Journal of Mathematical Psychology 53 (2009) 155–167

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Learning, risk attitude and hot stoves in restless bandit problems

Guido Biele [a,*], Ido Erev [b], Eyal Ert [b]

[a] Max Planck Institute for Human Development, Germany
[b] Technion–Israel Institute of Technology, Israel

ARTICLE INFO

Article history:
Received 16 March 2007
Received in revised form
21 January 2008
Available online 11 July 2008

Keywords:
Dynamic decision making
Probability matching
Underweighting of rare events
The recency/hot stove paradox
Case-based reasoning

ABSTRACT

This research examines decisions from experience in restless bandit problems. Two experiments revealed four main effects. (1) Risk neutrality: the typical participant did not learn to become risk averse, a contradiction of the hot stove effect. (2) Sensitivity to the transition probabilities that govern the Markov process. (3) Positive recency: the probability of a risky choice being repeated was higher after a win than after a loss. (4) Inertia: the probability of a risky choice being repeated following a loss was higher than the probability of a risky choice after a safe choice. These results can be described with a simple contingent sampler model, which assumes that choices are made based on small samples of experiences contingent on the current state.

Previous studies on decisions from experience (e.g. Denrell (2007), Erev and Barron (2005), Hertwig, Barron, Weber, and Erev (2004), March (1996) and Weber, Shafir, and Blais (2004)) have focused on simplified *static* environments. In a typical study (see review in Erev and Barron (2005)), the decision maker is faced with the same choice problem repeatedly, and the payoff for each selection is determined by a draw from the selected alternative's payoff distribution. The choice problem can be considered static in that the relevant distributions do not change during the experiment. The main goal of the current research is to extend this research paradigm to address more dynamic decision environments. The set of problems considered here comprises one-armed restless bandit problems (see Whittle (1988)). The key feature of such problems, as compared to conventional armed bandit problems, is that the payoff distribution of an alternative can change, even when this alternative is not chosen. In each trial of the experiments reported here the decision makers had to decide between a *safe prospect* that provided a *medium payoff* with certainty, and a *risky prospect* whose payoff depended on the state of nature. The payoff from risky choices was *high* if the state was positive, and *low* if the state was negative.

The exact state was determined with a two-state Markov process. If the state was positive at trial t, it stayed positive at trial $t+1$ with probability p. If the state was negative at trial t, it stayed negative with probability q. This process is summarized in Table 1. The goal of the decision maker confronted with this problem is to maximize the payoff over a given time horizon.

Table 1
The basic Markov process for a risky option

		Trial $t+1$	
		H	L
Trial t	H	p	1-p
	L	q	1-q

Note: The entries represent the transition probabilities between the two states of nature "high payoff" (H) and "low payoff" (L).

While the general setup of the problem seems relatively simple, achieving optimal outcomes is not. The important challenge of any armed bandit problem is to solve the exploration–exploitation dilemma. The exploration–exploitation dilemma refers to the fact that in each decision in an armed bandit task, the decision maker can choose between opting for the alternative with the known higher payoff, or opting for another alternative in order to increase knowledge about seemingly inferior alternatives. For conventional armed bandits this can be achieved by using the Gittins strategy (Gittins, 1979), which prescribes choosing in every round the alternative with the higher Gittins index. Intuitively, the Gittins index is the sum of an alternative's expected payoff and the information value from choosing this alternative. In restless bandit problems, however, because payoff distributions of nonchosen alternatives do not remain stable, solving the exploration–exploitation dilemma is intractable (Papadimitriou & Tsitsiklis, 1999). Therefore, heuristic solutions to the dilemma have been proposed (e.g. Glazebrook, Ruiz-Hernandez, and Kirkbride (2006)).

* Corresponding address: Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition & IJRG Neurocognition of Decision Making, Lentzeallee 94, 14195 Berlin, Germany.
E-mail address: biele@mpib-berlin.mpg.de (G. Biele).

0022-2496/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2008.05.006

Given the inherent difficulty of achieving maximal payoffs in restless bandit problems, successful behavior depends critically on the decision makers' prior information. For instance, providing decision makers with information about transition probabilities allows them to determine optimal choices in every round. The current analysis focuses on pure decisions from experience; the decision makers were told nothing about the payoff rule and the number of trials to be played. Indeed, they were not even told that they were facing a one-armed restless bandit problem. Instead, they had to base their decisions on their experience: the obtained payoff after each trial. This constraint implies that almost any behavior can be justified as "rational" given a particular set of beliefs.[1] Thus, the current analysis is not focused on evaluating people's rationality. Rather, it examines how they behave when their prior information is limited (and the available information does not allow computation of the optimal strategy).

Notice that under the constraint $p = q$, the restless bandit problem becomes a static choice environment. Thus, the set of problems considered here is a generalization of simple static decision problems. This generalization seems like a natural first step for a study that aims to extend static environment research, because it involves the addition of only one parameter (the parameter q).

Another attractive feature of restless bandit problems is that they are a natural abstraction of basic foraging problems. For example, consider a fisherman (or other fishing animal) who has to decide between fishing in a pond or in a river. The pond yields a relatively stable return, and the river's return depends on whether a school is in the area. Under the assumption that the schools' behavior is determined by a Markov process (where transitions can occur regardless of whether the fisherman goes fishing), the fisherman's dilemma is an example of the restless bandit problem examined here.

The current investigation starts with an experimental study of the robustness of one of the most important phenomena detected in the study of static choice environments, the hot stove effect (see Denrell and March (2001)) in the context of more dynamic situations (e.g., restless bandit problems). In the second part of the paper, we propose and evaluate alternative models that capture the main behavioral regularities observed in the study of static choice problems and the main results observed in the wider set of situations considered here.

1. The hot stove effect[2]

March (1996) showed that popular learning models proposed to capture human behavior in simple settings predict "learning to become risk averse". That is, the tendency to select risky (high variability) options is expected to decrease as the decision maker gains experience.

