, we sampled from a range of relative frequency thresholds in order to observe how both exploration and the amount of reward earned might be affected across the range of constraint values. The method of *representative design* implements representative stimulus sampling from among all possible values within the selected range of relative probabilities, and so allows a broad and random sampling of possible values, rather than a systematic design methodology that would dictate a more arbitrary selection of a few values (Brunswik, 1955; Dhami, Hertwig, & Hoffrage, 2004; Young, Cole, & Sutherland, 2012). It may also be informative to know the level of exploration in the absence of constraints, and so in addition to sampling from the full range of constraint values, we included a group of participants at the no-constraint level (relative frequency threshold of 1.0).

One of the factors that informs the optimal level of exploration in a non-stationary reinforcement learning problem is the nature of the changes the environment will undergo. We investigated two types of changes expected to require differing exploration levels in order to optimize overall reward. The first change left the payoff ratios unchanged for the best options, while the previously leanest options improved such that they paid off at a higher rate than the previously best options. A complete failure to explore these initially

leaner options will have detrimental effects after the change. The second change type only changed the best options, reducing them to payoff probabilities equal to the previously poorest options. The change under these conditions should be readily noticed and should prompt an increase in exploration after the change.

Method

Participants

The participants were 128 students enrolled in an introductory psychology course. Students received course credit for their participation.

Procedure

Participants were tested in groups of a maximum of four people. When entering the lab they were seated in front of a computer. They then read and signed a consent form and the experimenter read general instructions.

The response options appeared as buttons arranged in a circular pattern on a computer monitor (see Figure 1) and a computer mouse was the response operandum. The buttons were identical and unmarked, visually discriminable only by their position in the circular array. The participant clicked on a button to choose it.

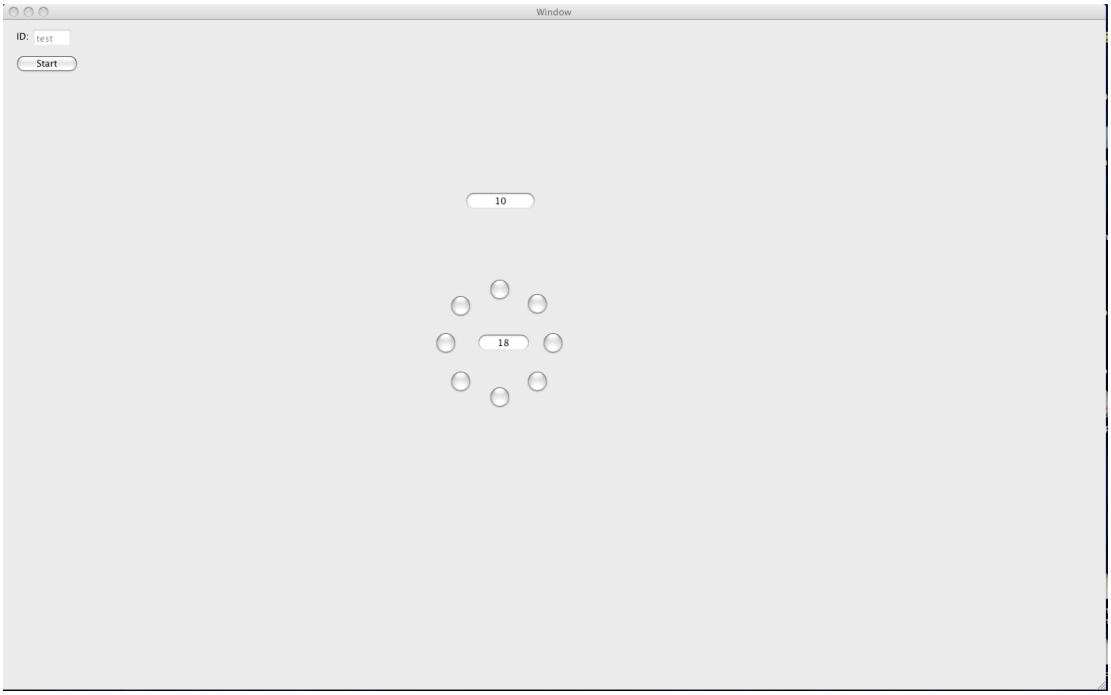


Figure 1. A screen shot of the experimental task.

A point counter appeared in the center of the array and a response tracker appeared at the top center of the screen. Participants were told that the study was a new kind of intelligence test and encouraged to do their best in order to help the experimenters evaluate the veracity of the new test. The experiment required 2000 responses dichotomized into Phase 1 (initial RRs) and Phase 2 (modified RRs) presented seamlessly to the participants without cueing. The response tracker began at 10 and counted down until the session terminated at zero; each of the 10 numbers on the count-down represented 200 responses completed. Responses that resulted in payoff caused one point to be added incrementally to the point counter. Participants were instructed that their task was to gain as many points as possible before their response tracker reached zero. A session required approximately 40 minutes to complete.

For Phase 1, two sets of eight RR schedules (wide vs. narrow) were selected based on the results of the pilot study. In the wide set, the Phase 1 RR schedules ranged from .05 to .75 (wide set); in the narrow set, Phase 1 RR schedules ranged from .05 to .40.

The RR values were pseudo-randomly assigned to the eight response options and were identical across participants in the same condition.

Conditions

Participants were assigned to one of four conditions (a 2 × 2 design). In the wide-increase condition, participants began Phase 1 with the wide RR set. In Phase 2, the leanest options from Phase 1 became the richest (see the first column of Table 1). Participants in the wide-decrease condition also began Phase 1 with the wide RR set. In Phase 2, however, the richest options from Phase 1 became equivalent to the two leanest options from Phase 1 (see the second column of Table 1). The narrow-increase and narrow-decrease conditions were similarly configured but with a much narrower range of initial RRs (see the third and fourth columns of Table 1).

Within each of the four conditions, participants were assigned a constraint value that remained constant throughout the experiment. Some participants (approximately eight in each condition) received no behavioral constraints. The remaining participants were assigned a randomly chosen constraint level by using a frequency threshold uniformly sampled from the .10 to 1.00 range with each participant receiving a different value.

