ELSEVIER

Journal of Economic Behavior & Organization
Vol. 39 (1999) 111–128

JOURNAL OF
Economic Behavior
& Organization

The effect of adding a constant to all payoffs: experimental investigation, and implications for reinforcement learning models

Ido Erev [a,*], Yoella Bereby-Meyer[1, b], Alvin E. Roth[2, c]

[a] *Faculty of Industrial Engineering and Management, Technion, Haifa, Israel*
[b] *Department of Education, Ben-Gurion University of the Negev, 84105 Beer Sheva, Israel*
[c] *Department of Economics, Harvard University, Cambridge, MA 02138, USA*

Received 16 March 1998; received in revised form 12 January 1999; accepted 16 March 1999

Abstract

This paper examines the effect on learning in simple decision tasks of the addition of a constant to all payoffs. Experiment 1 reveals that this effect, initially observed in a probability learning task, is not limited to single person decision making under uncertainty. Experiment 2 shows that the effect is not linear. Two additional experiments show that the non-linearity cannot be explained by whether zero is in the payoff range. The implications of these results for reinforcement learning models are evaluated and two models that capture the main results are proposed.

JEL classification: C72; C92

Keywords: Reinforcement; Adaptive learning; Reference point; Stimulus response; Probability matching

1. Introduction

In an earlier study two of us (Bereby-Meyer and Erev, 1998) found that learning speed, in a binary choice task, is affected by the sign of the possible payoffs . Maximal learning speed was observed when the optimal strategy had a positive expected value, and the alternative

* Corresponding author. E-mail: erev@tx.technion.ac.il
[1] E-mail: yoella@bgumail.bgu.ac.il
[2] Also affiliated at: Harvard Business School, Boston, MA 02163. E-mail: aroth@hbs.edu, http://www.economics.harvard.edu/~aroth/alroth.html

0167-2681/99/$ – see front matter
PII: S0167-2681(99)00028-1

strategy had a negative expected value. Thus, the addition of a constant to all payoffs that changes the signs of the expected payoffs can affect learning speed. This finding is potentially important because it may reflect a robust behavioral regularity that should be captured by descriptive models of learning.

The current research was designed to improve our understanding of the added constant effect and its implications to the development of descriptive learning models. The paper is organized as follows: Section 2 summarizes the experiment reported in Bereby-Meyer and Erev (BME, 1998), which found that the addition of a constant to all payoffs can have a significant effect on learning speed in a simple single-person decision task. Section 3 presents a new experiment that shows that the 'added constant' effect generalizes to strategic environments, and can be observed in a zero sum game.

Section 4 presents the adjustable reference point model supported by the BME results. It shows that although the model is consistent with the findings summarized in Sections 2 and 3, it cannot be general. The limitations of the model are highlighted in a thought experiment, and a new laboratory experiment. While the model predicts that the size of the effect will increase with the magnitude of the added constant, the results reveal that the effect is not monotonic. It seems that addition of a small constant can have large effect when all the payoffs are close to zero (the status quo outcome), and that a further increase in the added constant has little effect as it moves all the payoffs away from zero.

Section 5 presents two additional experiments to compare alternative modifications of the adjustable reference point model. Our analysis suggests that the effect is relatively insensitive to the range of the possible payoffs. Section 6 summarizes the observed regularities, and compares variants of the adjustable reference point model and an alternative average reinforcement model that can describe them. To reduce the risk associated with post hoc models, the models' parameters are fitted based on the (9) one-person decision tasks and then evaluated based on 13 two-person games.

2. Bereby-Meyer and Erev's (BME, 1998) main results

BME studied binary decision tasks under uncertainty. In each of the tasks (experimental conditions) the decision maker (DM) participated in 500 independent trials and was asked to guess which of two mutually exclusive events L or H will occur. In all trials the probability of H (P_h) was 0.7 (and the probability of L was 0.3). After each trial the DM received an immediate feedback concerning the realized event and his/her payoff.

The different tasks differed with respect to the payoffs. In Condition 0,4 the DM earned 4 points when he or she guessed correctly and lost nothing when he or she was wrong. The other two conditions were created by subtracting a constant from these payoffs. In Condition -2,2 the payoffs were 2 for a correct response and -2 for incorrect, and in Condition -4,0 the DM lost 4 points when he or she was wrong and earned nothing for a correct response. Subjects received an initial show up fee (in points), and at the conclusion of the experiment the accumulated points were converted to money (0.01 Shekel = \$0.003 to each point).

Simple decision problems of this type, referred to as probability learning tasks, were studied extensively by psychologists in the 1950s and 1960s. While the optimal response is always to choose the most common event (guess 'H'), the literature reveals that DMs are

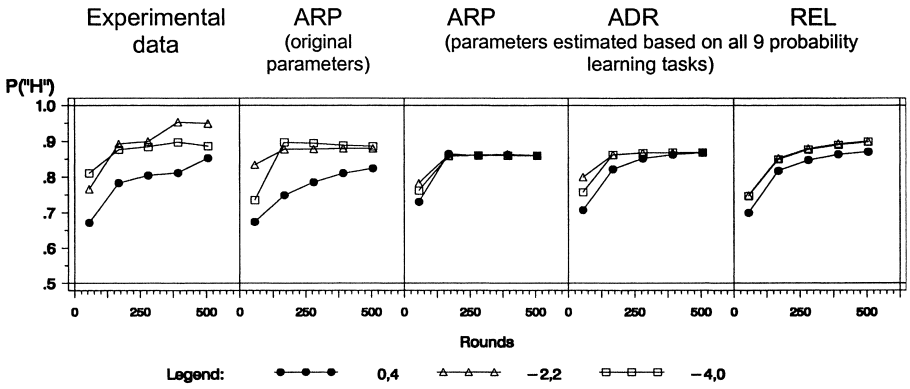

Fig. 1. Bereby-Meyer and Erev (1998). Proportion of optimal ('H') choices as a function of time (5 blocks) and payoff condition in the experiment (left) and the simulations.

slow learners. After 100 and 200 trials they tend to 'probability match;' that is, to select 'H' in 70 percent of the trials. With longer experience DMs slowly move toward the optimal choice (see Edwards, 1961). In addition, manipulation of the payoffs can affect the learning process (Siegel et al., 1964).