Denrell (2005, 2007) and Denrell and March (2001) extended this analysis and demonstrated that risk aversion is a product of the "hot stove effect": The hot stove effect is a natural consequence of the inherent asymmetry between the effect of good and bad experiences. Good outcomes increase the probability that a choice will be repeated and therefore facilitate the collection of additional information concerning the value of the alternative that has yielded the good outcome. Bad outcomes reduce the probability that the choice will be repeated and consequently impair the collection of additional information concerning the value of the alternative that has yielded the bad outcome. As a result, the effect of bad outcomes is stronger (lasts longer) than the effect of good outcomes. Since options with a high variability are more likely to produce bad outcomes, the hot stove hypothesis predicts a decreasing tendency to choose such options.

For example, consider Problem 1 with a constant payoff of 0 for the safe alternative and payoffs of $+1$ or -1 for the risky alternative, which uses Table 1's transition matrix with the parameters $p = q = 0.5$. A hot stove effect would be seen in a decrease in the proportion of risky choices with experience. This effect is expected because a sequence of bad outcomes from risky choices can reduce the average payoff from the risky alternative to below 0. This result is expected to cause the decision maker to prefer the safe alternative, and when the safe alternative is selected the fact that the risky alternative is perceived as having a lower average payoff does not change. Denrell (2007) showed that the hot stove effect is robust to the assumed level of rationality. It is the likely result of low-rationality, reinforcement-learning algorithms as well as high-rationality, utility-maximizing-learning algorithms.

The hot stove effect can be used to explain some of the best-known regularities observed in empirical behavioral research. Examples include taste aversion (see Garcia, Ervin, and Koelling (1966)), learned helplessness (see Seligman, Maier, and Geer (1968)), and the "dormitory effect" (see Denrell (2005)). In addition, it is consistent with the outcome of direct examinations of decisions from experience. When feedback is limited to the obtained payoffs the tendency to select the safe option is apt to increase with experience (see Erev and Barron (2005), Grosskopf, Erev, and Yechiam (2006), Munichor, Erev, and Lotem (2006) and Yechiam and Busemeyer (2006)). The availability of complete feedback that includes information concerning the foregone payoffs tends to increase risk taking (e.g. Grosskopf et al. (2006) and Yechiam and Busemeyer (2006)). However, the observed magnitude of the hot stove effect is not large. For example, Grosskopf et al. (2006) found that the availability of foregone payoff information increased maladaptive risk-taking behavior in the first 100 trials of their experiment, but this effect disappeared in the longer term.

We chose to start our analysis of behavior in restless bandit problems with a focus on the hot stove effect because the expected magnitude of this effect in these problems is highly sensitive to the assumed nature of the learning process. To clarify this point recall that Denrell (2007) proposed two justifications for the hot stove effect. The first involves the observation that basic learning models assume high sensitivity to recent outcomes. Models of this type suggest that restless bandit problems with positive sequential dependency ($p > .5, q < .5$) are expected to enhance the hot stove effect. The logic behind this prediction is simple: positive sequential dependency increases the probability of long sequences of bad outcomes, and sequences of this type create the hot stove effect.

The second justification involves rational considerations. In a static environment, rational exploration diminishes with time. In contrast, in the context of restless bandit problems, rational learning assumes a certain level of exploration even after

[1] Without any prior information, a decision maker could view the problem as an armed bandit, as a restless bandit, but also as a partially observable Markov decision process (POMDP; e.g. Monahan (1982)) or a Markov decision problem with imprecise probabilities (MDPIP; (e.g. White and Eldeib (1994)). In POMDPs the decision maker has only indirect information about the state of the world and in MDPIPs transition probabilities are not provided, and it is assumed that the decision maker interacts with a system that tries to minimize the decision maker's payoffs. Note that only approaches to bandit problems explicitly try to solve the exploration–exploitation dilemma, whereas solutions to POMDPs and MDPIPs aim to solve the planning problem in complex environments.

[2] An early reference to the hot stove effect can be found in Mark Twain's and Samuel L. Clemen's book "Following the Equator: A Journey Around the World", which the authors begin with "We should be careful to get out of an experience only the wisdom that is in it–and stop there; lest we be like the cat that sits down on a hot stove-lid. She will never sit down on a hot stove-lid again–and that is well; but also she will never sit down on a cold one any more". The authors cite "Pudd'nhead Wilson's New Calendar" as a source.

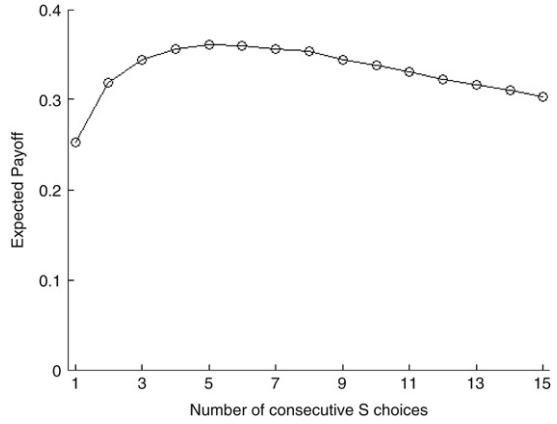

Fig. 1. The expected payoff from strategies of the type "Choose risky after high payoff, and after a sequence of n safe (S) choices; select the safe alternative otherwise" in Condition 0.95. The x-axis presents the value of n. The expected value is maximized with $n = 6$.

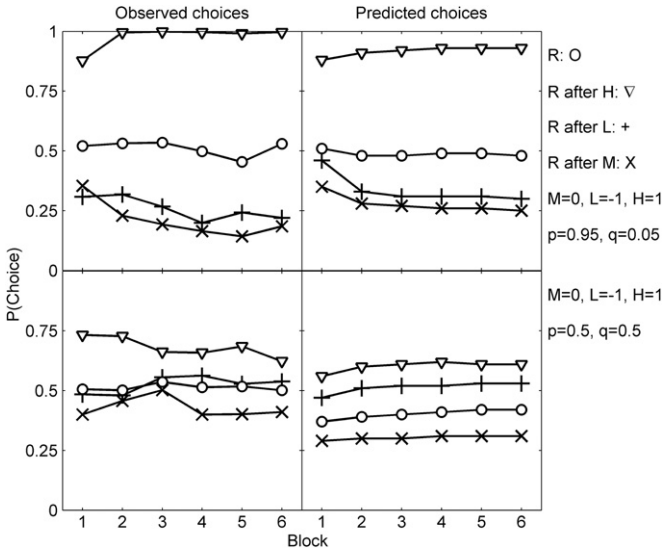

Fig. 2. The observed and predicted choice probabilities in Experiment 1 as a function of time (five blocks of 60 trials each). Each row depicts one problem. The left-hand graphs show the experimental results, the right-hand graphs the predictions of the model. The five parameters that define the problem are presented on the right: M (medium high) is the payoff from the safe option, H and L are high and low payoff from the risky option, and q and p are the transition probabilities (cf. Table 1). The R curve, designated with open circles, shows the mean choice rate of the risky option. The other three curves (R after H, R after M, and R after L) refer to the choice rate of the risky option following a high, medium, and low outcome, respectively.

