A Re-examination of Melioration and Rational Choice

RICHARD J. TUNNEY* and DAVID R. SHANKS
University College London, UK

ABSTRACT

We examined how people allocate choices between two alternatives when the payoff from each alternative varied as a function of the allocation of recent choices. On any one trial alternative A had a higher immediate payoff than alternative B, but across all trials B had a higher overall payoff than A. Rational choice theory requires that participants allocate all their responses to the alternative with the greatest overall payoff irrespective of which has the higher immediate payoff. Melioration, in contrast, proposes that participants are motivated to choose the alternative with the higher immediate payoff, irrespective of the consequences for future returns. We report four experiments in which we varied the nature of the payoffs. Participants exhibited self-control consistent with rational choice theory when payoffs varied in magnitude, but exhibited impulsiveness consistent with melioration when the payoffs varied in probability. Finally, we show that impulsivity when payoffs varied in probability can be overcome following un-reinforced practice. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS choice; inter-temporal choice; melioration; maximization; self-control; rationality; learning; reinforcement

Faced with a choice between two alternatives, one of which has a payoff with higher value or utility than the other, both animals and humans will predominantly choose the alternative with the higher value.1 This seems perfectly rational, in the sense that the agent gains as much from the situation as possible, but it would seem irrational if they continued to do so in a repeated-choice situation when this led to a decrease in overall utility. The problem is one of self-control—preference for a small immediate reward over a larger delayed reward is impulsive (Ainslie, 1975) and has been attributed to a principle of behavioral allocation known as melioration (Herrnstein and Vaughan, 1980). In many situations people find it easy to make decisions that maximize some future payoff. But in others the lure of an immediate reward can be difficult to overcome even when it can have negative consequences in the future. We know, however, that the tobacco addict can
overcome the desire to smoke and the gluttonous can diet. But, as these examples suggest, self-control is not always easy and must often be learned. In this paper we examine the conditions under which melioration prevails and how self-control can be learned. We begin by describing a repeated-choice situation designed to determine whether individual choices are motivated to maximize immediate or overall utility (the sum of all payoffs), and show how in some circumstances different theories of choice make different predictions concerning equilibrium behavior.

By way of introduction consider the following well-known menu problem. The hotel where you take your vacation has but one restaurant with just two options: fish cake and sushi. Since the two are relatively equal in nutritional value, and the meals are prepaid, the two options differ solely in hedonic value. The fish cake is relatively bland and its hedonic value changes little whether it is chosen every day or only occasionally. On the other hand, sushi is, to European tastes at least, relatively high in hedonic value. Due to its richness, however, if you were to choose sushi on several successive occasions its hedonic value would decrease rapidly and may even decrease your desire for fish, including fish cake, altogether leading to a low overall payoff (the total hedonic value experienced over the course of the vacation). Conversely, choosing fish cake more often than choosing sushi would increase the value of sushi on those occasions when it is chosen and maximize the overall payoff. Over the period of your vacation how should you choose what to eat on each day? Let us examine in more detail how choices might be allocated in situations such as these.

**BEHAVIORAL EQUILIBRIA AND CHOICE**

Consider the payoff schedules shown in Exhibit 1 in which hypothetical values for two alternatives A and B are presented as a function of the number of times alternative A is chosen. This example, sometimes known as the Harvard Game (Herrnstein et al., 1993), has the same structure as the menu problem. On the majority of occasions the payoff that can be expected by choosing alternative B (sushi) is higher than the payoff that can be expected by choosing A (fish cake). But each time alternative B is chosen the value of the payoff that can be expected for that alternative on subsequent occasions decreases, eventually reaching a level lower than that of alternative A. However, each time alternative A is chosen the value of the payoff that can be expected by choosing alternative B on subsequent occasions is increased.

Rational Choice Theory suggests that an organism is motivated to choose an allocation that maximizes the overall payoff, irrespective of which option has the highest immediate payoff. Although Rational Choice Theory is relatively silent about how each individual choice is made, it requires that they be allocated such that behavior reaches the maximizing equilibrium shown as a point on the dotted line in Exhibit 1. For instance, to maximize the overall payoff in a situation such as Exhibit 1, and indeed our menu problem, alternative A should be chosen on the majority of occasions and alternative B should only be chosen occasionally. That is, in order to maximize the overall payoff, one should abstain from consistently choosing the alternative with the highest immediate payoff.

In contrast to this prescriptive analysis, a good deal of research indicates that in situations similar to the one shown in Exhibit 1 animals often fail to maximize expected utility (Davison and McCarthy, 1988). Early studies indicated that humans behave in the same way. To explain these findings Herrnstein and Vaughan (1980) suggested that the motivation for each choice is to select the alternative that has the higher immediate payoff irrespective of which alternative (or combination of alternatives) would maximize the overall payoff. Herrnstein and Vaughan (1980) called this tendency *melioration*, although it may be more familiar to many readers as a restatement of The Law of Effect: ‘A rise or fall in the reinforcement of a response causes the rate of occurrence of the response to change in the same direction. Should there be an inequality in unit returns

---

2Where equilibrium is steady-state behavior that does not significantly change over time. In many studies behavior on the final trial block is taken as the dependent variable.
from two alternatives, behavior “meliorates”, redistributing itself toward the more lucrative, hence stronger, alternative’ (Herrnstein, 1990, p. 219). For example, if $B$ tends to have a higher immediate payoff than $A$ then the allocation of responses shifts towards $B$. If some distribution of responses between $A$ and $B$ earns equal immediate payoffs, then an equilibrium, even if it is sub-optimal, has been reached and is maintained as the individual continues to choose the alternative with the higher immediate payoff (see the melioration equilibrium in Exhibit 1). In situations such as Exhibit 1, melioration returns a lower overall payoff than the maximizing equilibrium. In a repeated situation the motivation for choice can only be inferred when the choices are consistent over time. We would not conclude, for instance, that a participant who for an initial period chose the meliorating alternative in Exhibit 1 and later chose the maximizing alternative had different motives at different times. Rather, we would conclude that the early behavior was, like the later behavior, motivated to maximize payoffs but that the participant’s expectations of future payoffs had changed as knowledge of the payoff structure had developed. This process enables us to examine the conditions under which people learn to maximize expected utility.