BME's experimental results are summarized in the left hand column of Fig. 1 by the proportion of 'H' choices ('optimal' choices) in five blocks of 100 trials in each condition (the right hand columns are simulation results to be discussed below). A two-way repeated measures ANOVA (with block as the repeated measure) on the choice of the dominant color (P('H')) revealed a significant condition effect. The proportion of 'H' choices was significantly lower in Condition 0,4 than in the other two conditions. The difference between Conditions $-2,2$ and $-4,0$ was insignificant.

3. The added constant effect in a constant sum game

The results summarized above appear to contradict previous findings by Rapoport and Boebel (1992). They found that the addition of a constant to all payoffs in a 5×5 constant sum two-person matrix game did not have a significant effect on behavior. This difference can be a result of a qualitative difference between single person decision making and decision making in the strategic environment presented by a game, but can also be a result of other differences between the BME and Rapoport and Boebel studies. Rapoport and Boebel's task was more complex, led to relative flat learning curves and was played for only 120 periods. Moreover, they did observe some (although insignificant) effect of the addition of a constant to the payoffs. Thus, it is possible that the added constant effect can also be found in games.

To evaluate this hypothesis the current section studies the added constant effect in a game played under the conditions used by BME with the exception that the payoffs are determined by a game (and players know that). Following Suppes and Atkinson (1960) we examined a 2×2 probabilistic constant sum game. In each trial of this game each player can either win or lose, and the winning probability is determined by the payoff matrix and the choices

made by the two players. Thus, as in a probability learning task each player makes a binary choice and then receives one of two payoffs.

3.1. Method

Participants. Fifty-six Ben Gurion University students served as paid participants in the experiment. They were assigned to one of two experimental conditions and run in pairs. The exact payoffs were contingent on performance and ranged from 21 to 24 Shekels ($7–$8).

Apparatus and procedure. The experiment was programmed and run using *Visual Basic* 3 for Windows 3.1. This system was installed on a 486PC, with a Super VGA 14" screen. Both pair members were seated in front of the same computer. They were separated by a plastic divider so each could see only his own part of the screen. They received 20–22.5 shekels for showing up and were told that they would play a game in which they could earn more money, but can also lose some of the show up fee.

Subjects were informed that they were playing a game against the person seated next to them. In each of the game's 500 trials they had to select one of two keys, and could either win (receive a payoff of W) or lose (a payoff of L). The payoffs were determined by the matrix in Fig. 2 which was not known to the subjects. They were only told that their payoff depends on their choice, their opponent's choice and on a chance event.

Each of the two keys was associated with one of the game alternatives (A or B) and a color (Blue or Red) that appeared on a box on the screen following the choice. The selected cell in the game's matrix determined the winning probabilities. For example if Player 1 chose Blue (A1) and Player 2 chose Red (B2), the probability of winning was 0.8 for player 1, and 0.2 for Player 2.

The subjects received two types of feedback after each period: an update in the accumulating payoff counter, and a graphical feedback. The graphical feedback was presented (in a feedback box for 3 seconds) to match the BME display. This display did not add information: winning was represented by presentation of the color the player had chosen. A loss was represented by the color not chosen.

Two payoff conditions were compared. In Condition −.5,.5 the payoffs were $L = -0.5$ and $W = 0.5$ and the initial endowment (showup fee) was 2250 points (The value of each point was 0.01 Shekels ($0.003)). In Condition 0,1 the payoffs were $L = 0$, $W = 1$ and the initial endowment was 2000 points (20 shekels).

3.2. Results

The left hand column in Fig. 2 presents the proportion of A choices in five blocks of 100 trials over the 14 subjects in each role in each condition. The results in Condition 0,1 are practically identical to the results observed in previous studies of this condition (Suppes and Atkinson, 1960; Erev and Roth, 1998). In Condition −.5,.5 Player 2 appears to be closer to the equilibrium, but Player 1 initially moves farther away from the equilibrium.

To evaluate if this pattern is consistent with the hypothesis of faster learning in Condition −.5,.5 a payoff sensitivity score was calculated for each participant as the proportion of

	A2	B2
A1	(.3, .7)	(.8, .2)
B1	(.4, .6)	(.1, .9)

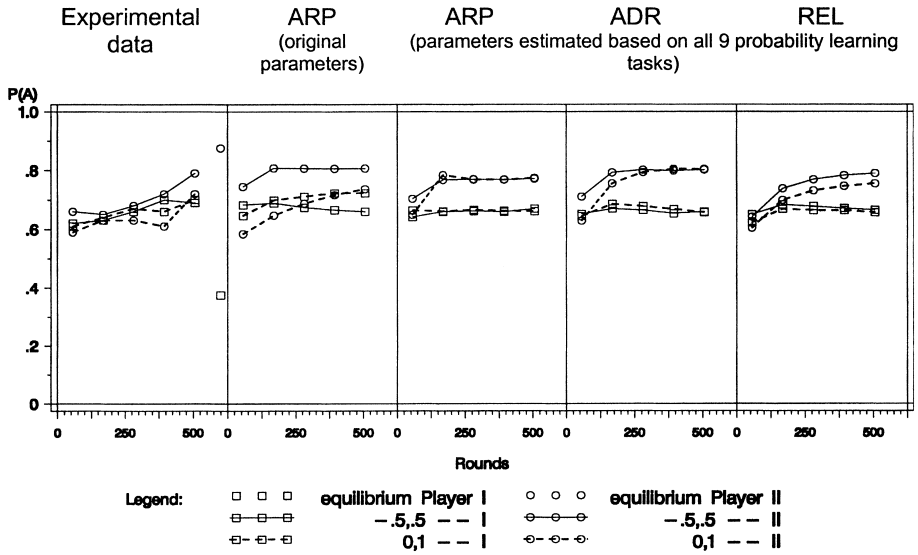

Fig. 2. The game experiment. The payoff matrix (top) presents the probability of winning for each subject given the possible choices. The figure shows the proportion of A choices as a function of time (5 blocks), payoff condition and player role in the experiment (left) and the simulations. The equilibrium predictions are presented at the right hand side of the experimental column.

choice of the alternative that yielded higher average payoff in the previous rounds of the game. Over the 500 blocks the score was 0.632 in Condition 0,1 and .679 in Condition .−.5,.5. The difference is small but significant ($t[26] = 2.0$, $p < 0.05$). These results are consistent with the BME findings. Players appear to be more sensitive to past outcomes when the payoff framing involves losses.