Although restricting exploitation by use of the relative frequency constraint may harm the reinforcers earned in Phase 1, we anticipated that moderate constraint levels may have a small detrimental impact and be compensated by faster adaptation to the change in the RR schedules in Phase 2. This benefit, however, was predicted to be strongest for the wide-increase and narrow-increase conditions because these were the conditions where people would be least likely to notice the change in the payoff structure if their exploitation levels were too high.

Table 1
Phase 1 and Phase 2 random ratios by condition

Wide-Increase	Wide-Decrease	Narrow-Increase	Narrow-Decrease
Phase 1, Phase 2	**Phase 1, Phase 2**	**Phase 1, Phase 2**	**Phase 1, Phase 2**
.05, .85	.05, .05	.05, .60	.05, .05
.15, .95	.15, .15	.10, .70	.10, .10
.25, .25	.25, .25	.15, .15	.15, .15
.35, .35	.35, .35	.20, .20	.20, .20
.45, .45	.45, .45	.25, .25	.25, .25
.55, .55	.55, .55	.30, .30	.30, .30
.65, .65	.65, .05	.35, .35	.35, .05
.75, .75	.75, .15	.40, .40	.40, .10

Table 2
Examples of response distributions that produce certain entropy levels

Entropy	Example Response Distribution
0.1	.98, .02, .00, .00, .00, .00, .00, .00
0.8	.75, .25, .00, .00, .00, .00, .00, .00
1.5	.50, .30, .20, .00, .00, .00, .00, .00
2.2	.30, .30, .15, .13, .12, .00, .00, .00
2.6	.30, .20, .12, .12, .12, .08, .06, .00
2.8	.20, .19, .12, .12, .12, .12, .12, .01
2.9	.19, .17, .12, .12, .12, .12, .12, .04
3.0	All .125

Analysis

The primary dependent variables were the level of exploration and the number of points earned (reinforcements). To assess exploration, we used entropy, a measure of categorical variability (Shannon & Weaver, 1949):

$$H(A) = -\sum_{a \in A} p_a \log_2 p_a \qquad \text{(Eq. 1)}$$

where $H(A)$ is the entropy of categorical variable A, a is a category of A, and p_a is the proportion of observed values within that category. If a participant was showing an exclusive preference for one button during a 100-trial block of training, the entropy for that block would be 0. If a participant were evenly distributing their responses across the eight buttons, the entropy for that block would approach 3 (the highest entropy possible when there are eight options). Because entropy is an unfamiliar measure for most psychologists, Table 2 shows some example response distributions and their computed entropy.

To ease visualization, the continuous relative frequency threshold was divided into 4 categories .10-.39, .40-.69, .70-.99, and 1.0 (i.e., no constraint). All analyses, however, treated this independent variable as continuous.

Results

A total of 127 participants successfully completed the experiment. Figure 2 plots the likelihood of choosing each button as a function of its payoff rate (RR schedule) in the first phase of the experiment for each of the conditions in each of the experimental phases. First, not surprisingly, people showed greater discrimination among the buttons as a function of their payoffs in the wide conditions (top figure) than in the narrow conditions (bottom figure). Second, the average participant showed a shift from the initially preferred buttons in the first phase to those buttons with the highest payoffs in the second phase thus demonstrating sensitivity to the change. Third, as predicted, this shift in preference was more marked for the decrease conditions (in which the richest buttons lost value) than for the increase conditions (in which the poorest buttons gained value), but only for the wide conditions (top figure). For the narrow conditions, the pattern was more complex and varied as a function of the constraint. Fourth, participants subjected to the weakest constraints (highest frequency thresholds) showed the strongest discrimination among the RR schedules in both phases of the experiment.

Phase 1 Details

To confirm these observations, we ran series of repeated measures Poisson regressions (a generalized multilevel model) in which we predicted the number of responses on a button as a function of its RR. The primary independent variable was thus the scheduled RR. A positive slope for this variable indicates sensitivity to the scheduled RRs in the phase. We began by analyzing performance in the first phase for the wide and narrow conditions separately and considered two moderators of the sensitivity to the RRs, the response frequency threshold (as a continuous predictor) and 100-trial block of training (1 to 10, using a logarithmic transformation).

In the wide condition, the estimated RR slope was 2.9 (SE = 0.3, z = 10.87) revealing strong sensitivity to the programmed contingencies. As expected, this sensitivity increased logarithmically across blocks (RR × block slope = 1.1, SE = 0.03, z = 41.01). Furthermore, the sensitivity was highest for the highest response thresholds (RR × threshold slope = 0.9, SE = 0.4, z = 2.20) in agreement with Figure 2.

In the narrow condition, the estimated RR slope was 1.5 (SE = 0.4, z = 3.65) revealing good sensitivity to the programmed contingencies, but much weaker than that observed in wide condition due to the greater difficulty of the discrimination. As expected, this sensitivity increased logarithmically across blocks (RR × block slope = 1.0, SE = 0.05, z = 20.33). Finally, the sensitivity was again highest for the highest response thresholds (RR × threshold slope = 2.9, SE = 1.2, z = 2.50), but this sensitivity to the threshold was much higher than that observed in the wide condition ($p < 0.05$).

Phase 2 Details

We conducted separate analyses for the four conditions: wide-decreasing, wide-increasing, narrow-decreasing, and narrow-increasing. Although we originally ran a complete model involving all of the predictors, the presentation of the results was much too complex due to the presence of 5-way interactions, so we used a staged approach with appropriate corrections for Type I error. In each regression, we examined the following predictors: Phase 1 RR, Phase 2 RR, changes in Phase 1 and Phase 2 RR sensitivity across blocks (i.e., their two-way interaction), frequency threshold, and frequency threshold as a moderator of Phase 2 RR sensitivity (i.e., their two-way interaction).

Due to the greater complexity of these results, the best fitting regression weights for the RR sensitivities are shown in Figure 3 along with the estimates derived from Phase 1. The graphs show the loss of behavioral control by the Phase 1 RR contingencies during Phase 2, as expected. Consistent with our hypothesis, in the wide condition the control by the new Phase 2 RR contingencies was stronger in the decreasing condition than in the increasing condition ($z = 2.86$, $p < 0.05$). In opposition to our hypothesis, in the narrow condition the control by the new Phase 2 RR contingencies was weaker in the decreasing condition than in the increasing condition, but this difference did not reach statistical significance ($z = -1.70$, $p < 0.10$) due to the large variability in the estimate of the sensitivity to the Phase 2 RRs.