In everyday life people are clearly able to make repeated choices that maximize expected utility. For instance, many people save rather than spend. But in many cases we fail to exhibit self-control: Despite being informed of the consequences for their long-term health many people continue to smoke, drink excessively, or eat a fatty diet.

Situations in which people meliorate despite knowledge of the consequences constitute a choice anomaly (Friedman, 1998) in the sense that behavior is inconsistent with rational choice theory. The dynamics of inter-temporal choice have also been examined in the form of one-shot decisions involving choices between hypothetical gambles or lotteries in which either the delay or probability of the payoffs vary (e.g. Kahneman and Tversky, 1979; Thaler, 1981; Loewenstein and Prelec, 1992). Such experiments have indicated a number of reasons why people might appear to meliorate: the value of a payoff might be discounted as a function of...
the delay of receiving it. However, experiencing the consequences of a behavior may be different to simply being informed of those consequences, and often the consequences of behavior only emerge over time. For example, base-rate neglect is diminished in situations where people experience rather than are simply informed of the probabilities (Goodie and Fantino, 1999). Situations in which the consequences of making choices are unknown or have not been experienced present a twofold problem: the payoff schedules associated with each choice must be learned before choices can be allocated strategically. Learning the payoff schedules inevitably involves making choices that may at first appear anomalous. As Friedman (1998, p. 942) points out, ‘irrational choices arising from incomplete learning do not imply the need to modify standard choice theory’. Similarly, Hertwig, and Ortmann (2001) note that before we conclude that human behavior is ‘irrational’ we must be sure that the rewards associated with optimal and sub-optimal equilibria are sufficiently different that the utility of adopting an optimal strategy is greater than the cost of finding that strategy.

Consequently the nature of the payoff plays a critical role in determining behavior, and this may in part explain why animals tend to meliorate. In the animal laboratory payoffs invariably consist of food received shortly after each choice. For food-deprived animals there may be some advantage to feeding whenever possible. Forzano and Logue (1994) found that humans also exhibited meliorating behavior when food was delivered as a reward during the course of the experiment. However, when the rewards consisted of points that could be exchanged after the experiment for either food or money then self-control prevailed. Thus both humans and animals may exhibit meliorating behavior when hungry and satiation is immediate.

Perhaps self-control is easier to acquire when complete abstinence from one option is required to maximize the overall payoff. Although many problems have the structure shown in Exhibit 1 where maximization requires the participant to distribute some proportion of responses to both schedules, parallel utility schedules, such as those shown in Exhibit 2, might be easier for subjects to discriminate because both

Exhibit 2. Parallel schedules used in Experiment 1. The payoff on both schedules is plotted as a function of the proportion of responses allocated to the Max button. The average payoff indicates the average payoff per trial as a function of the proportion of choices to each schedule over the preceding 10 trials. On any one trial choosing Mel returns a higher immediate payoff than Max. Consistently choosing Mel reduces the value of each subsequent payoff and consequently the overall or sum of payoffs. Consistently choosing Max increases the value of subsequent payoffs and maximizes the overall payoff. The maximizing equilibrium is reached when all responses are allocated to the Max button and over a 100 trial block returns an overall payoff of £1.40. The melioration equilibrium is reached when all of the responses are allocated to the Mel button and over a 100 trial block returns an overall payoff of £0.90
maximization and melioration require exclusive choice of one or the other schedule. Since the two schedules do not cross all meliorating responses will be allocated to the alternative that has the higher immediate value (Mel) despite the consequence of reducing the value of that and the other alternative. In contrast, maximization requires that all responses be allocated to the alternative with the lower immediate value (Max) because consistently choosing the Max alternative returns a higher payoff than consistently choosing the Mel alternative. In this way parallel schedules such as these sharply distinguish between the two equilibria making them easier for participants to find.

Herrnstein et al. (1993, Experiment 3) report such a comparison. In one condition, maximization required exclusive preference for just one alternative and melioration for the other. In another condition maximization required 70% of responses to be allocated to one alternative and the remaining 30% to the other. When the task required exclusive preference for one button (parallel schedules similar to Exhibit 2), participants allocated the majority of their responses to the maximizing alternative, but when the task required a distribution of responses to both alternatives (crossing schedules similar to Exhibit 1) they allocated about 50% of their responses to each alternative suggesting a failure to discriminate between schedules. Clearly the structure of the task influences how easy it is for subjects to maximize their overall payoff. Self-control is easier to learn when complete abstinence is required than when it requires a more subtle redistribution behavior.

As is the case with many decision problems, feedback can play an important role. Receiving payoffs alone may do little more than indicate how well a participant is doing. In order to make choices that maximize payoffs people may need to be informed of how well they could be doing (Balzer, Doherty, and O’Connor, 1989). Recently, Warry, Remington, and Sonuga-Barke (1999) investigated the extent to which feedback encouraged participants to maximize utility on parallel payoff schedules. In the prospective information group participants were told the values of the payoffs they would receive for each alternative. For example the screen would indicate that choosing left would return 5 points and right 8 points, but not that choosing the higher alternative would decrease its value on the next choice. In the social comparison group participants were shown, after each 10 trial block, how many points ‘an expert’ would have won over the same period where expert behavior was the maximizing strategy. In addition one group of participants received both additional sources of information and another received none. Each group that received feedback allocated increasingly more responses to the maximizing alternative as learning progressed, although they did so at different rates and did not reach a maximizing equilibrium at the end of the experiment (200 trials). On the other hand, the group of participants who received no feedback settled on a meliorating equilibrium.