4. The model supported by BME and its limitations

BME showed that the added constant effect, described above, is not captured by most learning models proposed in recent years (e.g. Roth and Erev, 1995; Tang, 1996; March, 1996; Mookherjee and Sopher, 1997; Borgers and Sarin, 1995; Fudenberg and Levine, 1998; Cheung and Friedman, 1998; Camerer and Ho, 1998). Only the adjustable reference point (ARP) reinforcement model (initially proposed in Erev and Roth, 1996 and

studied in Rapoport et al., 1997; Erev and Rapoport, 1998) captures the observed effect. [3] This model is an extension of Roth and Erev's (1995) linear quantification of Thorndike's (1898) Law of Effect (similar quantifications were suggested by Bush and Mosteller, 1955; Luce, 1959; Herrnstein, 1970; Harley, 1981). It can be described by the following assumptions.

A1 Initial propensities: At time $t=1$ (before any experience has been acquired) each player n has an initial propensity to play his kth pure strategy, given by some non-negative number $q_{nk}(1)$. In the current context it is natural to assume equal initial propensities for all pure strategies, that is for each player n,

$$q_{nk}(1) = q_{nj}(1) \text{for all pure strategies } kj.$$

A2 Reinforcement function: The reinforcement of receiving a payoff x in trial t is given by an increasing function $R(t, x)$. Specifically, the reinforcement is assumed to be the difference between the payoff and the player's reference point.

$$R(t,\ x) = x - \rho_n(t),$$

where the reference point is a weighted average of previous payoffs. The weighted average is computed as

$$\rho_n(t+1) = (W(t))(\rho_n(t)) + (1 - W(t))x \tag{1}$$

where $W(t)$ is determined by the sign of the reinforcement and two parameters ($0 < w$, $\alpha < 1$): $W(t) = w$ if $x \leq \rho_n(t)$ and $W(t) = 1 - \alpha(1 - w)$ if $x > \rho_n(t)$.

A3 Updating of propensities: If player n plays his kth pure strategy at time t and receives a reinforcement of $R(t,x)$, then the propensity to play strategy j is updated as a function of $R(t,x)$.

$$q_{nj}(t+1) = \text{MAX}[\upsilon, (D(t))q_{nj}(t) + E_k(j, R(t, x))] \tag{2}$$

where $\upsilon > 0$ is a technical parameter that insures that all propensities are positive, $D(t)$ is a discounting function which slowly reduces the importance of past experience, and E is a function which determines how the experience of playing strategy k and receiving the reward $R(t,x)$ is generalized to update each strategy j.

The model assumes a fixed discounting $D(t) = 1 - \phi$ where ϕ is a forgetting parameter. In the case of binary decisions the generalization function is reduced to a 'two-step' function:

$$E_k(j, R(t, x)) = \begin{cases} R(t, x)(1 - \varepsilon) & \text{if } j = k \\ R(t, x)\varepsilon & \text{otherwise} \end{cases}$$

A4 Probabilistic choice rule: Following Luce (1959) the probability $p_{nk}(t)$ that player n

[3] The other models predict no effect (Roth and Erev, 1995; Cheung and Friedman, 1996; Mookherjee and Sopher, 1997; Fudenberg and Levine, 1997) or a different pattern (Tang, 1996; March, 1996; Borgers and Sarin, 1995; Camerer and Ho, 1998).

plays his *k*th pure strategy at time *t* is

$$p_{nk}(t) = \frac{q_{nk}(t)}{\Sigma q_{nj}(t)}, \tag{3}$$

where the sum is over all of player *n*'s pure strategies *j*.

Predictions. BME derived the model's predictions using computer simulations given the parameters selected by Erev and Roth (1996, and utilized by Rapoport et al., 1997; Erev and Rapoport, 1998). The value of these parameters are: $s(1) = 3$, $\epsilon = 0.2$, $\phi = 0.001$, $\rho(1) = 0$, $w = 0.98$, and $\alpha = 0.5$. The model's predictions for the three conditions compared by BME are presented in the second column in Fig. 1. This figure shows that the model captures the slower learning in the gain domain (Condition 0,4).

The second column in Fig. 2 presents the model's predictions for the experiment described in Section 3. In this case too the model appears to capture the added constant effect.

Limitations. Although Erev and Roth's (1996) adjustable reference point model provides a reasonable fit to the results presented above it is clear that this model is limited. To see this, think about the effect of adding a very large constant to all payoffs in the BME study. For example, consider the addition of 100 points to the 0,4 condition to create a 100,104 condition. The model predicts extremely slow learning in this condition. Even after 500 periods the model predicts choice probabilities around random choice. Moreover, slow learning is expected even in the trivial case of decision making under certainty ($P_h = 1$).

This unlikely prediction is a result of the assumed slow reference point adjustment process. Under the current model (and parameters) more than 100 periods are needed for the reference point to adjust to a value above 100. During this slow adjustment period both alternatives receive relatively high reinforcements. Thus, almost half the reinforcement is received for a choice of the dominated alternative. The ratio of reinforcement changes once the reference point is high enough, but at this stage all the reinforcements are small relative to the initial reinforcements and as a result the learning is very slow.

To provide empirical support for this thought experiment we ran a replication of the BME study with more extreme conditions (although not as extreme as the thought experiment). Two conditions were compared: payoffs 2,6 and -6, -2.

4.1. Experimental test

Method:

Participants: Twenty-eight Technion students served as paid participants in the experiment. They were randomly assigned to the two experimental conditions.

Apparatus and Procedure: The apparatus and procedure used in the experiment described in Section 3 were utilized again with two exceptions: (1) The payoff probabilities for each choice were fixed throughout the experiment (were not affected by the choice made by another subject), and (2) the subjects were told that their task is to guess which of the two events will occur.

Specifically, in each of the 500 trials the subjects were asked to predict the appearance of one of two colors (Blue or Red). The participants were told that their payoff would be *W* for a correct prediction and *L* for an incorrect prediction. The value of *W* and *L* defined the

experimental conditions. In Condition −6,−2 the payoff for inaccurate prediction (L) was −6, the payoff for accurate prediction (W) was −2, and the show up fee was 4000 initial points.