extensive experience. Continuous exploration is necessary to collect information concerning the state of nature. Thus, rational considerations can negate the hot stove effect in restless bandit problems. For example, consider Problem 1 with the parameters $p = 1 - q = 0.6$. Under the assumption that the decision maker's prior beliefs about the payoff distributions are accurate, the optimal strategy in this problem involves 75% risky choices independent of the experience length.[3]

2. Experiment 1: The effect of sequential dependencies on the hot stove effect

The first experiment examined the effect of the transition probabilities (the values of p and q) in two variants of Problem 1. It considered the cases $p = 1 - q = 0.95$ (Condition 0.95), and $p = 1 - q = 0.50$ (Condition 0.50). The two conditions were studied in a repeated choice task for 300 trials with immediate feedback about the chosen option. The high value of parameter p in Condition 0.95 implies a positive sequential dependency. The state is not likely to change between sequential trials. Assuming that the payoff rule is known, the optimal (expected value maximizing) strategy in this condition implies risky choices in three cases: in the first trial; after obtaining high payoff (+1); and after a sequence of six choices of the safe alternative. The expected payoff given the optimal strategy is 0.35, as demonstrated in Fig. 1.[4] Condition 0.50 is a static environment without sequential dependency.

2.1. Apparatus and procedure

Participants were informed that they would be playing some games on a "computerized money machine" (see a translation of the instructions in Appendix A) with two unmarked buttons for an unspecified number of trials, but they did not receive prior information about the games' payoff structure. The participants' task was to select one of the machine's unmarked buttons in each of 300 trials. Payoffs were contingent upon the button chosen and were determined by the payoff distributions and Markov process described above. Two types of feedback immediately followed each choice: (1) The payoff for the last choice, which appeared on the selected key for 2 s, and (2) the accumulated-payoff counter, which was displayed constantly. The participants were told that their goal was to maximize their earnings, and the accumulated points for one of the games would be converted to cash at the conversion rate of 10 agorot (2.22 US cents) per point. The assignment of risky and safe alternatives to buttons was randomly determined for each participant at the beginning of the experiment and was fixed during the experiment.

2.2. Participants

Fifty-four Technion students served as paid participants in the experiment. Twenty-seven participants were randomly assigned to each of the conditions (in their first game).[5] They were given a payment of 25 shekels for attending, from which they could win or lose money according to their obtained payoffs in the experimental task. Final payoffs ranged between 19 and 39 shekels (about \$4.20–\$8.70).

2.3. Results

The left column of Fig. 2 presents the main experimental results (the right column presents the prediction of the model discussed below). It shows the observed proportion of risky choices over trials and conditioned on the last payoff, as a function of time and experimental condition. The results highlight four main observations: (1) almost no evidence for the hot stove effect, (2) high sensitivity to the transition probabilities in Condition 0.95,

[3] To obtain this result we simulated a decision rule that chooses the risky option again after a high outcome, switches to the safe option after a low outcome, and switches back from safe to risky after n consecutive safe choices, where n was constrained to be within 1 and 15. For the problem parameter $p = 1 - q = 0.6$, the best switching time is after only one safe choice.

[4] The result was obtained with a simulation as described in footnote 3. The participants in the current study did not know the payoff rule. Thus, there was no reason to expect that they would follow the optimal strategy. This strategy is used here as a benchmark.

[5] Upon completion of the first game the participants were asked to participate in a second game (that was designed to examine an order effect). This second game is not relevant to the current investigation and is not reported here.

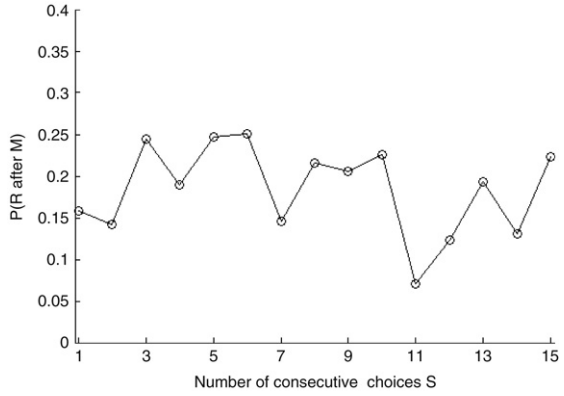

Fig. 3. The proportion of risky choices following n consecutive safe (S) choices in Condition 0.95.

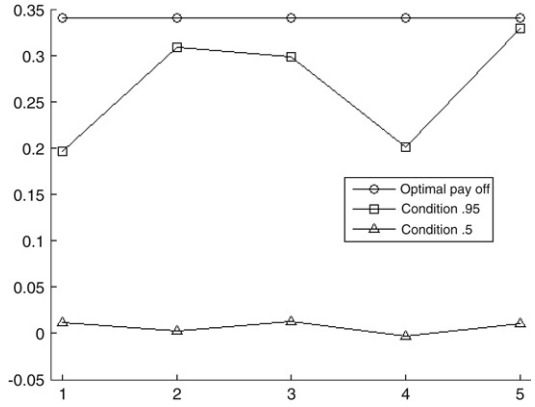

Fig. 4. Observed payoffs in six blocks of 50 trials in Experiment 1.

(3) a positive recency effect, and (4) inertia. We discuss these observations in the following section.