OVERVIEW OF EXPERIMENTS

In contrast to the animal literature the few experiments conducted with humans are encouraging in that given an appropriately structured environment participants can learn to allocate their responses strategically in order to maximize overall payoff. Laboratory research has indicated some of the factors by which people maximize their payoffs: people must be aware that the payoff for one behavior is lower than the payoff for another with the difference between the two being sufficiently motivating. Self-control is learned through sufficient exposure to the environment and with informative feedback. Despite these encouraging laboratory findings, outside the laboratory people persist in impulsive and often self-destructive behaviors, even when as is the case with health issues the consequences are transparent. The present research was conducted to further analyze the claim that human choice is necessarily sub-optimal and to examine whether there are indeed environments in which melioration leads to a stable equilibrium. Our research is motivated by Friedman’s (1998, p. 941) assertion that ‘every choice “anomaly” can be greatly diminished or entirely eliminated in appropriately structured environments’. We therefore constructed simple choice environments in which participants received substantial financial incentives, long periods of training over several days, and were provided with feedback concerning optimal rates of payoff. In particular we examined behavior in two
environments with different payoff structures. In Experiment 1 participants’ behavior varied the magnitude of the payoffs they would receive, and in Experiments 2–4 their behavior varied the probability of payoff. In each of the experiments described below participants were asked to make a number of consecutive choices between two alternative buttons on a computer screen. Each experiment had a payoff structure similar to that shown in Exhibit 2. The payoff on one button (Max) increased with the proportion of responses allocated to that button over the past ten choices. The payoff on the other button (Mel) decreased with the proportion of responses allocated to it. Allocating a small proportion of responses to the Mel button returned a higher immediate rate of reward than allocating all of the responses to the Max button. However, because the payoff on the Mel button decreased with the proportion of responses allocated to it the returns quickly diminished to a level below that returned by the Max button. The payoff schedules were arranged so that consistently choosing the Max button and abstaining from the Mel button returned a higher cumulative payoff than any other combination of choices.

**EXPERIMENT 1**

In Experiment 1 participants were trained for 1000 trials and received a payoff for each choice that they made. The magnitude of each payoff varied according to how participants allocated their choices over the preceding ten trials. Participants received feedback about how well they were doing, relative to a maximizing strategy, after every 100 trials. This was similar to one condition of an experiment described by Herrnstein et al. (1993, Experiment 3). Herrnstein et al. tested participants over 400 trials using parallel payoff slopes that varied the magnitude of payoff. They observed that participants tended towards maximization.

**Method**

**Participants**

Eight members of the University College London community volunteered for this study, three male and five female. Participants did not receive a turn-up fee but instead received the cumulative payoffs from the experiment. They were informed that in all cases they would win at least £3.00 but that they could win as much as £7.00 on each session.

**Apparatus**

Participants were tested individually in a sound-proofed testing cubicle. An IBM compatible PC controlled the experiment. The screen display is presented in Exhibit A1 in the Appendix. Two buttons marked ‘Left’ and ‘Right’ were displayed horizontally and separated by a gap of 3 cm. Above these two buttons and in the center of the screen a large outcome box, marked ‘Payoff on last trial’, indicated the magnitude of payoff in pounds for the immediately preceding choice. To the right of the payoff box a smaller box, marked ‘Last choice’ indicated whether the last choice was left or right. At the top center of the screen a large outcome box, marked ‘Cumulative winnings’ indicated the sum of all payoffs, and a smaller box to the right indicated the last trial number. Participants made their choices by selecting one button or the other using the mouse.

**Design and procedure**

At the start of the experiment the following instructions were displayed on the computer screen. These instructions were also available on a sheet of paper throughout the experiment.

Thank you for agreeing to take part in this experiment. Your task is simple. On each of 500 trials you have to choose between 2 buttons, marked Left and Right. Simply click on a button to register your choice. Each choice will earn you a small payment which is given in fractions of a pound: thus 0.01 is 1 p. You
will be shown your payment on each trial as well as your cumulative winnings. That’s all there is to it—just try to win as much money from the computer as you can. At the end of the experiment you will be paid the amount you have won. Take as much time as you wish and please do not write anything down during the experiment.

To initiate each trial a yellow label marked ‘CHOOSE!’ appeared on screen and the two buttons were enabled (see Exhibit A1). Following each choice the two buttons were disabled and the last choice box and trial box were updated. To ensure that participants were aware that the payoff box indicated payoff received on the preceding trial rather than expected payoff the box cleared for half a second before being updated. Participants were then prompted to make another choice and the buttons were enabled. The experimenter remained in the testing cubicle for the first five trials to ensure that participants understood the task. Participants were given 500 consecutive trials on each session, lasting approximately 20 minutes. At the end of each session a screen displayed the cumulative payoff which participants received as payment. The experiment consisted of 1000 trials spread over two sessions that were separated by an interval of one week. Each session was identical: For each participant the Max button was the same in both Session 1 and Session 2. Feedback was provided every 100 trials where participants were prompted to take a short break and informed of their cumulative winnings and what their winnings would have been had they adopted the optimal strategy, but not what the optimal strategy was (see Exhibit A2). The winnings from the optimal strategy were calculated as the cumulative payoffs from consistently choosing the Max button. As Exhibit 2 shows, the cumulative payoff for consistently choosing Max was £1.40 for every 100 trials.