Condition 2,6 was created by adding 8 points to each outcome. Thus, the payoffs were $L = 2$, $W = 6$. To insure identical objective incentives the 'added points' ($8 \times 500 = 4000$) were 'deducted' from the initial endowment; so that there was no showup fee in that condition.

For each participant one of the two colors (Red or Blue) was selected to be the 'high probability' ('H') response. This color was the correct response in 70 percent of the 500 trials. The order of the high probability events was randomized independently for each participant across the 500 trials.

Results. The experimental results are summarized in the left hand column of the top panel in Fig. 3, which has the same format as Fig. 1 (proportion of 'H' choices ('optimal' choices) in five blocks of 100 trials in each condition). The predictions of the model presented above are graphed in the second column. The results show that in violation of the large condition effect predicted by the model, very little effect was observed in the experiment. In fact, the difference between the two conditions is not close to being significant ($F(1,26) = 0.44$, ns). Moreover, The proportion of 'H' choices in both conditions is not significantly different from the proportion in Condition 0,4 of the BME study ($F(1,39) = 0.32$, ns), but is significantly lower than the proportions observed in Conditions −2,2 and −4,0 ($F(1,52) = 9.86$, $p < 0.002$).

One possible explanation of this pattern is that the initial reference point is influenced by the range of possible payoffs (so that the model with a fixed initial reference point predicts different behavior than was observed). This hypothesis is examined in the next section.

5. The effect of irrelevant outcomes

The adjustable reference point model presented above can account for all the results summarized above under the assumption that, for example the initial reference point parameter equals zero ($\rho(1) = 0$) when zero is in the payoff range, and equals the worst possible payoff ($\rho(1) = X_{\min}$) otherwise. Two experiments designed to test this kind of explanation by manipulating the payoff range via the addition of irrelevant outcomes are presented below.

5.1. A White trials experiment

In a first test of the 'payoff range' hypothesis we ran a replication of the experiment reported in Section 4 with the addition of 50 trials in which the state of nature was White and the payoff was zero independently of the subject's choice.

Method: Eighteen Technion students participated in this study. Nine subjects were assigned to Condition 0,2,6W which was identical to Condition 2,6 in Section 4 with the exception of the added 50 trials. Thus, the experiment lasted 550 trials. The added 50 trails were randomly distributed among the original 500 trials. The subjects were told that in some of the trials there will be no correct answer and the payoff will be zero. The color in the State of Nature feedback for these trials was White.

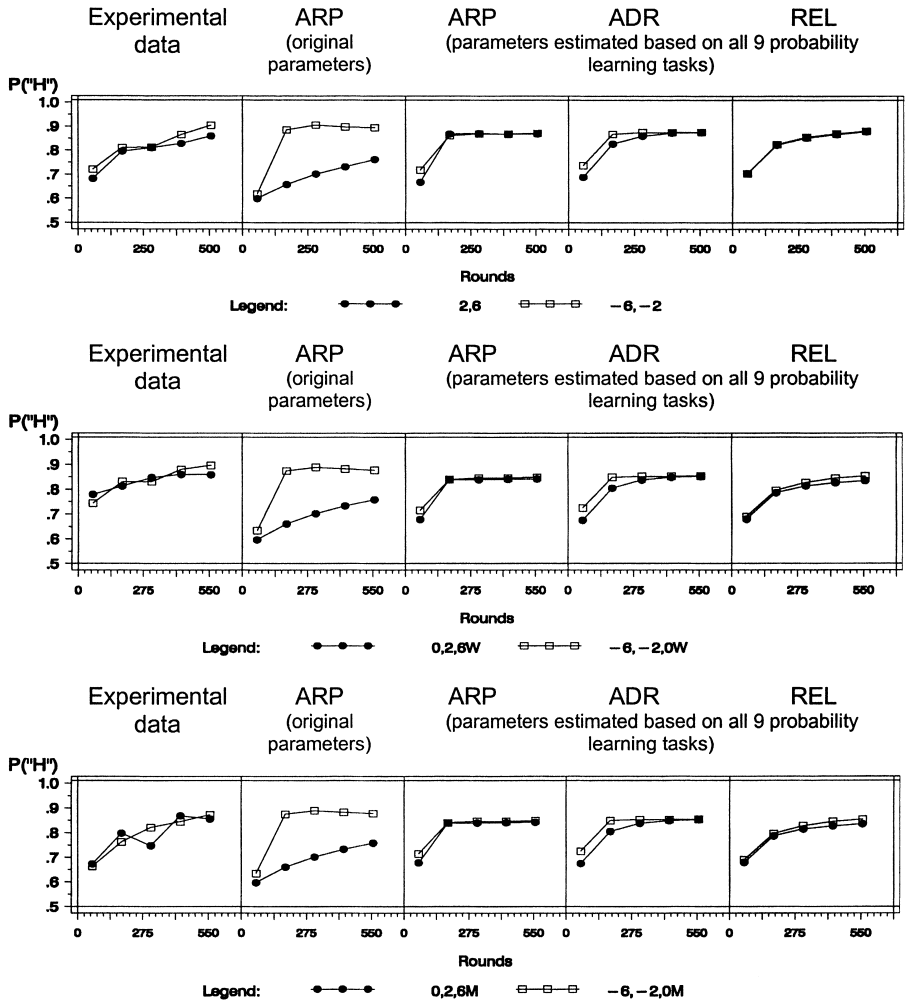

Fig. 3. The 2 and 6 conditions in Fig. 1's format.

A second group of nine subjects was assigned to Condition −6,−2,0W which was a replication of Condition −6,−2 with the addition of the white trials described above.

Results: The experimental results summarized in the left hand column of the second panel in Fig. 3 show no support for the 'payoff range' hypothesis. While this hypothesis implies the large condition effect predicted by the model presented in the second column, no condition effect was observed ($F(1,16) = 0.14$, ns).