Effects on Exploration (Entropy)

Figure 4 shows the differences in exploratory behavior (entropy) across threshold categories for each condition and phase; entropy was calculated for each block for each subject. In the wide conditions (top figure), participants showed a range of natural variability levels when not subject to a response constraint. Stricter constraints (lower thresholds) increased the level of exploration as required by the task. It appears that participants explored less in the second phase, especially in the increasing condition. In the narrow conditions (bottom figure), participants showed a range of variability levels in the absence of a threshold that was much higher than that produced in the wide conditions. Indeed, in the narrow decreasing condition, the level of exploration was quite high regardless of the threshold; only the strictest constraint had a visible effect and then only by removing the lower tail of the distribution. In the narrow increasing condition, higher constraints (i.e., lower frequency thresholds) had a more visible effect on increasing exploration, and the strongest effect occurred in the second phase.

These observations were confirmed by a repeated measures regression of frequency threshold, phase, increasing/decreasing, wide/narrow, and their interactions as predictors of entropy within each block for each participant. We used a model comparison approach to identify the simplest model that eliminated a number of the highest order interactions. For the best model, there was a main effect of threshold ($F(1, 127) = 106.0$, $p < 0.01$), phase ($F(1, 127) = 56.2$, $p < 0.01$), wide/narrow ($F(1, 127) = 25.5$, $p < 0.01$), but not increasing/decreasing ($F(1, 127) = 2.7$, $p > 0.05$). These effects were qualified by the following significant interactions: Constraint × Phase ($F(1, 127) = 37.9$, $p < 0.01$), Phase × Increasing/Decreasing ($F(1, 127) = 21.6$, $p < 0.01$), Constraint × Wide/narrow ($F(1, 127) = 7.3$, $p < 0.01$), and Constraint × Phase × Increasing/decreasing ($F(1, 127) = 7.0$, $p < 0.01$); the Constraint × Increasing interaction did not reach statistical significance ($F(1, 127) = 3.1$, $p > 0.05$) but was included due to the presence of the 3-way interaction. In sum, entropy was significantly lower when the threshold was higher, in Phase 2, and for the wide conditions. The interactions revealed that combining these variables produced even lower entropy than would be predicted by their main effects.

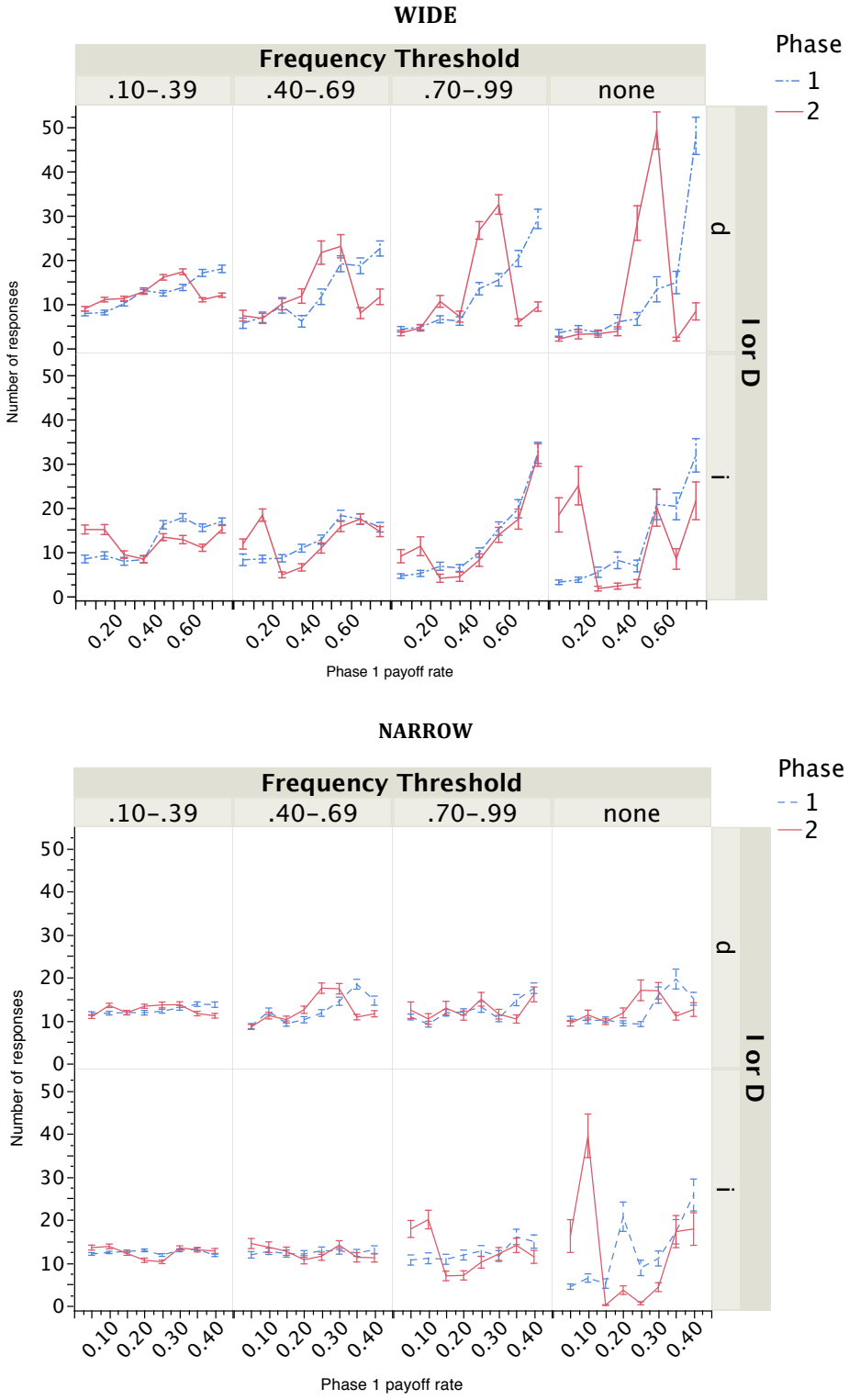


Figure 2. Number of responses on the eight buttons across phases (1 and 2) and conditions (wide/narrow, increasing/decreasing). The buttons were labeled based on their assigned RR in the first phase of the experiment (x-axis). Error bars are one standard error.