At the start of session two participants were informed that the second session was in every way identical to the first.

Payoff schedules
Participants received a payoff for every choice that they made. Consistently choosing one button (Max) over the other (Mel) returned the highest cumulative payoff. For half the participants left was allocated as the Max button and right as the Mel button.

The magnitude of payoff for any one choice was determined by the current choice and the proportion of responses allocated to the Max button over the preceding ten trials. Exhibit 2 shows the payoff schedules associated with each choice button. The magnitude of payoff on the Max button = 0.004 + (0.01* (proportion of responses allocated to the Max button in the preceding 10 trials)). The magnitude of payoff on the Mel button = 0.009 + (0.01* (proportion of responses allocated to the Mel button in the preceding 10 trials)). For example, if participants consistently chose the Max button over a period of ten consecutive trials, they would receive a payoff of £0.014 if they chose Max again. If, after 10 consecutive Max choices, participants decided to switch to the Mel button they would receive a payoff of £0.019. On the other hand if participants consistently chose the Mel button over a period of ten consecutive trials they would receive a payoff of £0.009 if they chose Mel again. If, after a period of 10 consecutive Mel choices, participants decided to switch to the Max button they would receive a payoff of £0.004.

Over the 500 trials of each session consistently choosing Max would return a cumulative payoff of £7.00, whereas consistently choosing Mel would return a cumulative payoff of £4.50, a difference of £2.50.

Results and discussion
Participants earned a mean of £5.63 (SD = 0.48) over Session 1 and a mean of £6.39 (SD = 0.63) over Session 2. The mean proportion of responses allocated to the Max button during each block are shown in

---

3Participants in each experiment were asked not to talk about the experiment to other potential volunteers.
Exhibit 3. A repeated-measures ANOVA with Block as the within-subjects factor revealed clear evidence of a trend towards maximization as learning progressed: $F(9, 63) = 4.66, MSE = 0.06, p < 0.01$.4

The distribution of $P(\text{max})$ values for the final block of each session (5 & 10) is shown in Exhibit 4. At the end of Session 1 participants had clearly not reached asymptote and only two participants allocated more than 95% of their responses to the Max button during the last block. Indeed the overall proportion of responses allocated to the Max button during the final block did not differ from a chance distribution ($t < 1$). However, each participant’s allocation of maximizing responses increased and overall reliably more Max button responses than would be expected by chance were observed during the last block of Session 2: $t(7) = 3.18, SD = 0.27, p < 0.02$, and five (63%) participants allocated over 95% of their responses to the Max button during this period.

The results of Experiment 1 indicate that after 1000 trials the tendency to maximize payoff is greater than the tendency to meliorate on schedules that vary magnitude of payoff. Indeed the majority of participants plainly found the optimizing strategy. Although there were individual differences and some participants did not reach the maximizing equilibrium these are likely to be due to differences in learning rate rather than...
strategy since each participant’s allocation of maximizing choices increased between Sessions 1 and 2. In the next experiment we repeat the procedure used in Experiment 1 but use a different schedule in which the magnitude of payoff is fixed and instead vary the probability of payoff.

**EXPERIMENT 2**

Unlike the schedules examined in Experiment 1, outside the laboratory the payoffs for choices in which people fail to exhibit self-control are often stochastic. Sub-optimal decision making about, and based on, probability of outcome is well documented in both one-shot (e.g. Loewenstein and Prelec, 1992; Thaler, 1981) and repeated-choice situations (e.g. Edwards, 1961; Myers, 1976; Shanks, Tunney, and McCarthy, 2002; Vulkan, 2000). In Experiment 2 we examine whether behavior differs when the payoffs for choices are probabilistic. Participants did not receive a payoff on every trial. Instead the probability of receiving a fixed magnitude reward on any one trial differed according to a schedule (shown in Exhibit 5) that would return similar cumulative payoffs as Experiment 1. Thus the motivation to maximize would be similar in both environments.

**Method**

**Participants**

Twelve members of the University College London community volunteered for this study, six male and six female. Participants did not receive a turn-up fee but instead received the cumulative payoffs from the

Exhibit 5. Parallel schedules used in Experiment 2. These schedules have an identical structure to those shown in Exhibit 2 with the exception that the allocation of choices varies the probability of receiving a fixed magnitude payoff of £0.02. The payoff on both schedules is plotted as a function of the proportion of choices to each schedule over the preceding 10 trials. On any one trial the probability of payoff for choosing Mel is 0.33 higher than choosing Max, but consistently choosing Mel reduces the probability of payoff to 0.33, while consistently choosing Max increases the probability of payoff to 0.66. Over the course of 100 trials meliorating behavior returns an average of 33 payoffs, while maximizing behavior returns an average of 66 payoffs.
experiment. They were informed that in all cases they would win at least £3.00 on each session but that they could win up to £7.00.

**Apparatus**
The screen displays were identical to the ones used in Experiment 1 (Exhibits A1 and A2).

**Design and procedure**
At the start of the experiment participants received the same instructions that were used in Experiment 1 with the exception that the phrase ‘each choice will earn you a small payment’ was replaced with the phrase ‘each choice may earn you a small payment’ to allow for the change in payoff schedules. These instructions were also available on a sheet of paper throughout the experiment. Each of the two sessions consisted of 500 consecutive trials lasting approximately 20 minutes. As in Experiment 2 feedback was provided every 100 trials when participants were prompted to take a short break and informed of the cumulative payoffs and what their cumulative payoffs would have been had they adopted the optimal strategy. The value of each payoff was fixed at £0.02 and varied in probability. Payoff from the optimal strategy was determined by choosing Max on each choice. As Exhibit 5 shows, the probability of receiving payoff for choice when consistently choosing Max was 0.66, so the average optimal payoff after every 100 trials was £1.32.