5.2. A minimal and noisy information study

An additional experiment was run to evaluate if the failure of the 'payoff range' hypothesis can be explained by the subjects' ability to ignore the White trials. This study was a direct

replication of the previous study with respect to the payoff manipulation, but the instructions and the information were modified. In particular, the subjects (nine in each conditions) were not told that their task was to predict events. Rather, they were simply asked to choose between two keys. The feedback was limited to the obtained payoffs. Thus, a priori the subjects could not tell that the '0' payoff trials are irrelevant. That is, this experimental condition is designed to limit the information of the players to that used by the reinforcement learning model. If the results differ from the previous results, the difference can be attributed to an effect of the additional information the players had when they knew that some trials would give them zero no matter what they chose.

We refer to these two minimal information conditions as 0,2,6M and −6,−2,0M.

Results: The experimental results (left hand column of the lower panel in Fig. 3) are surprisingly similar to the results of the previous experiment. In violation of the 'payoff range' hypothesis and the model's predictions there was no significant condition effect ($F(1,16) = 0.01$, ns). Thus subjects seem able to ignore the outcomes that they do not influence in this case also, even though they do not know that some of their outcomes are not influenced by their actions.

6. Implications and alternative models

The results presented in Sections 4 and 5 suggest that a minimal variation of the ARP model in which the initial reference point is sensitive to the payoff range is not sufficient to account for the non-linear effect of the addition of a constant to the payoff. The current section extends the search for a descriptive model that can account for the current results. Two directions are taken: Minimal modifications of the adjustable reference point (Sections 6.1 and 6.2), and a more dramatic modification that assumes that DMs consider a loss aversion strategy (Section 6.3). We understand that the current search is not likely to result in finding the 'right model,' yet we hope to find potentially general post hoc models. Specifically, they should account for behavior in the tasks for which the original model performed well (the 12 games considered in Erev and Roth (1998) and the three conditions studied in BME), as well as account for the current data. To facilitate ex ante predictive power this generality should be achieved without fitting parameters to specific games. Thus, the models should account for the following observations:

Observation 1: When the average payoffs are 'close' to 0, the original model (that assumes the initial reference point is 0 and the adjustment process is slow) provides a good approximation of behavior. In line with the model's prediction the addition of a constant to the payoffs affects learning speed. Learning is slower when all payoffs are non-negative.

Observation 2: When the average payoffs are 'farther away' from 0, the original model fails. In violation of the model's predictions the addition of a constant to the payoffs does not appear to affect learning speed.

Observation 3: The failure cannot be accounted for by a 'payoff range' hypothesis. It seems that the distinction between payoffs that are 'close to 0' and 'farther away from 0,' will have to be based on a quantitative measure.

6.1. The ARP model with estimated parameters

The failure of the ARP model with the parameters estimated in Erev and Roth (1996) does not imply that this model cannot account for the data with other values of its parameters. The model's parameters have to be re-estimated to evaluate this possibility. To facilitate evaluation of the estimated model's potential generality, we decided to estimate the parameters based on the nine probability learning tasks summarized above and then test its ability to capture the 12 games considered in Erev and Roth (1998)[4] and the new game studied here.

Although the model has seven parameters, the current analysis focuses on five. The value of two parameters, the technical parameter ν and the initial reference point parameter $\rho(1)$ are taken from Erev and Roth (1996) and are not estimated here. The value of the technical parameter is left at $\upsilon = 0.0001$. This value was selected by Erev and Roth to be 'small enough' so that further reduction does not affect the model's fit. Thus, it is not a free parameter that should be estimated. The value of the initial reference point is left at $\rho(1) = 0$ as this choice of parameter can be seen as a meaningful psychological assumption rather than an arbitrary choice. (Indeed an initial reference point at 0 is also assumed by other models including Prospect Theory (Kahneman and Tversky, 1979).)

A grid search with a mean squared deviation (MSD) criterion was conducted to estimate the value of the remaining five parameters. At the first step of this analysis computer simulations were run for a wide set of parameter combinations.[5] One hundred simulations were run with each set of parameter values in each of the nine tasks. In each task the simulated players 'participated' in the same number of rounds as the experimental subjects. At each round of each simulation the following steps were taken:

1. The simulated players' strategies were randomly determined via Eq. (3).
2. Payoffs were determined using the payoff rule employed in the experiment in question.
3. Propensities were updated according to Eq. (2).
4. The reference point for the next period was determined.

The results of the simulations were summarized by the same statistics that were used to summarize the experimental data. That is, the simulations of the probability learning tasks by the proportion of 'H' choices in blocks of 100 trials.

At a second stage of the analysis, the squared distance between the experimental and each set of simulated choice proportions in each block was calculated. The average of these squared distances is the MSD score of the relevant set of parameters. The set of parameters that minimize this score was found to be $s(1) = 3$, $\epsilon = 0.18$, $\phi = 0.075$, $w = 0.96$ and $\alpha = 0.5$. The MSD score (multiplied by 100) of the model with these parameters is 0.19.

The third panel in Figs. 1–3 shows the predictions of the current tasks. It shows that although the quantitative fit scores are relatively good (this statement will become clearer in the model comparison section below), the model does not capture Observation 1; with

[4] These games, described in detail in Erev and Roth, include all the published (before 1998) experimental games with unique mixed strategy equilibria that were run for more than 100 trials under conditions that eliminate the possibility of reciprocation.

[5] This parameter estimation approach is less efficient but more robust than traditional approaches to violations of the assumption of a well specified model (see a discussion in Roth et al., 1998).

the parameters that account for Observation 2 (the similarity of the 2,6 and the −6,−2 conditions) this model does not capture the higher proportion of optimal choices in Condition −2,2 relative to 0,4 (with the exception of the first block).

6.2. Variants of the ARP model with sensitivity to relative reinforcement size

The failure of the ARP model with estimated parameters suggests that at least one of the model's basic assumptions has to be modified to account for Observations 1 and 2 with a single set of parameters. In a search for a sufficient modification we estimated the parameters of the models that best fit the six new probability learning tasks (all the 2 and 6 conditions) and compared the estimated parameters to the parameters that best fit the BME results. This comparison shows that larger forgetting (ϕ) and adjustment (w) parameters are needed to capture the 2 and 6 conditions. The ARP model requires small forgetting and adjustment parameters to capture Observation 1 (BME results) but requires larger values of these parameters to capture Observation 2 (the 2 and 6 conditions). These results suggest that a modified model, whose discounting and/or adjustment speed are a function of factors that distinguish BME and the 2 and 6 conditions, may be able to account for the two observations with a single set of parameters.