A (weak) hot stove effect. The curves marked by circles in Fig. 2 depict the overall proportion of risky choices per block. Over the 300 trials, the observed rates in the two conditions were virtually the same: .51(SD = .17) in Condition 0.95 and .51(SD = .23) in Condition 0.50. Interestingly, the learning curve in Condition 0.95 shows a weak nonmonotonic V shape; risky choices decrease initially and then start increasing toward the end of the session. This finding, and the finding of more than 50% choices of the risky option in both conditions, is inconsistent with the notion of the hot stove effect. In the long run, participants did not learn to become risk averse.

Partial sensitivity to transition probabilities. The "risky after high" and "risky after low" curves in Fig. 2 represent the choice rate of the risky option immediately after receiving high and low payoffs, respectively, and indicate sensitivity to the incentive structure. Over the 300 trials, the observed rate of risky choice after high payoff in Condition 0.95 (when the conditional probability for high payoff was 0.95) was .98(SD = .07). The observed rate of risky choice after low payoff in Condition 0.95 (when the conditional probability for high payoff was 0.05) was .27(SD = .29). This difference is significant, $t(26) = 14.33$, $p < .0001$, and clearly larger than the difference between the "risky after high" and "risky after low" curves in Condition 0.50, $t(52) = 8.25$, $p < .0001$ for the difference between the two differences.

Fig. 3 presents the probability of choosing the risky option as a function of the number of consecutive choices of the safe option in Condition 0.95. In this case, the participants exhibited limited sensitivity to the incentive structure. They did not tend to switch after 6 consecutive choices of the safe alternative, which is, as Fig. 1 shows, the best time to switch. In contrast, the results reveal a relatively flat curve with a higher tendency to explore after 4–10 safe choices than after other sequences. However, converging to the optimal switching time seems difficult, as Fig. 1 also shows that payoffs for switching after 3 or more choices of the safe alternative are similar.

Fig. 4 presents the average payoff as a function of time. Recall that the expected payoff under the optimal choice strategy in Condition 0.95 is around.35. The participants appear to approach this value. The mean payoff in the last block is.32.

A recency effect. The difference between the risky after high and risky after low curves in Condition 0.50 is smaller than the corresponding difference in Condition 0.95, but it is still significant. Over the 300 trials, the proportions of risky choices in Condition 0.50 are .70(SD = .28) after high payoff and .55(SD = .32) following low payoff. The significant difference, $t(26) = 2.45$, $p < .03$, suggests a positive recency effect: Recent good outcomes increase the tendency to select the risky alternative.

Inertia. Additional examination of Condition 0.95 reveals that the choice rate of the risky option following low payoff outcome (the risky after low curve) is higher than the choice rate of the risky option following a safe choice (the risky after medium curve). A similar pattern was observed in Condition 0.50. These results appear to contradict the positive recency effect, which predicts more switching after low outcomes.[6] A natural explanation of these results is that they reflect inertia (see Erev and Haruvy (2005)): a tendency to repeat the last choice, irrespective of the obtained outcome.

2.4. Relationship to previous studies of decisions from experience

In a recent paper, Erev and Barron (2005) reviewed the main behavioral regularities observed in experience-based decisions. Ten of the static choice problems of their analysis are members of the problem space studied here. These problems involve a choice between a sure medium payoff and a risky alternative with (up to) two possible outcomes. As in the current study, the information available to the decision makers was limited to the payoffs they experienced. Table 2 presents these problems, and the observed choice proportions in the last 100 trials of the experiments.

Erev and Barron (2005) did not report the conditional choice proportions or the sequential dependencies. This analysis, however, is highly relevant to the current context. Therefore, we reanalyzed their data to evaluate whether the sequential dependencies described by positive recency and inertia can be detected in static environments as well. The learning curves are presented in Fig. 5.

The results highlight the robustness of the positive recency and inertia patterns discussed above. Recency, as estimated above, was observed in 8 of the 10 problems (the other two, E&B3 and E&B4, involved two safe prospects). Six of these 8 problems reveal a positive recency effect: The probability of risky choices is higher after high payoff than after low payoff. All 8 problems reveal inertia: The probability of risky choices is higher after low payoff than after medium payoff (a choice of the safe prospect).

3. A quantitative summary

Recent research demonstrates that decisions from experience in static settings can be captured with simple models that assume reliance on small samples of experiences (see Erev and Barron

[6] Notice that these results cannot be captured with a simple exploration assumption since it predicts similar rates of risky choice after high and low outcomes. Nor can they be captured with simple exploitation since it predicts an equal rate of risky choices after low outcomes and after safe choices in the 0.5 condition, and lower rate after low outcomes than after medium outcomes (that follow safe choices) in the 0.95 condition.

Table 2
The payoff distributions, observed results (Obs), and predictions of the three sampler models for Experiment 1 and for 10 selected problems from Erev and Barron (2005)

Problem description						Choice proportion															
Probl.	Med (Safe)	Low (Risky)	High (Risky)	p	q	Risky				Risky after low				Risky after medium				Risky after high			
						Obs	Nv	Ct	4m	Obs	Nv	Ct	4m	Obs	Nv	Ct	4m	Obs	Nv	Ct	4m
Cond. 0.95	0	−1	1	.95	.05	.51	.51	.49	.49	.33	.23	.23	.31	.25	.15	.15	.26	.98	.50	.93	.93
Cond. 0.50	0	−1	1	.5	.5	.51	.49	.54	.42	.54	.59	.59	.53	.31	.38	.38	.31	.68	.49	.50	.61
E&B1 (21)	3	0	4	.8	.8	.67	.62	.66	.67	.68	.52	.52	.69	.49	.38	.38	.50	.88	.62	.65	.78
E&B2 (40)	−3	−4	0	.2	.2	.43	.36	.26	.30	.46	.64	.64	.46	.22	.32	.32	.23	.55	.36	.13	.54
E&B3 (MI)	11	10	10	1	1	.10	.07	.05	.07	.29	–	–	.25	.05	.06	.06	.05	–	–	–	–
E&B4 (MI)	−10	−11	−11	1	1	.05	.07	.05	.07	.28	–	–	.25	.05	.04	.04	.05	–	–	–	–
E&B5 (2)	11	1	19	.5	.5	.29	.41	.36	.39	.52	.47	.47	.51	.28	.19	.19	.29	.74	.41	.27	.59
E&B6 (6)	−11	−19	−1	.5	.5	.50	.56	.55	.55	.63	.49	.49	.63	.40	.37	.37	.42	.79	.56	.52	.72
E&B7 (23)	3	0	32	.1	.1	.32	.34	.29	.35	.48	.81	.81	.49	.24	.11	.11	.28	.80	.37	.46	.65
E&B8 (25)	−3	−32	0	.9	.9	.61	.62	.61	.60	.67	.55	.55	.65	.44	.46	.46	.44	.88	.62	.60	.72
E&B9 (3)	10	1	21	.5	.5	.57	.57	.54	.54	.61	.65	.65	.62	.40	.25	.25	.41	.68	.56	.52	.70
E&B10 (5)	−10	−21	−1	.5	.5	.45	.41	.35	.40	.52	.44	.44	.51	.30	.30	.30	.31	.88	.42	.27	.60