**Payoff schedules**
Unlike Experiment 1, in Experiment 2 participants did not receive a payoff for every choice that they made. The magnitude of payoff associated with each button was fixed at £0.02, but the probability of receiving that reward on any one was trial was varied. The probability of receiving payoff for Max choice = 0 + (0.66* proportion of time on Max), and the probability of receiving payoff for Mel choice = 0.33 + (0.66* proportion of time on Max). As Exhibit 5 shows, the distribution for Mel fell in the interval [0.33, 0.99] and the distribution for Max fell in the interval [0.00, 0.66] with the shape and mean of the distributions determined by the participant’s choice allocation.

For example, if participants consistently chose the Max button over a period of ten consecutive trials, then the probability of receiving a payoff of £0.02 if they chose Max again would be 0.66. If, after 10 consecutive Max choices, participants decided to switch to the Mel button the probability of receiving £0.02 would be 0.99. On the other hand, if participants consistently chose the Mel button over a period of ten consecutive trials the probability of receiving a payoff of £0.02 would be 0.33 if they chose Mel again. If, after the same period of 10 consecutive Mel choices, participants decided to switch to the Max button the probability of receiving £0.02 would be 0.00.

Over the 500 trials of each session consistently choosing Max would return an average cumulative payoff of £6.60, consistently choosing Mel would return a cumulative payoff of £3.30, a difference of £3.30.

**Results and discussion**
Participants earned a mean of £4.39 ($SD = 0.44$) over Session 1 and a mean of £4.49 ($SD = 0.59$) over Session 2. The proportion of maximizing responses for each block is shown in Exhibit 6. A repeated-measures ANOVA with Block as the within-subjects factor revealed no reliable change in the allocation of responses across blocks ($F < 1$): Unlike the previous experiment there was no trend towards maximization on either session.

Exhibit 7 shows the distribution of maximizing responses during the last block of each session. In the final block one participant allocated more than 95% of responses to the Max key, but only two allocated more than half. The remainder allocated the majority of their responses to the Mel key. Overall participants allocated
more responses to the Mel button than would be expected by chance during the last block of each session ($t(7) = -2.71$, $SD = 0.27$, $p < 0.02$ for Session 1 and $t(7) = -2.31$, $SD = 0.26$, $p < 0.04$ for Session 2).

Exhibit 8 shows the allocation of responses across experiments. A comparison of the change in response allocation between Experiments 1 and 2 revealed a reliable effect of Block ($F(9, 162) = 4.18$, $MSE = 0.04$, $p < 0.01$), a reliable effect of Experiment ($F(1, 18) = 19.12$, $MSE = 0.20$, $p < 0.01$) and a reliable interaction between Block and Experiment ($F(9, 162) = 4.32$, $MSE = 0.04$, $p < 0.01$), indicating that the trend towards melioration was much slower in Experiment 2 than was the trend towards maximization in Experiment 1. The allocation of responses in Experiments 1 and 2 were reliably different at the end of Session 1 ($t(18) = 2.34$, $p < 0.03$), and at the end of Session 2 ($t(18) = 3.95$, $p < 0.01$). Although these analyses involve cross-experiment comparisons the experiments were identical in all respects apart from the fact that in one case the payoffs differed in magnitude and in the other in probability. Otherwise, all that differed were the participants but even these were recruited via the same sampling process.

In Experiment 2 we were unable to detect changes in behavior over time despite a larger sample size than Experiment 1 (where such an effect was observed). However, the sample was sufficiently powerful to demonstrate that participants were reliably closer to the melioration than the maximization equilibrium at the end of Session 2. We conclude from this that learning of the payoff schedules is slower in a stochastic than a deterministic environment. In the next experiment, to increase the likelihood of detecting changes in behavior

<table>
<thead>
<tr>
<th>Block</th>
<th>Mean $P_{(max)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.35 (0.03)</td>
</tr>
<tr>
<td>2</td>
<td>0.31 (0.03)</td>
</tr>
<tr>
<td>3</td>
<td>0.33 (0.06)</td>
</tr>
<tr>
<td>4</td>
<td>0.30 (0.07)</td>
</tr>
<tr>
<td>5</td>
<td>0.29 (0.08)</td>
</tr>
<tr>
<td>6</td>
<td>0.32 (0.04)</td>
</tr>
<tr>
<td>7</td>
<td>0.35 (0.07)</td>
</tr>
<tr>
<td>8</td>
<td>0.35 (0.04)</td>
</tr>
<tr>
<td>9</td>
<td>0.30 (0.04)</td>
</tr>
<tr>
<td>10</td>
<td>0.33 (0.08)</td>
</tr>
</tbody>
</table>

Note: Figures in parentheses are standard errors.
over time, we increased the number of choices that participants made and introduced negative payoffs to facilitate this process.

EXPERIMENT 3

Experiment 1 demonstrated that participants could learn to maximize their payoffs when the magnitude of those payoffs varied. In Experiment 2, however, participants were unable to maximize the probability of payoff. In Experiment 3 we again used probability as the payoff dimension but introduced negative payoffs that penalized participants for meliorating behavior (see Exhibit 9). Negative payoffs influence decision making in a number of ways. First, Bereby-Meyer and Erev (1998) report that learning in a probability learning experiment is faster when participants receive negative payoffs for incorrect decisions than when they receive no payoff. Second, due to asymmetric value functions, losses are known to be more motivating than gains of the same value (Kahneman and Tversky, 1979). Although in Experiment 3 the payoffs associated with maximization remained similar to those used in our earlier experiments (approximately £1.32 per block), the payoff associated with melioration becomes a loss of approximately £0.66 per block (compared to a gain of £0.66 per block in Experiment 2). For instance, if participants allocated less than 50% of their responses to the Max button then the payoff for further Max responses became a loss of —£0.02. In addition we also gave participants slightly longer training periods of 700 trials on each session compared to 500 trials for Experiments 1 and 2.