The finding that the minimal information experiment gave similar results to the white trials experiment suggests that the distinguishing factors should be computed from the obtained payoffs (that were the only feedback available in the minimal information experiment). The most obvious difference (excluding the payoff range) is the absolute average size of the reinforcements. While the initial average absolute reinforcements were around 2 in the 0,4 and −4,0 conditions, they were around 4 in the 2,6 and −6,2 conditions.

In what follows, we construct and test a reinforcement model with the property that, when the most recent reinforcements received are far from the running average of reinforcements, behavior adjusts quickly, but adjusts slowly when the recent reinforcements are in the range of the running average.

To distinguish between 'small' and 'large' absolute average reinforcement size, it is convenient to focus on relative size. (Recall that reinforcement size is different from payoff size, since reinforcements are measured compared to the current value of the reference point.) Relative reinforcement size at time t is defined here as

$$\Delta(t) = \frac{\mathrm{AR}(t)}{\mathrm{RV}(t)}$$

where AR(t) is an estimate of the average reinforcement size and RV(t) is an estimate of the variability of the reinforcements. The average reinforcement size is estimated as AR(1) = R(1,x), and for $t > 1$:

$$\mathrm{AR}(t) = (w')\mathrm{AR}(t-1) + (1-w')R(t,x).$$

The reinforcement variability is estimated here by a weighted average of the change in accumulated reinforcements. Specifically, we assume that RV(1) = |R(1,x)|, and for $t > 1$:

$$\mathrm{RV}(t) = (w'')\mathrm{RV}(t-1) + (1-w'')|\mathrm{AR}(t-1) - R(t-1,x)|$$

Three variants of the ARP model that assume sensitivity to relative reinforcement size ($\Delta(t)$) are studied here. The first assumes that only the discounting is sensitive to $\Delta(t)$, the second assumes that only the reference point adjustment speed is affected, and the third assumes that both discounting and adjustment speed are affected.

An accelerated discounting (AD) model. To allow faster discounting given relatively large absolute average reinforcements, this modification of the ARP model replaces the discounting function in Eq. (2) with:

$$D(t) = (1 - \phi)^{|\Delta(t)|}$$

Note that since $0 < (1 - \phi) < 1$, large discounting is implied when $|\Delta(t)| > 1$ and small discounting is implied when $|\Delta(t)| < 1$. So the model will 'forget' previous propensities quickly when recent reinforcements are far from the running average.

The modified model, as described above has the five free parameters of the original model ($s(1)$, ϵ, ϕ, w and α) and two new averaging parameters w' and w''. To facilitate model comparison we start the current investigation with the simplification assumption $w' = w'' = w$ that reduces the number of free parameters to five.

An accelerated reference (AR) point model. The assumption of an 'accelerated reference point' can be added to the ARP model by a modification of the definition of $w(t)$ (see Eq. (1) in Section 4). Specifically, we assume:

$$W(t) = w^{|\Delta(t)-\alpha|}$$

where $0 < w < 1$ is an initial adjustment parameter, $0 < \alpha < 1$ is a 'pessimism' parameter that determines the relative reinforcement size that minimize the adjustment speed. This assumption implies fast adjustment (low $W(t)$) when $|\Delta(t) - \alpha| > 1$ and slow adjustment when $|\Delta(t) - \alpha| < 1$. Here it is the reference point (rather than the propensities to choose each strategy) that is being adjusted more quickly.) And, in line with the original model faster adjustment is predicted for negative reinforcements.

With the simplification assumption $w' = w'' = w$ the current model has five free parameters ($s(1)$, ϵ, ϕ, w and α) like the models discussed before.

An accelerated discounting (AD) and reference (ADR) model. The third modification of the ADR model includes the accelerated forgetting (as in the AD model) and accelerated reference point (as in the AR model). So in this model, when recent reinforcements are far from the running average, both the propensities to choose each action and the reference point against which payoffs are measured to assess how reinforcing they are adjust quickly. Note that in this case too, with the assumption $w' = w'' = w$, the model has five free parameters: $s(1)$, ϵ, ϕ, w and α.

6.3. A reinforcement average model with a 'loss aversion' strategy (REL)

In this section we briefly explore two distinct kinds of modifications. On the one hand, we will review a modification of the learning model itself, which makes propensities depend on average reinforcements rather than accumulated reinforcements. Such a model performed well on the games studied by Roth et al. (1998). However we will also consider how the effect of adding a constant can be modeled, not in the learning model itself, but in the

choice of strategies over which the learning model operates. We do that in the present case by positing a 'loss aversion' strategy, in the spirit of Prospect Theory (Kahneman and Tversky, 1979).

The basic reinforcement average (REA) model, studied by Roth et al. (1998) (see Mookherjee and Sopher, 1997; Camerer and Ho, 1998; Fudenberg and Levine, 1998, for similar models) can be described by Assumption A1 as stated above, a simplification of A2 that assumes that the reinforcements are identical to the obtained payoffs, and the following modified updating and response rules:

A3′ Updating of propensities: If player *n* plays his *k*th pure strategy at time *t* and receives a reinforcement of *R(t,x)*, then the propensity to play strategy *j* is updated as a function of *R(t,x)*.

$$q_{\mathrm{nj}}(t+1) = \left[\frac{q_{\mathrm{nj}}(t)[C_{\mathrm{nj}}(t) + N(1)] + R(t,x)}{[C_{\mathrm{nj}}(t) + N(1) + 1]}\right] \tag{4}$$

where $C_{\mathrm{nj}}(t)$ is the number of times that strategy *j* has been played in the first *t* trials and $N(1)$ is a free parameter that determines the strength of the initial propensities. (Roth et al. simplified the model by the assumption $N(1) = 0$, this simplification is not imposed here.)

A4′ Exponential response rule: The probability $p_{\mathrm{nk}}(t)$ that player *n* plays his *k*th pure strategy at time *t* is

$$p_{\mathrm{nk}}(t) = \left[\frac{\exp[\lambda q_{\mathrm{nk}}(t)]}{\Sigma\ \exp[\lambda q_{\mathrm{nj}}(t)]}\right], \tag{5}$$

where the sum is over all of player *n*'s pure strategies *j* and λ is a free parameter that determines reinforcement sensitivity.