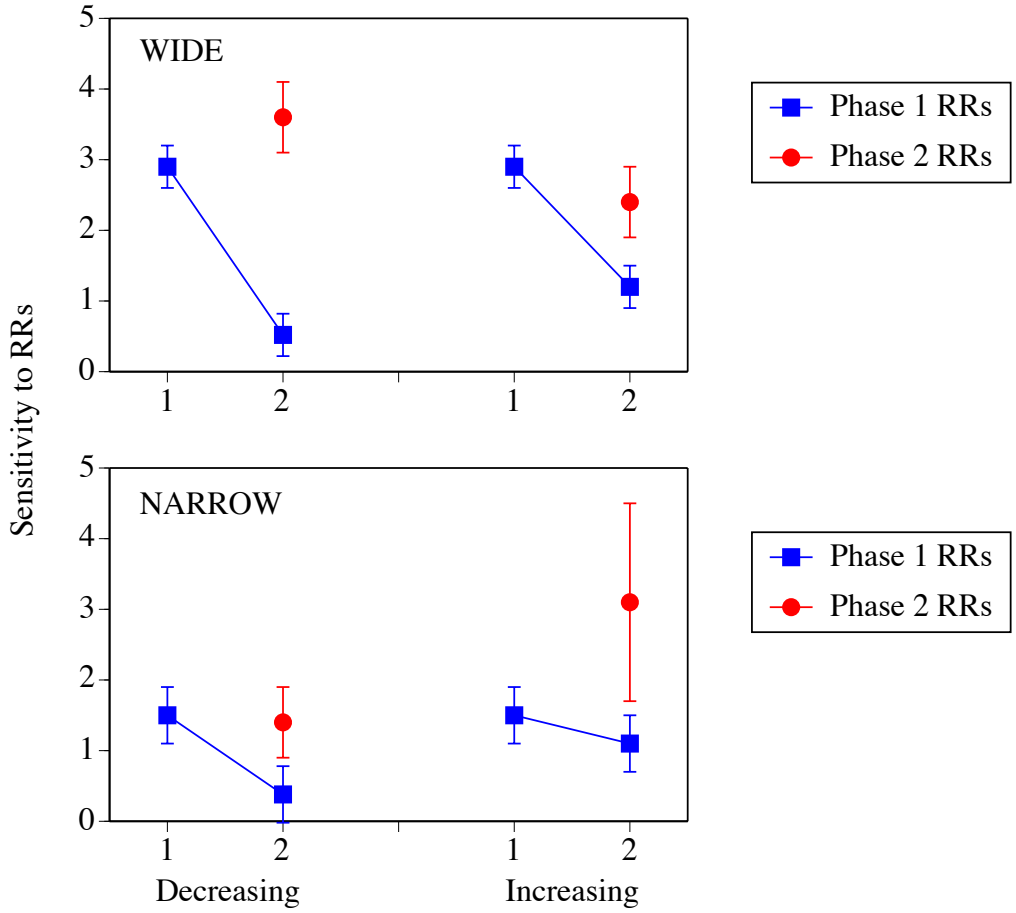


Figure 3. Sensitivity to the Phase 1 RR schedule in Phase 1, and sensitivity to the Phase 1 RR and the Phase 2 RR schedules in Phase 2. Phase (1 vs. 2) is plotted on the x-axis separately for the decreasing and increasing conditions. Error bars are one standard error.

A finer level of analysis is shown in Figure 5 in which entropy is plotted as a function of block. As hypothesized, exploration increased following the change in payoff structure for the decrease conditions but not for the increase conditions. The effect was confirmed by a 2 × 2 × 2 analysis of variance of wide/narrow, increasing/decreasing, and block (10 vs. 11) which revealed significant main effects and interactions. Central to the hypothesis is the significant three-way interaction ($F(1, 127) = 48.5$, $p < 0.01$) for which follow-up tests revealed an increase in entropy from block 10 to 11 for the decreasing conditions (increase for wide = 0.57, narrow = 0.25, $t(127) = 2.28$, $p < 0.05$ for the difference) but not the increasing conditions (decrease for wide = -0.15, narrow = -0.11, $p > 0.05$).

Finally, we examined individual differences across Phase 1 to determine whether the data shown in Figure 5 are representative of individuals or was the product of averaging a group of high exploiters with a group of high explorers. Figure 6 reveals that nearly all participants reduced their exploration as the phase continued and that individual performance spanned the entire range of entropy levels observed under each constraint level. Interestingly, nearly all participants were exploring more than was required by the frequency threshold. For example, under a 0.7 frequency threshold, a participant could avoid losing their preferred key with an entropy of about 0.8 to 1.0 (see Table 2). And yet, nearly every participant was still exploring at a higher level than required by the end of the first phase. Similarly, a participant subject to the 0.4 threshold could meet this criterion with a response distribution with an entropy as low as 1.5. Yet, every participant who

was assigned a threshold between 0.40 and 0.69 produced entropies much higher than 1.5 throughout the phase.

Reinforcement Earned

The key research questions involved the impact of the behavioral constraints on the amount of reinforcement earned. Because some people naturally showed more exploratory behavior even when there was no constraint to require it, we examined the relationship between entropy and reinforcement and not the frequency threshold and reinforcement. Figure 7 shows a scatterplot of the reinforcement earned in the 20 blocks of training for each participant in a condition.

Figure 7 reveals that there was a generally negative relationship between the level of exploration and the amount of reinforcement earned. This was true in all conditions, although weakest in the narrow conditions where exploration was naturally quite high. In none of the conditions or phases was there clear evidence that a level of entropy greater than zero increased the average reinforcement, in direct opposition to our hypothesis. The only benefit of higher exploration was its prevention of the lowest reinforcement rates that could occur if a participant fixated on a button with an RR schedule leaner than the average. As participants' entropy neared 3.0 (whether naturally or under constraint), the reinforcement earned converged on a single point that differed across conditions – the payoff rate for choosing equally among the buttons regardless of their payoffs.