Note: Med (Safe) are payoffs for the safe alternative, Low (Risky) and High (Risky) are the high and low payoffs for the risky alternative. The models are naïve sampler (Nv), contingent sampler (Ct), and 4-mode contingent sampler (4m). Reported results are for the last 100 trials of each Problem. No positive recency effect was observed in problems E&B5, E&B9, and E&B10. The 10 selected problems from Erev and Barron (2005) are designated E&B1 through E&B10; original IDs are in parentheses.

(2005), and related ideas in Fiedler (2000), Kareev (2000) and Osborne and Rubinstein (1998)). To clarify the relationship of the current results to this research we chose to try to capture the observed trends with three abstractions of the sampling assumption.

3.1. The naïve sampler model

Similar to empirical (e.g. Daw, O'Doherty, Dayan, Seymour, and Dolan (2006)), theoretical (March, 1991), and normative (e.g. Gittins (1979)) approaches to experience-based decision making, the naïve sampler model (see Erev, Ert, and Yechiam (in press)) assumes two decision modes (or cognitive strategies): exploration and exploitation. Exploration implies random choice. Exploitation at trial t implies a draw of $m(t)$ past recalled experiences with each alternative and a selection of the alternative with the highest sample mean. The value of $m(t)$ is assumed to be randomly drawn from the set $\{1, 2 \ldots \kappa\}$ where κ is a free parameter that determines the maximal sample size.

The probability of exploration is 1 in the very first trial, and it reduces toward an asymptote with experience. The value of this asymptote, referred to as ε, is a free parameter. The effect of experience on the probability of exploration depends on the expected length of the experiment (T). Exploration diminishes quickly when T is small, and slowly when T is large. This assumption is quantified as follows:

$$P\,(\text{Explore}_t) = \varepsilon^{\frac{t-1}{t+T^{\delta}}},$$

where δ is a free parameter that captures the sensitivity to the length of the experiment.

The experience in the very first trial is assumed to be encoded as the first experience with both alternatives. Thus, in the first exploitation trial the decision maker has at least one "experience" with each alternative. The sampling from the relevant experience is performed with replacement.[7]

Estimation and results. Notice that one of the model's three parameters, the value of the asymptote (ε), can be directly estimated from the results of Problems E&B3 and E&B4 reported in Table 2. Under the current model, the deviations from maximization in these trivial problems in the long term reflect persistent exploration. Specifically, because the payoffs for the two alternatives are always the same, not choosing the high payoff option after long-term experience is most readily explained by the assumption of some minimum exploration in a stable environment. Thus, assuming that the observed rate converges N toward 5% deviations, the implied value is $\varepsilon = 0.1$.

The estimation of the remaining two parameters was based on a grid search with two criteria. The first was the mean square deviation (MSD) between the predicted and the observed (*unconditional*) proportions of risky choices in the last 100 trials of the 12 problems summarized in Table 2. The second criterion was the MSD between the predicted and observed *conditional* proportions of risky choices in the last 100 trials of the problems summarized in Table 2. We searched for parameters that minimized the sum of the two MSD scores. The estimated values were maximal sample size of $\kappa = 8$, and $\delta = 0.6$. The implied predictions of the naïve sampler model are presented in Table 2 under the "Nv" columns. The corresponding MSD scores (between the model and the mean over participants) are 0.005 for the unconditional choice rates, and 0.082 for the conditional rates. These values suggest that the model fits the unconditional rates reasonably, but it clearly fails to capture the conditional choice rates. The largest failure occurs in Condition 0.95 of Experiment 1. The naïve sampler model cannot capture the large difference between the proportions of risky choices after low and high payoffs.

3.2. The contingent sampler model

The second model considered here is a refinement of the naïve sampler model that incorporates two ideas: sensitivity to payoff contingencies and partial sensitivity to the average experience. The first change involves the abstraction of instance-based or case-based reasoning (see related concepts in Gilboa and Schmeidler (1995), Gonzalez, Lerch, and Lebiere (2003), Hochman and Erev (2007) and Logan (1988)). To capture the observed sensitivity to sequential dependencies, the refined model assumes that the sampling process is contingent on the last choice and its payoff. The sample is contingent in the sense that experiences are considered only if they occurred after the same sequence (contingency) of events as the current contingency.

A contingency is defined by the previous choice and its outcome. Thus, there are three relevant contingencies: (safe

[7] In addition to the assumptions listed above, Erev et al. (in press) and Erev, Roth, Slonim, and Barron (2007) assumed a subjective value function that captures diminishing sensitivity relative to a reference point. We ignore this assumption because Erev et al.'s results imply an approximately linear function for low absolute payoffs, like the payoffs considered here.