When the initial propensities are assumed to equal the average reinforcement from random choice, the basic reinforcement average model predicts no added constant effect. Thus, it captures the 2 and 6 results but fails to account for the findings summarized in Sections 2 and 3. In addition to this failure the model has two obvious limitations. First, it predicts that multiplying all payoffs by a positive constant will have a large effect. This prediction is inconsistent with previous results (e.g. Myers et al., 1963). A more important limitation involves the prediction concerning the effect of payoff variability. The REA model predicts that choice probability in probability learning tasks converges to a value that is solely determined by the difference between the expected values of the two alternatives. [6] A larger difference leads to more extreme choice probability independently of payoff variability. For example, it predicts that the probability of 'H' choices in Condition 0,4 considered above will be higher than in a 'noise-free' variant of this condition in which 'H' yields 1 point with certainty ($P_{\mathrm{h}} = 1$) and 'L' yields 0 with certainty.

Two modifications of the model are introduced here to address these limitations. First, the choice rule is modified to imply sensitivity to payoff variance and insensitivity to payoff

[6] It predicts that the probability of choosing alternative H will converge to $1/[1 + \exp(\lambda(1 - 2P_{\mathrm{h}})D)]$ where *D* is the difference between the good and bad outcomes.

magnitude. Under the modified choice rule:

$$p_{\mathrm{nk}}(t) = \left[\frac{\exp[\lambda q_{\mathrm{nk}}(t)/\mathrm{PV}(t)]}{\Sigma \exp[\lambda q_{\mathrm{nj}}(t)/\mathrm{PV}(t)]} \right], \tag{6}$$

where PV(t) is a measure of the payoff variability. PV(1) > 0 is assumed to equal the expected absolute difference between the obtained payoff from random choice and the average payoff from random choice. For example, in Condition 0,4 $\mathrm{PV}(1) = .(0.3|0-2| + 0.7|4-2| + 0.7|0-2| + 0.3|4-2|)/2 = 2$. It is then undated as an average absolute difference between the recent payoff and the accumulated average payoff:

$$\mathrm{PV}(t+1) = \left[\frac{\mathrm{PV}(t)(t + mN(1)) + |x - \mathrm{PA}(t)|}{t + mN(1) + 1} \right]$$

where PA(t) is the accumulated payoff average in trial t and m is the number of strategies (2 in the current probability learning tasks). PA(t) is calculated in a similar manner. PA(1) is the expected payoff from random choice (thus, $q_{\mathrm{nk}}(1) = \mathrm{PA}(1)$ for all n and k), and

$$\mathrm{PA}(t+1) = \left[\frac{\mathrm{PA}(t)(t + mN(1)) + x}{t + mN(1) + 1} \right].$$

To address the effect of the addition of a constant (payoff sign) the current model distinguishes between strategy selection and observed alternative selection. It assumes that DMs learn among three unobserved cognitive strategies: L, H and a 'loss aversion' (LA) strategy. A selection of one of strategies L or H implies a selection of alternatives 'L' or 'H' respectively. A selection of strategy LA in trial t implies a choice of the alternative that led to lower proportion of losses in the first $t-1$ trials. When the loss proportions cannot be reliably ranked (because they are identical as in Condition 0,4, or because the rank changes from period to period as in Condition −6,−2,0) this strategy is ignored. This model, referred to as REL, has two parameters: λ, $N(1)$.

6.4. Model comparison

Descriptive fit. The grid search procedure described in Section 6.1 was utilized to estimate parameters of the post hoc models described above. The estimated parameters and the MSD scores (multiplied by 100) over the nine probability learning tasks are presented in Table 1 (center). Since the three acceleration models have the same number of parameters the comparison of these models is simple: The best model is the one with the lowest MSD score. Table 1 shows that among the adjustable reference point models the best fit was obtained by the ADR variant that assumes that both the forgetting and the reference point functions are sensitive to the relative reinforcement size. This should not be too surprising, since this is the model that allows both perceived reinforcements and actions to respond most quickly when they are far from the average range. The fourth column in Figs. 1–3 presents the predictions of this model with the estimated parameters.

Table 1 also shows that the modified reinforcement average (REL) model fit of the probability learning data outperforms all the adjustable reference point models. The success of this model in accounting for the probability learning data is remarkable as it has fewer free

Table 1
MSD scores (100 × mean squared deviation — smaller is better) between the different predictions and the experimental results. The scores measure the average distance between block *n* of the data compared to round *n* of the prediction

Model	Estimated parameters					MSD (×100) score		
						Best fit of the 9 probability learning tasks	Prediction of the new game	Prediction of the 12 games studied in Erev and Roth
	$S(1)$	ϵ	ϕ	ω	α			
ARP	30	0.18	0.075	0.96	0.4	0.18	0.29	0.86
AD	1	0.2	0.02	0.94	0.5	0.19	0.68	1.01
AR	3	0.2	0.02	0.96	0.5	0.25	0.38	0.87
ADR	3	0.29	0.09	0.935	0.5	0.17	0.49	0.71
	$N(1)$	λ						
REL	30	2.8				0.12	0.25	0.95

parameters. The fifth column in Figs. 1–3 presents the predictions of this model. Comparison of this column to the fourth column (the ADR model) reveals that the main advantage of the REL is its ability to capture the average learning speed. The ADR incorrectly predicts that almost all the learning will occur in the first two blocks.

Generality. To evaluate the potential generality of the post hoc models and the estimated model we computed their predictions for the 12 games studied in Erev and Roth and the new game condition studied here. (Condition 0,1 was not considered as it is identical to one of the conditions in Erev and Roth.) Two hundred simulations of each of the 13 experimental games were run. Each simulation was a direct replication of the original experiment (including the number of trials and the matching rule for the players). The players in the simulations were programmed to behave according to the relevant model with the parameters that best fit the probability learning data (second column in Table 1). The simulation results are summarized by the same statistics that summarize the experimental data in Erev and Roth (choice proportions in 4–10 blocks). The model's MSD scores (right hand column in Table 1) measure the distance between the experimental and the simulation choice proportions. The results show that the ADR model provides the most accurate predictions of published data. Moreover, the fit provided by this and the REL model for the games studied by Erev and Roth (0.71 and 0.95) is good even relative to the fit of the models estimated in that paper (0.59–1.0).