Method

Participants

Twelve members of the University College London community volunteered for this study, five male and seven female. Participants were warned that they could lose as well as win money as a result of the choices that they made, but that they could win up to £9.00.
Apparatus

The screen displays were identical to the ones used in Experiment 2 (Exhibits A1 and A2).

Design and procedure

At the start of the experiment participants received the same instructions that were used in Experiment 2 with the exception that the phrase ‘As a result of your choices you can both win and lose money’ was added and highlighted in red. These instructions were also available on a sheet of paper throughout the experiment. Each of the two sessions consisted of 700 consecutive trials lasting approximately 20 minutes. As in Experiment 2 feedback was provided every 100 trials when participants were prompted to take a short break and informed of the cumulative payoffs and what their cumulative payoffs would have been had they adopted the optimal strategy. Payoff from the optimal strategy was determined by choosing Max on each choice. As Exhibit 9 shows, the probability of receiving a payoff when consistently choosing Max was 0.66, so the average optimal payoff after every 100 trials was £1.32, the same as in Experiment 3.

Payoff schedules

As in Experiment 2 participants did not receive a payoff for every choice that they made. The magnitude of payoff associated with each button was fixed at £0.02, but the probability of receiving that reward on any one trial was varied. In this experiment participants also received negative payoffs of −£0.02 when the proportion of responses allocated to the Max button fell below 0.5.

As in Experiment 2 if participants consistently chose the Max button over a period of ten consecutive trials, then the probability of receiving a payoff of £0.02 if they chose Max again would be 0.66. If, after
10 consecutive Max choices, participants decided to switch to the Mel button the probability of receiving £0.02 would be 0.99. The probability of receiving payoff for Max choice = ((1.32* proportion of time on Max) −0.66), and the probability of receiving payoff for Mel choice = ((1.32* proportion of time on Max) −0.33). If however, the probability of receiving a payoff on either button was negative then the payoff sign became negative. For example, if participants choose the Mel button 90% of the time then the probability of receiving a negative payoff would be 0.2, and if they switched to max, it would be 0.53. If participants consistently chose the Mel button over a period of ten consecutive trials the probability of losing a payoff of −£0.02 would be 0.33 if they chose Mel again. If, after the same period of 10 consecutive Mel choices, participants decided to switch to the Max button the probability of losing −£0.02 would be 0.66.

Over the 700 trials of each session consistently choosing Max would return an average cumulative payoff of £6.60, while consistently choosing Mel would return a cumulative loss of £4.62.

Results and discussion

Participants earned a mean of £3.10 (SD = 1.61) over Session 1 and £4.03 (SD = 2.59) over Session 2. None of the participants finished any session in debt; the minimum earned on any session was £1.84. The proportion of maximizing responses for each block is shown in Exhibit 10. A repeated-measures ANOVA with Block as the within-subjects factor revealed no reliable change in the allocation of responses across blocks (F(13, 143) = 1.63, MSE = 0.02, p < 0.08).

Exhibit 11 shows the distribution of maximizing responses during the last block of each session. In the final block two participants allocated more than 95% of responses to the Max button, and five allocated more than half. The remainder allocated the majority of their responses to the Mel button. Overall participants allocated no more responses to the Max button than would be expected by chance during the last block of each session (t < 1 for Session 1 and for Session 2 t(7) = 1.24, SD = 0.58, p < 0.24). The allocation of choices did not differ between the final block of each session (blocks 7 and 14): t(11) = −1.33, SD = 0.25, p < 0.211, indicating that behavior did not change over the two sessions.

Exhibit 8 shows the mean proportion of maximizing responses over each block for each experiment. A comparison of Experiments 2 and 3 reveals that participants in Experiment 3, who received both positive and negative payoffs, allocated more responses to the Max button than did those in Experiment 2 who received positive payoffs alone. In Experiment 3 we were unable to detect an effect of learning despite more trials than Experiment 2 and a manipulation (negative payoffs) that has been shown to increase motivation (Bereby-Meyer and Erev, 1998). Moreover, participants’ allocation of choices were indistinguishable from a chance distribution despite having the same sample size as Experiment 2 in which participants were closer to the melioration equilibrium than either chance or maximization. An inspection of Exhibit 11 indicates that an increase in sample size would be unlikely to influence these results. Perhaps negative payoffs discouraged participants from allocating consecutive choices to either alternative for long periods. If so, this would affect participants’ subsequent allocation of choices as participants would not have learned the payoff.

<table>
<thead>
<tr>
<th>Block</th>
<th>Mean P(max)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Session 1</td>
<td>0.49 (0.03)</td>
</tr>
<tr>
<td>Session 2</td>
<td>0.58 (0.07)</td>
</tr>
</tbody>
</table>

*Note:* Figures in parentheses are standard errors.
schedules. In Experiment 4 we introduce a further manipulation designed to encourage learning of the payoff schedules.