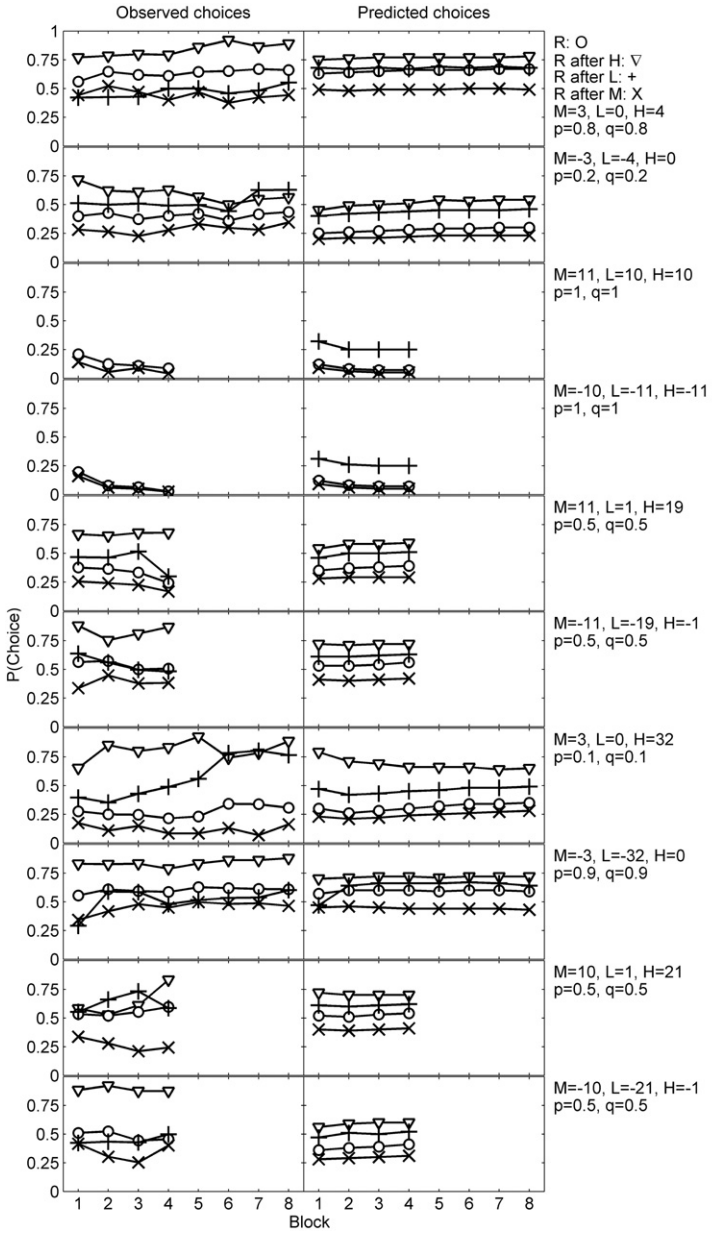

Fig. 5. The observed and predicted results in the 10 problems studied in Erev and Barron (2005) using Fig. 2's format. Each block summarizes 50 trials. The left and right columns show participants' behavior and predictions of the 4-mode contingent sampler model, respectively. The five problem parameters are presented on the right: M (medium) is the payoff from the safe option, H and L are high and low payoff from the risky option, and q and p are the transition probabilities. The R curve (open circles) shows the mean choice rate of the risky option. The other three curves, R after L, R after M, and R after H, show the choice rate of the risky option immediately after low, medium, and high outcomes, respectively. Note that a recency effect is present when the R after H curve is above the R after L curve (a risky choice is more frequently repeated after high payoff), and an inertia effect is present when the R after L curve is above the R after M curve (the repetition of a risky choice after a low payoff is more frequent than a switch from a safe to a risky choice). No positive recency effect was observed in problems E&B5, E&B9, and E&B10.

choice→medium payoff), (risky choice→low payoff), and (risky choice→high payoff). That is, during exploitation under a particular contingency the decision maker is assumed to recall κ experiences with each action under this contingency. Past experiences are assumed to be grouped in six distinct sets. Each set includes the outcome of a particular choice (risky or safe alternative) after one of the three contingencies. In all other respects, the contingent sampler model is identical to the naïve sampler model.

For example, assume that a decision maker makes the following sequence of choices with corresponding outcomes in Trials 1 to 10 (T1 to T10):

$T1$: *safe* $\rightarrow 0$, $T2$: *safe* $\rightarrow 0$, $T3$: *safe* $\rightarrow 0$,
$T4$: *risky* $\rightarrow 1$, $T5$: *risky* $\rightarrow 1$, $T6$: *risky* $\rightarrow -1$,
$T7$: *risky* $\rightarrow 1$, $T8$: *risky* $\rightarrow 1$, $T9$: *safe* $\rightarrow 0$,
$T10$: *risky* $\rightarrow 1$,

where each choice is indicated before the arrow and its corresponding outcome is after the arrow. Consider now the decision maker's decision in Trial 11 assuming an exploitation mode. When assessing the value of selecting the risky alternative in this trial, the decision maker is assumed to sample κ times from the set $\{0, 1, -1, 1\}$. The four outcomes in this set reflect the outcomes of italicized Trials 1, 5, 6, and 8 respectively, which are the outcomes of choosing the risky option after receiving a high payoff. When evaluating the value of selecting the safe alternative, the decision maker considers the outcomes of italicized Trials 1 and 9 (trial 9 is

the only case of selecting the safe alternative after obtaining a high payoff from the risky option). Thus, the relevant set is {0, 0}.

The second modification builds on the work of Lebiere, Gonzalez, and Martin (2007) and involves the relaxation of the assumption that the exploitation decisions are entirely determined by small samples.[8] Under the instance-based interpretation of the reliance on small samples the decision in trial t is made based on the sample of the $m(t)$ experiences in cases that seem most similar to the situation in trial t. That is, only the most similar experiences are considered. Lebiere et al. (2007) showed that this strong assumption is not necessary to explain the results reviewed by Erev and Barron (2005). The data can be captured with the assertion that small samples of similar experiences receive more weight than other experiences. The current model implements this idea with the assumption that one of the experiences sampled by the subject is the "average experience". The payoff associated with each alternative given this experience is the average observed payoff of this alternative during the experiment. Thus, one of the $m(t)$ experiences that is used to evaluate each alternative at trial t is the average experience with this alternative, and the remaining $m(t) - 1$ experiences are specific experiences, as in the original naïve sampler model.