7. Conclusions

The current research examined the implication of the finding that the addition of a constant to all payoffs can affect learning (Bereby-Meyer and Erev, 1998). Five main conclusions have been reached:

1. The effect is not limited to decision making under uncertainty. A small but significant effect of the addition of constant to all payoffs was observed in a 2×2 constant sum game.
2. The effect is not linear and/or trivial. While the addition of 2 units (from -2 and 2 to 0 and 4) had a large effect (in the BME study), the addition of 8 units (from -6 and -2 to 2 and 6) had no effect.

3. The non-linearity of the effect cannot be accounted for by the assumption that the initial reference point (that determines if an outcome increases or decreases the probability of choosing an action again) is a simple function of whether zero is in the payoff range.
4. The ARP model has to be modified to account for this non-linear effect. A modification that assumes that both discounting and the reference point adjustment process are sensitive to the relative reinforcement size outperforms simpler modifications. The modified model captures the current results as well as the results of the 12 matrix games studied by Erev and Roth (1998).
5. The results are better described by a simpler two-parameter average reinforcement model that assumes that DMs consider a loss aversion strategy. Yet, the modified reference point model appears to provide better account for the games considered by Erev and Roth.

In addition to these relatively 'direct' conclusions, the current research provides some support to the optimistic assertion that descriptive game-theoretic models can be rather general. The finding that the models proposed (and estimated) to account for behavior in simple decision tasks under uncertainty provides a good fit for behavior in two-person games, suggests that the models capture general principles. Thus, models that capture these principles can be used to derive useful ex ante predictions of behavior.

Examination of the common features of the two successful models suggests that the important principles are likely include: slow probabilistic adjustment process to previous outcomes that is sensitive to the addition of constant to all payoffs around 0 and to payoff variability; and relatively insensitivity to the addition of constants that do not change payoff sign, and to multiplication of all payoffs by a positive constant.

Finally, the fact that the two post hoc models are rather distinct suggests that even a rough approximation of these principles is sufficient to capture behavior, and that future research is needed to improve our understanding of the best approximation. In particular, we think that future research is needed on players' cognitive strategies, to provide an empirical basis on which to model particular strategies as being those among which players learn.

Acknowledgements

This research was supported in part by a grant from National Science Foundation to the Univ. of Pittsburgh. It has benefited from related research and insightful conversations with Bob Slonim, Joachim Meyer and Sharon Gilat and from the comments of the participants of the workshop on Economics and Psychology at the Univ. of British Columbia on June, 1997.

References

Bereby-Meyer, Y., Erev, I., 1998. On learning to become a successful loser: A comparison of alternative abstractions of learning in the loss domain. Journal of Mathematical Psychology 42, 266–268.

Borgers, T., Sarin, R., 1995. Naive reinforcement learning with endogenous aspirations Mimeo, University College, London.

Bush, R.R., Mosteller, F., 1955. Stochastic Models for Learning, Wiley, New York
Camerer, C.F., Ho, T.-H., 1998. EWA learning in games: Preliminary estimates from weak-link games. In: Budescu, D., Erev, I., Zwick, R. (Eds.), Games and Human Behavior: Essays in Honor of Amnon Rapoport, LEA, Hillsdale, NJ.
Cheung, Y.-W., Friedman, D., 1998. A comparison of learning and replicator dynamics using experimental data, Journal of Economic Behavior and Organization.
Edwards, W., 1961. Probability learning in 1000 trials. Journal of Experimental Psychology 62, 385–394.
Erev, I., Rapoport, A., 1998. Magic, reinforcement learning, reinforcement learning and coordination in a market entry game. Games and Economic Behavior 23, 146–175.
Erev, I., Roth, A.E., 1996. On the need for low rationality, cognitive game theory: reinforcement learning in experimental games with unique, mixed strategy equilibria, working paper, University of Pittsburgh.
Erev, I., Roth, A.E., 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. American Economic Review 88, 848–881.
Fudenberg, D., Levine, D., 1998. Theory of Learning in Games, MIT Press, Cambridge, MA
Harley, C.B., 1981. Learning the evolutionary stable strategy. Journal of Theoretical Biology 89, 611–633.
Herrnstein, R.J., 1970. On the law of effect. Journal of the Experimental Analysis of Behavior 13, 244–266.
Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47, 263–291.
Luce, D.R., 1959. Individual Choice Behavior, Wiley, New York.
March, J.G., 1996. Learning to be risk averse. Psychological Review 103, 309–319.
Mookherjee, D., Sopher, B., 1997. Learning and decision costs in experimental constant sum games. Games and Economic Behavior 19, 97–132.
Myers, J.L., Fort, J.G., Katz, L., Suydam, M.M., 1963. Differential monetary gains and losses and event probability in a two-choice situation. Journal of Experimental Psychology 66, 521–522.
Rapoport, A., Boebel, R.B., 1992. Mixed strategies in strictly competitive games: A further test of the minmax hypothesis, Games and Economic Behavior 4, pp. 261–283.
Rapoport, A., Erev, I., Abraham, E.V., Olson, D.E., 1997. Randomization and adaptive learning in a simplified poker game. Organizational Behavior and Human Decision Processes 69, 31–49.
Roth, A.E., Erev, I., 1995. Learning in extensive-form games: experimental data and simple dynamic models in intermediate term. Games and Economic Behavior, Special Issue: Nobel Symposium 8, 164–212.
Roth, A.E., Erev I., Slonim, R.L., 1998. Learning and equilibrium as useful approximations: accuracy of prediction on randomly constant sum games, discussion Paper.
Siegel, S., Siegel, A.E., Andrews, J.M., 1964. Choice, Strategy, and Utility, McGraw-Hill, New York
Suppes, P., Atkinson, R.C., 1960. Markov Learning Models for Multiperson Interactions, Stanford University Press, Stanford, CA
Tang, F.-F., 1996. Anticipatory learning in two-person games: An experimental study, Discussion paper B-363, University of Bonn.
Thorndike, E.L., 1898. Animal intelligence: An experimental study of associative processes in animals, Psychological Monographs, 2.