EXPERIMENT 4

Experiment 3 failed to find evidence for maximization despite negative payoffs penalizing melioration. One possibility is that negative payoffs discouraged participants from exploring how their behavior affected the payoff schedules. In Experiment 4 participants received one block of un-reinforced practice in order to allow them to learn something of the structure of the payoff schedules without being penalized for doing so. Silberberg, Thomas, and Berendzen (1991) report an experiment in which participants were asked to choose between alternatives that affected room temperature, cold being associated with melioration and warmth with maximization (on the assumption that participants would want to maximize warmth in a cold environment). Despite the motivating nature of these payoffs Silberberg et al. (1991) found participants only maximized when given prior training to discriminate between schedules.

Method

Participants
Twelve members of the University College London community volunteered for this study, five male and seven female. Participants were warned that they could lose as well as win money as a result of the choices that they made, but that that they could win up to £9.00.

Apparatus
The screen displays were identical to the ones used in Experiment 3 (Exhibits A1 and A2).
Design and procedure
The design and procedure were identical to Experiment 3 with the exception that participants were given an additional training block of 100 trials at the beginning of each session. The practice block was identical to the test blocks except that payoffs were in points rather than pence.

Payoff schedules
The schedules were identical to those used in Experiment 3 (see Exhibit 9).

Results and discussion
Participants earned a mean of £3.74 ($SD = 2.04$) over Session 1 and £5.46 ($SD = 2.53$) over Session 2. None of the participants finished any session in debt; the minimum earned on any session was £0.18. The proportion of maximizing responses for each block is shown in Exhibit 12. A repeated-measures ANOVA with Block as the within-subjects factor revealed a reliable change in the allocation of responses across blocks ($F(15, 165) = 3.31$, $MSE = 0.03$, $p < 0.01$) indicating that learning took place.

Exhibit 13 shows the distribution of maximizing responses during the last block of each session. In the final block five participants allocated more than 95% of responses to the Max button, and nine allocated more
than half. Overall participants made more responses to the Max button than would be expected by chance during the last block of each session ($t(11) = 2.37, SD = 0.23, p < 0.04$ for Session 1 and $t(11) = 3.89, SD = 0.25, p < 0.01$ for Session 2), indicating a robust tendency towards maximization. All but two participants allocated more responses to the maximizing alternative at the end of Session 2 than at the end of Session 1.

The mean proportion of maximizing responses for each block (except the practice blocks) are shown in Exhibit 8. Inspection reveals that participants in Experiment 4, following un-reinforced practice, were substantially closer to maximization than those in Experiments 3 where there was no practice.

GENERAL DISCUSSION

In this series of experiments we examined how people allocate choices between two alternatives that differ in payoff. On the one hand, melioration suggests that participants tend to choose the alternative that has the higher immediate payoff irrespective of the future consequences for overall payoff, while on the other, rational choice theory requires that participants allocate choices in order to maximize the overall payoff irrespective of which has the higher immediate return. In many situations the option that has the highest immediate payoff also leads to the highest overall returns, but in cases where there is an internality between choice and payoff, such that choosing one alternative reduces the payoff on one or both alternatives, the two theories make divergent predictions. Animal behavior in such situations is consistent with melioration as the motive for choice, in that although the value of each individual choice is maximized, the overall utility is suboptimal. Research with humans has produced mixed results. Some data are consistent with rational choice theory (e.g. Silberberg et al., 1991), while other data suggest melioration may be the motive for choice (e.g. Herrnstein et al., 1993; Savastano and Fantino, 1994).

However, no study has unequivocally demonstrated equilibrium (or asymptotic) behavior, in part because participants have received relatively little training, and because the payoff for adopting one strategy over another may have been too small to motivate participants. Our aims were to examine choice behavior in situations with different payoff structures. In the present research we gave participants long training sessions and substantial financial incentives in order to establish whether people could learn to allocate choices that maximize their expected payoffs, and if not under what conditions behavior might become optimal.

Exhibit 8 summarizes the results of all the experiments. In Experiment 1 participants received a payoff for each choice that they made. The financial magnitude of one alternative was always higher than the other, but consistently choosing the higher magnitude alternative reduced the magnitude of both options. Consistently choosing the lower magnitude alternative increased the magnitude of both alternatives, such that consistently choosing the lower magnitude alternative maximized the overall payoff from the experiment. Melioration suggests that participants will consistently choose the higher magnitude alternative disregarding the consequences for overall payoff. Rational Choice Theory requires participants to consistently choose the lower magnitude alternative in order to maximize overall payoff. Experiment 1 demonstrates that participants can allocate choices to maximize the magnitude of payoff but that they must learn to do so: after 1000 trials, participants allocated 80% of their responses to the optimal alternative. These data are consistent with Rational Choice Theory.

In contrast, in Experiment 2 we observed a different pattern of results. Although this experiment had a similar structure to Experiment 1, participants did not receive a payoff for every choice that they made. Rather the allocation of choices affected the probability of receiving a fixed magnitude payoff. In this case participants’ behavior was closer to melioration than maximization: at the end of Session 1 participants allocated an average 29% of the responses to the optimal alternative, and only 33% at the end of Session 2.

In Experiment 3 we examined whether negative payoffs might discourage participants from melioration. The overall payoff for maximization was the same as in the previous three experiments, but negative payoffs reduced the overall payoff for meliorating behavior to a loss. Although the distribution of participants’ responses was further from melioration than was observed in Experiment 2 (at the end of Session 1
participants allocated an average of 49% of the responses to the optimal alternative and 58% at the end of Session 2) this distribution did not differ from chance. One possibility is that negative payoffs discouraged participants from sufficiently exploring the payoff schedules that would enable them to allocate their choices strategically. However, the results of Experiment 3 were sufficiently encouraging to ask whether pre-training might encourage maximization.