Estimation and results. The estimation procedure described above was used again. The estimated parameters are $\varepsilon = 0.1$, $\kappa = 9$, and $\delta = 0.4$. The corresponding MSD scores are 0.009 for the unconditional choice rate, and 0.019 for the conditional rates. These results suggest that the contingent sampling assumption improves the fit of the conditional probabilities but impairs the fit of the overall choice rates. Moreover, the predicted rates of risky choice (cf., the "Ct" columns in Table 2) reveal that the model under-predicts the inertia and recency effects discussed above.

3.3. The 4-mode contingent sampler model

The third model considered here is an extension of the contingent sampler model with an explicit abstraction of recency and inertia as additional decision modes. The mode selection process is assumed to involve three steps. The first step determines whether exploration is selected. The probability and the implications of this mode are the same as in the models described above.

The second stage determines if the inertia mode is selected. The probability of inertia at this stage ι, where $0 < \iota < 1$ is a free parameter. Notice that the second stage can be reached only if the exploration mode was not selected. Thus, the unconditional probability of inertia is $[1 - P(\text{Explore}_t)]\iota$.

The third stage determines if the recency mode, implemented as a simple win-stay, lose-shift choice rule (WSLS; e.g., Nowak and Sigmund (1993)), is selected. The probability of this mode, given that neither the exploration nor the inertia mode is selected, is determined by the free parameter ω, and the last choice. The probability (conditioned on reaching the third mode) is ω if the risky option was selected in the last trial, and 0 otherwise. The unconditional probability after a risky choice is $\{1 - [1 - P(\text{Explore}_t)]\iota\}\omega$. As implied by its name, the WSLS rule requires repeated risky choice if and only if the last payoff was high; the safe alternative option is chosen otherwise. Notice that WSLS requires only reliance on short-term memory.

If none of the first three modes was selected, the decision is made in the exploitation mode as in the basic contingent sampler model. Notice that the exploitation mode is the only mode that uses long-term memory, as we assume that the sampling probability is independent of time and order of experiences. Exploration requires no memory, whereas inertia and WSLS require reliance on short-term memory.

Estimation and results. Extension of the logic of the estimation procedure used above allows direct sequential estimation of three of the five parameters from the data. The first step in this sequential estimation involves the estimation of ε and is as explained above (with value $\varepsilon = 0.1$).

Two additional parameters can be estimated from the long-term choice proportions in Condition 0.50 of Experiment 1 ($p = q = 0.5$). When the experience approximates the true payoff distributions, the implications of the different strategies in this problem are known. In the long term, when exploration converges to ε, the probabilities of selecting the different strategies are determined by the parameters ε, φ, and ρ. Appendix B shows that the model assumes the following constraints (C1–C4):

(C1) $P(\text{risky after high}) - P(\text{risky after low}) = P(\text{WSLS})$

(C2) $P(\text{risky after low}) - P(\text{risky after medium}) = P(\textit{inertia}) - 0.5 \cdot P(\text{WSLS})$

(C3) $P(\text{inertia}) = (1 - \varepsilon)\iota$

(C4) $P(\text{WSLS}) = [1 - \varepsilon - (1 - \varepsilon)\iota]\omega$.

In the final block of Condition 0.50, the difference of P(risky after high) and P(risky after low) equals .09, and the difference of P(risky after low) and P(risky after medium) equals .22, which indicates an inertia level of $\iota = 0.31$, and WSLS level of $\omega = 0.14$.

The estimation of the remaining two parameters was based on the two-criterion grid search procedure described above. The estimated values were sample size of $\kappa = 10$, and $\delta = 0.2$. The corresponding MSD scores are 0.0058 for the unconditional rates, and 0.0017 for the conditional rates.

The "4m" columns in Table 2 and the right-hand column in Figs. 2 and 5 present the models' predictions. A qualitative evaluation shows that the model reproduces the four effects considered above: limited hot stove, sensitivity to the transition probabilities, inertia, and recency. Additional indication of the quantitative fit is provided by the high correlations between the observed and predicted choice proportions over the 12 problems. The exact Pearson correlations are .95 for P(risky), .36 for P(risky after low), .84 for P(risky after medium), and .76 for P(risky after high).

3.4. Summary

The current analysis suggests that the simpler naïve sampler model does not provide a satisfactory account of the present findings. Yet, this model can be easily extended to capture the results. The supported extension involves three assumptions: contingent sampling, recency, and inertia.

4. Experiment 2: Extension of the dynamic environments

Experiment 1 was designed to study behavior under a "clear" dynamic environment. Three features of the dynamic condition ($p = 0.95$) facilitated this clarity: (1) The state was changed repeatedly but not too often (only once in 20 trials on average); (2) the changes were symmetric; and (3) the effort to respond to the changes could increase expected payoff (from 0 to 0.35) but was not likely to impair expected earnings (the expected earning from random choice was 0).

The main goal of Experiment 2 was to examine the generality of the results in environments that are less friendly to decision makers who make the effort to respond to sequential dependencies. To achieve this goal we studied 20 randomly selected problems from a wider space of the dynamic choice problems considered here.

[8] The motivation to relax this assumption comes from a thought experiment in which the decision makers are asked to select between a safe alternative that yields 1000 with certainty, and a risky option that yields 1001 with probability 0.95, and 0 otherwise. Models that assume reliance on small samples (e.g., the naïve sampler models with the parameters estimated above) predict a tendency to select the risky option (because most samples are not likely to include the 0 outcome).