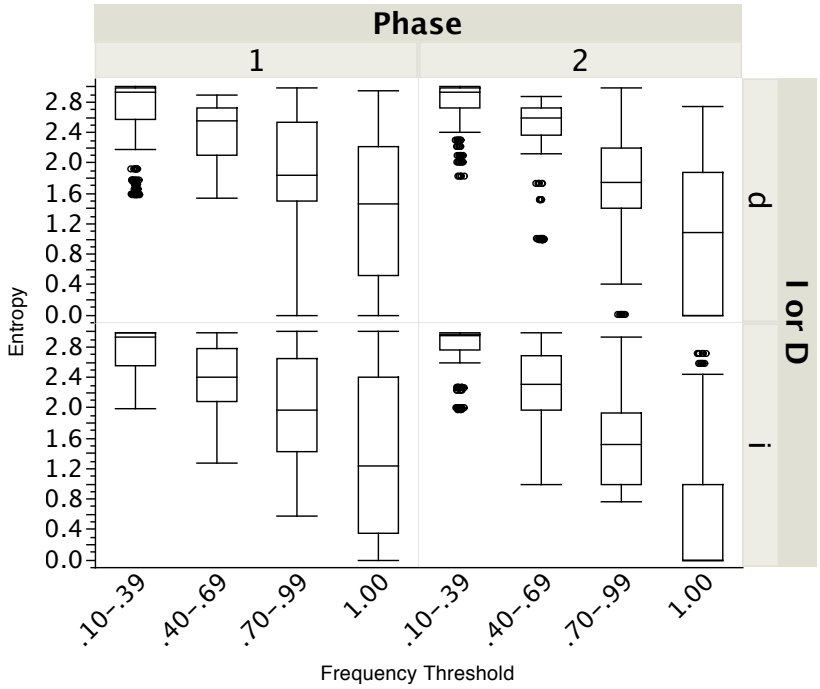
A closer look at this relationship is shown in Figure 8 in which we focus on behavior immediately preceding the change in the payoff structure (block 10) and the first three blocks after the change. Although the negative impact of greater exploration largely disappeared in the first block after the change (especially in the wide conditions), there was again no consistent evidence of an advantage of levels of exploration greater than zero.

Discussion

The experiment provided evidence that the relative frequency threshold constraints served to increase exploration while still producing discrimination among the button payoffs. However, we uncovered no evidence that introducing a frequency constraint actually helped our participants to learn the payoff structure after a change. Although participants were more likely to notice a decrease in the payoffs of the richest buttons by increasing their exploration, they eventually discovered the change even when the leanest buttons became richer (see Figures 2 and 5).

The absence of a beneficial effect of exploitation constraints may be a byproduct of the persistent exploration of nearly all of our participants in the first phase of the experiment. Participants only rarely showed pure exploitation in the first phase and then only toward the end of the phase (see Figure 6). This natural tendency to explore, even when it is not optimal to do so, is evidenced in the human tendency to probability match (for a recent review, see Vulkan, 2000). In contrast, many other species show maximizing (i.e., exclusive exploitation) when discriminating RR schedules (e.g., Herrnstein & Loveland, 1975). This species difference raises the possibility that response frequency constraints might be more beneficial when training non-human species. However, we are not quite willing to concede that frequency constraints are not effective at producing beneficial levels of exploration in humans; our results may be the consequence of certain aspects of our task.

WIDE



NARROW

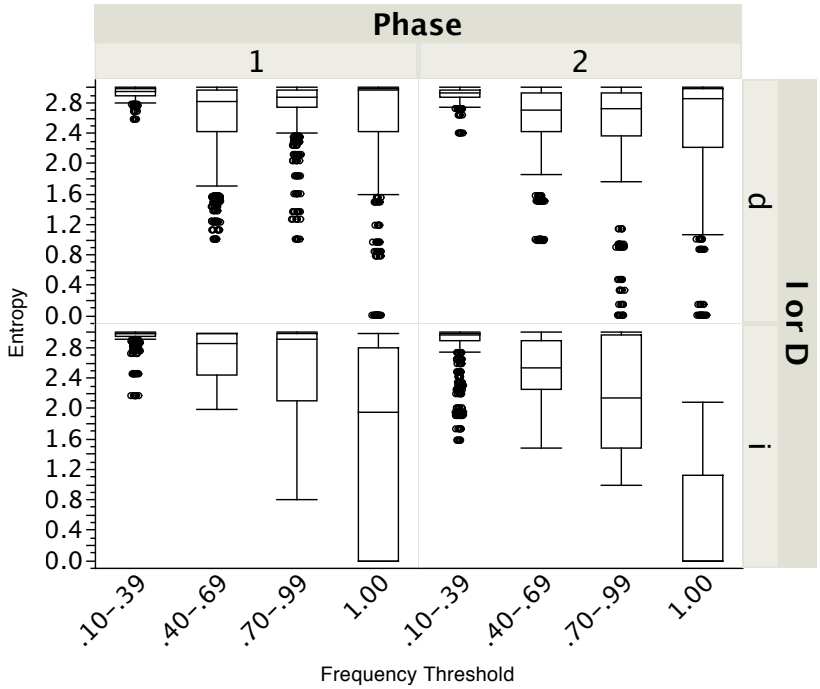


Figure 4. Boxplot of the entropy produced in each of the 20 100-trial blocks for each participant as a function of phase, wide (top) and narrow (bottom), and increasing/decreasing.