Experiment 4 was identical to Experiment 3 with the exception that participants received an additional block (100 trials) of un-reinforced practice to familiarize themselves with the payoff schedules. At the end of Session 1 participants allocated an average of 66% of their responses to the max button, and at the end of Session 2 this exhibit had risen to 78%. This experiment demonstrates that the seemingly stable sub-optimal behavior observed in Experiments 2 and 3 can be avoided if participants are given an opportunity to learn how their choices might affect their payoffs without being penalized for doing so. We believe that this is the first demonstration of this kind.

Our sample sizes were relatively small, yet although individual differences are apparent, in three of the four experiments the samples were sufficient to reach statistical significance. Moreover, in the experiments in which maximization was the dominant strategy the majority of participants allocated more choices to the maximizing alternative in Session 2 than in Session 1. We conclude from this that any individual differences there may have been reflect differences in the rate at which participants learned the payoff schedules, rather than differences in strategy. In sum, when the payoffs from each alternative differed in magnitude the majority of participants soon reached the maximizing equilibrium, but when the payoff from each alternative was fixed and differed in probability participants appeared to meliorate except where they received unreinforced practice.

Why might magnitude and probability of payoff make a difference to choice behavior? A number of workers (e.g. Mazur, 1997; Prelec, 1982; Rachlin, 2000) suggest that probabilistic payoffs are discounted in the same way as delayed payoffs. For instance, it is well known that people tend to prefer £100 tomorrow to £100 next week. The two payoffs are discounted as a function of their relative delay periods. The payoffs used in Experiments 2 and 3 were fixed at £0.02 but may have been discounted as a function of the probability of receiving that payoff. Indeed, Herrnstein et al. (1993) report that participants are more likely to meliorate than maximize when payoffs vary in delay than when they vary in magnitude, and this comparison is similar to our observations and reflects behavior in one-shot decisions (e.g. Kahneman and Tversky, 1979; Thaler, 1981). In our experiments, however, whatever participants’ discount rates might be the melioration equilibrium always has a substantially lower value than the maximizing equilibrium. The probability of payoff at the maximizing equilibrium was 0.66, while at the meliorating equilibrium was 0.33, and since the discount rate remains constant across both schedules the value of the maximizing equilibrium remains higher than that of the meliorating equilibrium.

Herrnstein et al. (1993) provide a rather more prosaic interpretation of the difference between magnitude of reinforcement and delay until reinforcement that can, in principle, be extended to probability. In their Experiment 3 Herrnstein et al. compared a fixed magnitude payoff with varying delays versus a fixed delay with varying magnitude payoffs (for our purposes we are concerned with the conditions where the payoff slopes were parallel). Participants in the varying magnitude condition allocated on average 70% of their responses to the optimal alternative, while those in the varying delay condition allocated a mere 22% of their responses to the optimal alternative. Herrnstein et al. suggested that allocating choices in the magnitude condition was relatively easier than in the delay condition because, although participants must still perform the same calculations to determine the future consequences of each choice, the first step is provided by the apparatus. That is, the immediate consequences of behavior are readily apparent to the subject on screen in numerical form. In the case of delay, and indeed of probability, no such information is present and must instead be inferred by the participant. Nonetheless, over long periods of training we might still expect participants to learn the probability of reinforcement and how it was contingent upon their behavior. Yet Experiments 2 and 3 provide little evidence of this: it is hard to detect any evidence of learning in these experiments (see Exhibit 8). Thus it is not obvious that Herrnstein et al’s. explanation is appropriate.
We trained participants in well-structured environments for long periods in the hope that, at equilibrium, participants could maximize their payoffs. Participants learned to maximize their overall payoffs when choice affected the magnitude of payoff, but tended to meliorate when behavior affected the probability of payoff. Thus choice behavior can differ considerably in similar environments. However, we also found that un-reinforced practice that encouraged participants to examine the problem without being penalized for doing so induced maximization. This is an important result that goes beyond other laboratory findings that typically seek to observe rather than change behavior (e.g. Herrnstein et al., 1993). We conclude that choice in the Harvard Game is, like probability matching (Shanks et al., 2002), the Three Doors problem (Friedman, 1998), and base-rate neglect (Goodie and Fantino, 1999), an example of a choice anomaly that is heavily context dependent and which can be made to disappear under appropriate conditions.

Exhibit A1. Screenshot of Experiments 1–4. For half the participants Max was the left button and Mel the right, for the other half the button assignments were reversed.

Exhibit A2. Screenshot of the blocked feedback used in Experiments 1–4. Participants were informed of their cumulative payoff over the preceding 100 trials relative to what that payoff would have been if they had adopted an optimal strategy. Participants were not informed that the optimal strategy was to consistently choose the Max button.
ACKNOWLEDGEMENTS

The research described here was supported by grants from the United Kingdom Economic and Social Research Council (ESRC) and the Leverhulme Trust. The work is part of the programme of the ESRC Centre for Economic Learning and Social Evolution, University College London. We thank Ken Binmore, Nigel Harvey, Darrell Collins, Ben Newell, and Magda Osman for their helpful comments.

REFERENCES

Authors’ biographies:

Richard Tunney (DPhil 1999, University of York) is Senior Research Fellow at the ESRC Centre for Economic Learning and Social Evolution. His interests include human learning, judgment and decision making, and economic psychology.

David Shanks (PhD 1985, University of Cambridge) is Professor of Experimental Psychology at University College London and Scientific Director of the ESRC Centre for Economic Learning and Social Evolution. His research interests include human learning, the implicit–explicit distinction, computational modeling, especially with neural network models, economic psychology, and game theory.

Authors’ address:

Richard Tunney and David Shanks, Department of Psychology, University College London, Gower Street, London WC1E 6BT, UK.