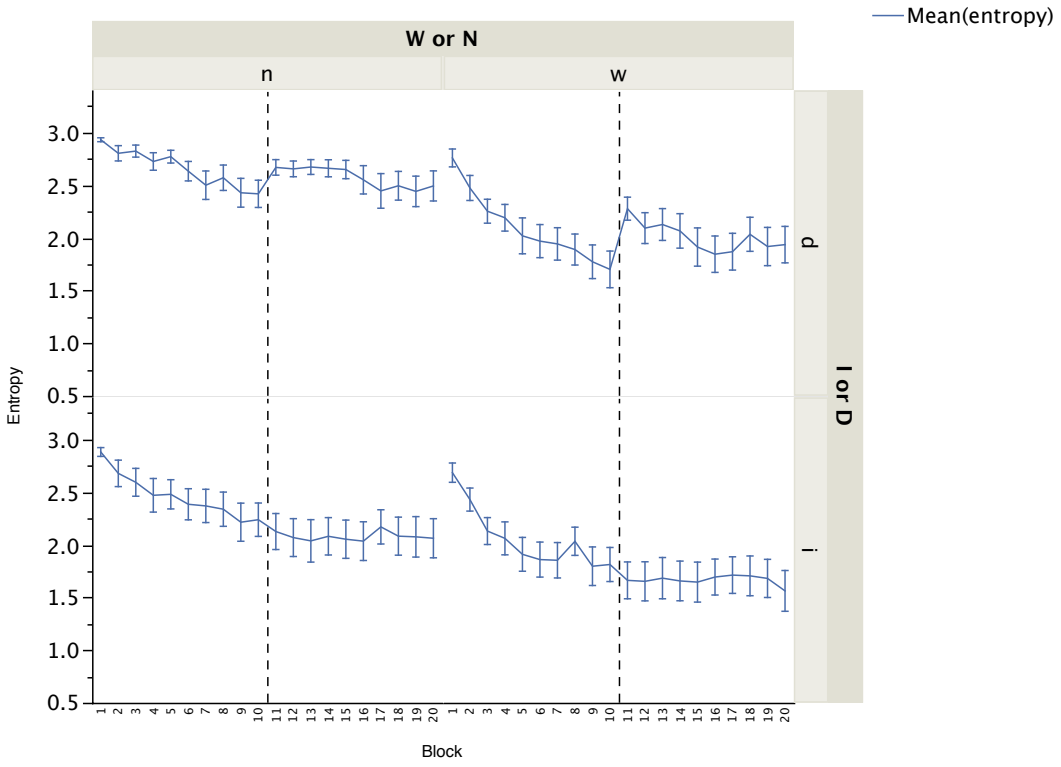


Figure 5. Changes in entropy as a function of 100-trial block of training in each of the four conditions. Error bars are one standard error.

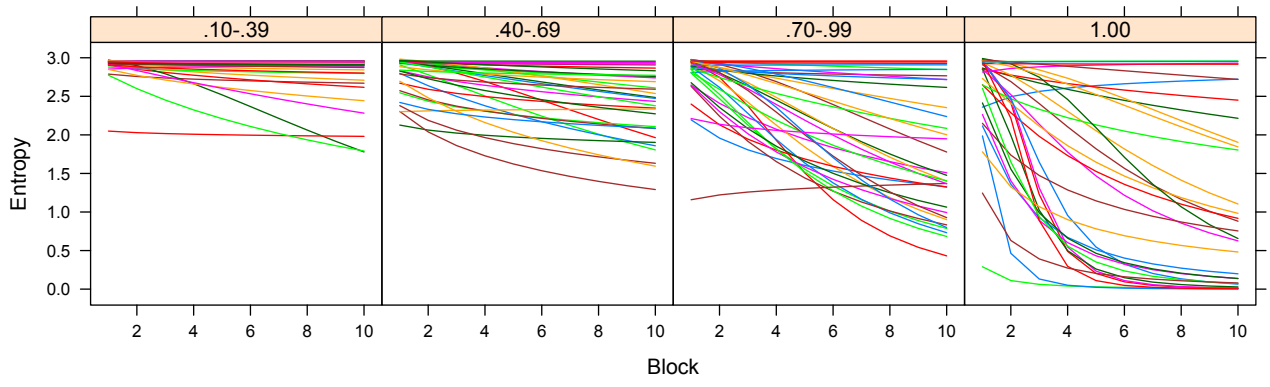
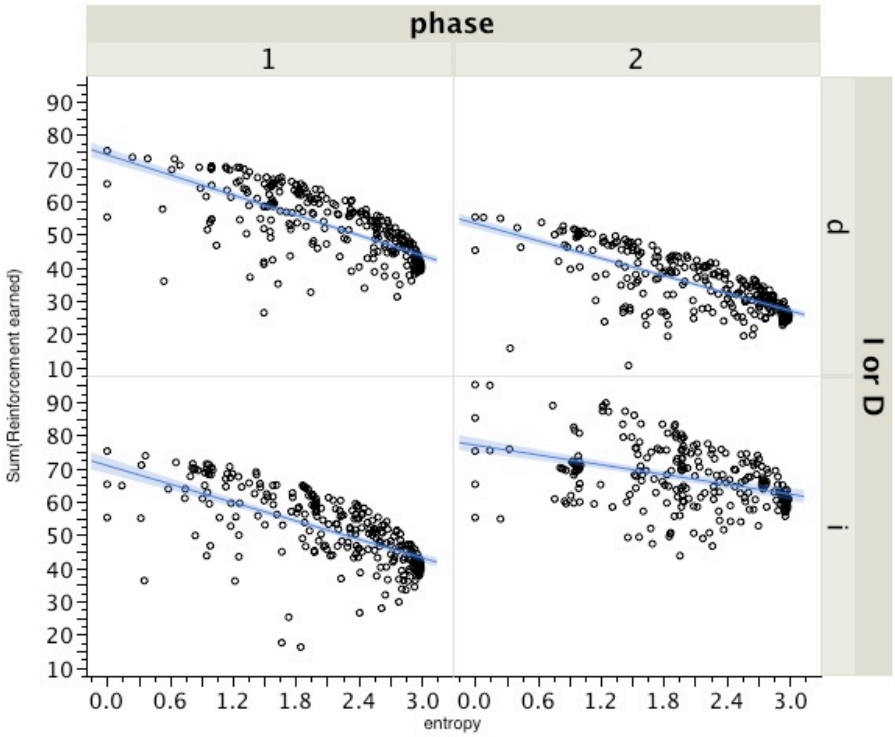


Figure 6. Logistic model fits of each participant's change in entropy as a function of block of training for each of the four response frequency threshold categories.

WIDE



NARROW

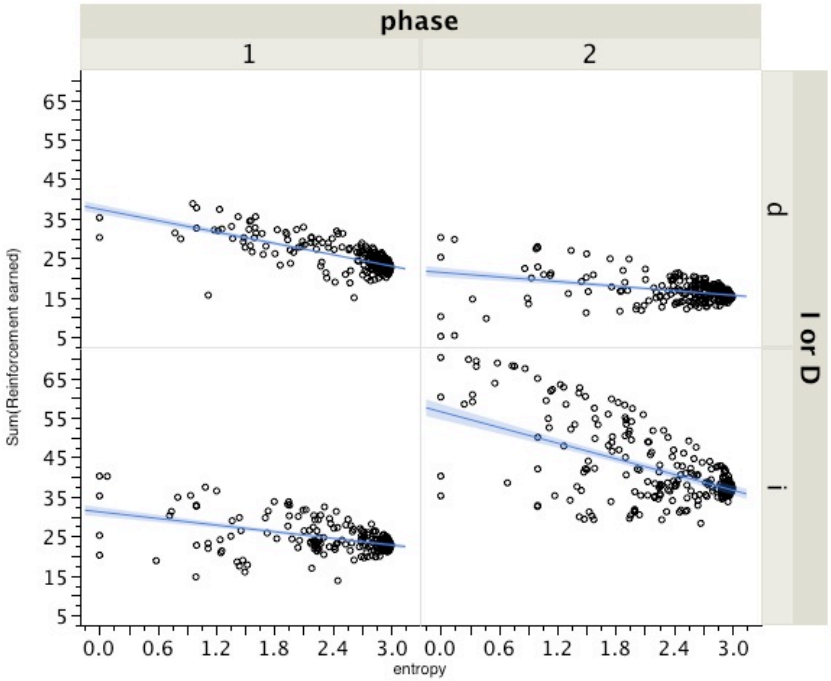


Figure 7. Reinforcement earned in each of the 20 100-trial blocks for each participant across the four experimental conditions.

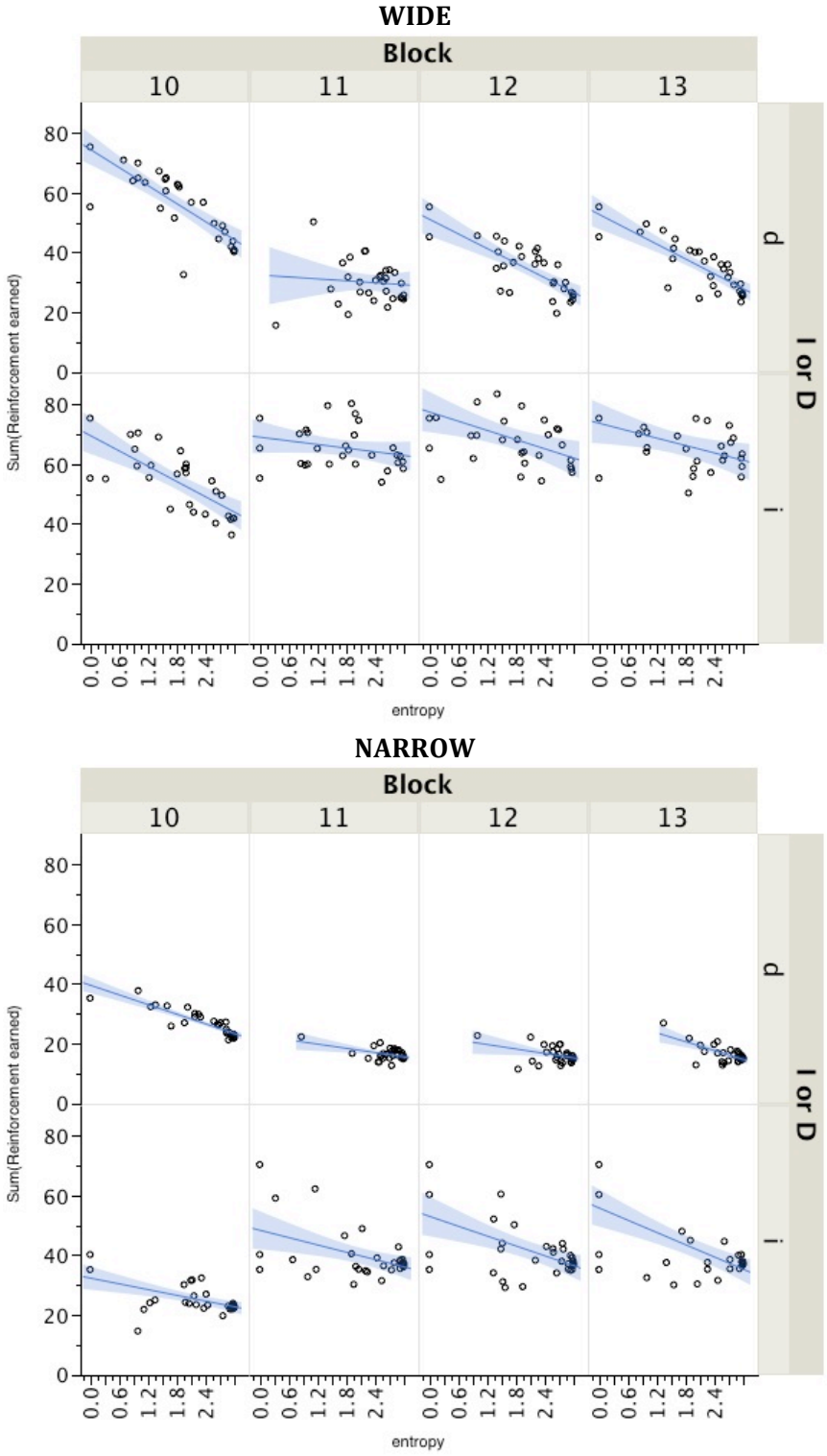


Figure 8. Reinforcement earned by each participant in the block immediately preceding the change in the payoff structure (block 10) and the three blocks immediately following the change (blocks 11-13).

Possible Reasons why Exploration in our Task was not Beneficial

Given prior evidence that there are environments in which reinforcement of variability has benefits (for a review, see Neuringer, 2004), the absence of such an effect in the 8-armed bandit task may be the byproduct of (a) the immediate availability of reward, (b) the payoff structure, (c) participants' limited task experience, (d) disappearance aversion, or (e) low switch costs. We will consider each of these possibilities in turn.

In the bandit task, the consequences of a response are immediately known, but in many other tasks a reward may not be available until a particular sequence of actions is produced (e.g., Neuringer, Deiss, & Olson, 2000). Reinforcement of variability has been a successful strategy to increase exploration for difficult sequences because the production of the correct sequence or sequences may occur too rarely to facilitate their reinforcement. Furthermore, in some complex contexts encouraging behavioral variability is tantamount to increasing creativity in the products of that behavior which can increase artistic value (Neuringer, 2004; Stokes, 2001).

It is also possible that certain payoff structures not included in our study may be more likely to produce exploitation at detrimental levels. For example, if one of the buttons in Phase 1 was reinforced at a high rate and the other seven buttons were never reinforced, participants may quickly exploit the only button that produces rewards and thus explore so rarely that changes to the reward structure for the non-reinforced buttons would go unnoticed. Our payoff structures in the wide and narrow conditions required a fairly high degree of exploration in order to identify the buttons' outcomes in a probabilistic environment, and this behavior may have created a tendency toward exploratory behavior even after the richest buttons had been identified.

Relatedly, this tendency to explore may be the result of the participants' relatively short exposure to the reward contingencies. During learning, exploration of RRs is a necessary part of even short-term success in a stationary environment. As Figure 6 revealed, participants' exploration steadily dropped during the first phase of the experiment when not under constraint. Thus, it is possible that extended training with the multi-armed bandit task may eventually produce high levels of exploitation in most of the participants that would be detrimental in certain non-stationary environments like our increasing conditions.

The excessive exploration in the face of a response frequency threshold might also be the result of avoidance of losing an option, a behavior termed *disappearance aversion* by Shin and Ariely (2004). Participants may have been so averse to losing their favored option that they overcompensated by exploring the other options at a level not required by the task constraints. The resulting high levels of exploration may have undermined our ability to detect any beneficial effects of sustained low levels of exploration.

Finally, the persistent exploration may be the result of the ease with which the chooser could move among options. Working a job that pays well enough rather than looking for a better paying one, continuing to shop at a particular store when you suspect better options may be out there, or leaving your investment portfolio unchanged when it is performing well enough are everyday examples of a switching cost that may suppress the level of exploration (cf. travel costs between resource patches in foraging, MacArthur & Pianka, 1966).

Implications

Whether individuals have greater experience with a radically changing environment, a completely stable one, or something intermediate may influence the level of exploration they adopt. Assuming people adopt a level best suited to their experience, this level cannot serve them optimally in all environments. If for

instance a person is accustomed to an environment that occasionally changes and the change is typically moderate, then a low level of exploration may be adopted even after the current set of contingencies have prevailed for an extended period of time. If the options never change, then they have given something up in their continued exploration. And if the options do change, but the change is a difficult one to detect (such as an improvement in a previously very poor option) their low level of exploration may not be enough to take advantage of it, and again they have given something up. There are countless possible change types and magnitudes, each with an attendant appropriate level of exploration for maximum reward to be earned. How adaptable we are is an important part of how well we survive and prosper. It stands to reason that we should pursue an understanding of the conditions that control our learning in the trial-and-error environments in which we so often find ourselves.

References

Antonitis, J. (1951). Response variability in the white rat during conditioning, extinction, and reconditioning. *Journal of Experimental Psychology, 42*, 273-281.

Balsam, P. D., Deich, J. D., Ohyama, T., & Stokes, P. D. (1998). Origins of new behavior. In W. O'Donohue (Ed.), *Learning and behavior therapy*. Boston, MA: Allyn & Bacon.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193-217.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959-988.

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, NY: Oxford Press.

Jensen, G., Stokes, P. D., Paterniti, A., & Balsam, P. D. (2013). Unexpected downshifts in reward magnitude induce variation in human behavior. *Psychonomic Bulletin and Review, 21,* 436-444. doi: 10.3758/s13423-013-0490-4

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research, 4*, 237-285.

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.

MacArthur, R. H., & Pianka, E. R. (1966). On optimal use of a patchy environment. *The American Naturalist, 100*, 603-609.

Neuringer, A. (2004). Reinforced variability in animals and people: Implications for adaptive action. *American Psychologist, 59*, 891-906.

Neuringer, A., Deiss, C., & Olson, G. (2000). Reinforced variability and operant learning. *Journal of Experimental Psychology: Animal Behavior Processes, 26*, 98-111.

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

Racey, D. E., Young, M. E., Garlick, D., Pham, J. N., & Blaisdell, A. (2011). Pigeon and human performance in a multi-armed bandit task in response to changes in variable interval schedules. *Learning & Behavior, 39*, 245-258.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Shin, J., & Ariely, D. (2004). Keeping doors open: The effect of unavailability on incentives to keep options viable. *Management Science, 50*, 575-586.

Stokes, P. D. (2001). Variability, constraints, and creativity: Shedding light on Claude Monet. *American Psychologist, 36*, 355-359.
Stokes, P. D., & Balsam, P. (2001). An optimal period for setting sustained variability levels. *Psychonomic Bulletin and Review, 8*, 177-184.
Stokes, P. D., & Balsam, P. D. (2003). Effects of early strategic hints on sustained variability levels. *Creativity Research Journal, 15*, 331-341.
Stokes, P. D., & Harrison, H. M. (2002). Constraints have different concurrent effects and after effects on variability. [Comparative Study]. *Journal of Experiment Psychology: General, 13*, 552-566.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. London, England: MIT Press.
Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*, 101-118.
Young, M. E., Cole, J. J., & Sutherland, S. C. (2012). Rich stimulus sampling for between-subjects designs improves model selection. *Behavior Research Methods, 44*, 176-188.

Financial Support: This work had no external sources of financial support.
Conflict of Interest: All authors of this paper declare no conflict of interest.

Submitted: *September 6th, 2013*
Resubmitted: *December 5th, 2013*
Accepted: *January 20th, 2014*