Table 3
The payoff distributions, observed results (Obs), and predictions of the three models for the 20 problems of Experiment 2

Problem description						Choice proportion													
Nr.	Med (Safe)	Low (Risky)	High (Risky)	p	q	Risky				Risky after low				Risky after medium				Risky after high			
						Obs	Nv ENO = 14.18	Ct ENO = 13.22	4m ENO = 10.50	Obs	Nv ENO = 8.75	Ct ENO = 2.50	4m ENO = 7.49	Obs	Nv ENO = 4.40	Ct ENO = 3.66	4m ENO = 22.96	Obs	Nv ENO = 1.26	Ct ENO = 1.06	4m ENO = 15.38
1	−5	−8	6	.9	.5	.91	.84	.88	.84	.70	.77	.77	.69	.63	.68	.68	.70	.97	.86	.92	.92
2	1	−3	6	.2	.9	.46	.57	.55	.54	.80	.58	.58	.80	.42	.37	.37	.47	.55	.60	.28	.54
3	−6	−10	0	.1	.2	.22	.23	.16	.18	.39	.42	.42	.34	.14	.15	.15	.15	.80	.32	.11	.44
4	−4	−10	−1	.2	.9	.25	.35	.28	.32	.64	.45	.45	.65	.23	.21	.21	.25	.55	.39	.13	.42
5	1	−4	3	.7	.6	.37	.46	.36	.42	.49	.54	.54	.47	.31	.21	.21	.34	.68	.49	.31	.61
6	7	5	9	.1	.2	.16	.20	.12	.15	.37	.44	.44	.31	.11	.09	.09	.12	.64	.27	.09	.43
7	7	3	10	.8	.2	.44	.42	.45	.44	.42	.46	.46	.39	.27	.26	.26	.32	.79	.45	.69	.80
8	0	−6	7	.1	.3	.27	.27	.17	.21	.42	.43	.43	.39	.15	.21	.21	.17	.54	.34	.10	.45
9	−5	−7	4	.9	.7	.92	.85	.90	.87	.76	.80	.80	.75	.65	.71	.71	.72	.97	.87	.92	.92
10	2	−8	5	.9	.7	.71	.68	.71	.72	.62	.59	.59	.61	.51	.47	.47	.57	.83	.70	.74	.82
11	−1	−5	10	.9	.7	.95	.85	.90	.86	.74	.88	.88	.74	.65	.71	.71	.71	.98	.86	.92	.92
12	7	1	9	.4	.1	.07	.15	.09	.12	.34	.27	.27	.28	.08	.06	.06	.09	.62	.22	.14	.43
13	4	−11	6	.8	.1	.21	.24	.15	.22	.35	.40	.40	.30	.14	.10	.10	.15	.79	.31	.24	.60
14	−8	−10	9	.6	.1	.36	.46	.51	.47	.53	.53	.53	.50	.38	.26	.26	.41	.82	.56	.77	.84
15	−1	−3	10	.8	.4	.84	.80	.83	.80	.71	.79	.79	.69	.60	.58	.58	.66	.98	.83	.90	.91
16	−1	−4	2	.8	.4	.73	.66	.70	.67	.55	.67	.67	.53	.49	.58	.58	.54	.90	.68	.82	.86
17	0	−3	1	.1	.1	.12	.14	.08	.11	.33	.44	.44	.27	.08	.06	.06	.08	.67	.24	.12	.42
18	1	−9	10	.7	.7	.65	.64	.68	.67	.69	.63	.63	.68	.50	.46	.46	.55	.80	.67	.68	.78
19	1	0	8	.2	.9	.70	.76	.74	.71	.83	.79	.79	.83	.58	.60	.60	.63	.79	.79	.58	.71
20	−6	−9	−5	.5	.9	.25	.40	.31	.37	.63	.44	.44	.63	.27	.19	.19	.29	.56	.43	.20	.48

Note: Med (Safe) are payoffs for the safe alternative, Low (Risky) and High (Risky) are the high and low payoffs for the risky alternative. The models are naïve sampler (Nv), contingent sampler (Ct), and 4-mode contingent sampler (4m). See main text for the definition of the ENO values.

Table 4
Equivalent number of observations (ENO) statistics and regression of observed and predicted choice probabilities for the 4-mode contingent sampler model

Dep. variable: Choice proportion	ENO statistics			Regression statistics					
				Coefficients		p values		R^2	$F(1)$
	S^2	MSE	ENO	Intercept	Predicted probability	Intercept	Predicted probability		
Risky	.0507	.0553	10.50	.101	.773	$< .0001$	$< .0001$	0.949	336.55
Risky after high	.0586	.0624	15.37	.031	.880	.539	$< .0001$	0.886	140.36
Risky after low	.0862	.0981	7.49	.250	0.552	.010	.002	0.436	13.92
Risky after medium	.0599	.0624	23.0	.088	.741	.551	$< .0001$	0.935	260.4
Overall	.0639	.0695	11.27	–	–	–	–	–	–

Each problem is based on five randomly selected values (three outcomes and two transition probabilities). The medium (for the safe alternative), low, and high (for the risky alternative) outcomes were randomly drawn from a uniform distribution with values from -10 to 10 and the following restriction: $-10 \leq \text{low} \leq \text{medium} \leq \text{high} \leq 10$. The two transition probabilities p and q were determined by sampling their values from a uniform set $\{.1, .2, .3, \ldots, .9\}$. Table 3 shows the 20 problems.

Notice that the exact effect of sensitivity on the dynamic nature of the environment is likely to vary among the different problems. To address this difficulty the current analysis uses the models presented above. It tests the generality of these models with Busemeyer and Wang's (2000) generalization criteria. That is, the predictions of the models were derived before running Experiment 2 based on the parameters estimated above.

4.1. Apparatus and procedure

The apparatus and procedure were identical to that used in Experiment 1, with the exception that each participant faced 10 different problems, with 100 trials per problem. The conversion rate in this experiment was 1 a (about .22 U.S. cents) per point.

4.2. Participants

Forty Technion students served as paid participants in the experiment. Twenty participants were faced with Problems 1–10, and 20 were presented with problems 11–20. The participants received a fee of 20 shekels for attending ($4.50), from which they could win or lose money according to their obtained payoffs in the experimental task. Final payoffs ranged between 12 and 28 shekels (about $2.70–$6.30).

4.3. Results

Table 3 presents the observed (unconditional and conditional) proportions of risky choices and the predictions of the three models (with the parameters estimated as above based on Table 2's data). The first line in the lower panel presents the mean squared error (MSE) of the different predictions.[9] Note that these predictions were generated with the parameters that were fitted to Experiment 1.

To clarify the implication of the observed MSE values we have translated them to the equivalent number of observations (ENO) of the models (see Erev et al. (in press, 2007)). In short, the ENO of a model is an estimation of the minimal size of experiment that has to be run to obtain predictions that are more accurate (have a smaller squared error) than the model's prediction. For example, if a model has an ENO of 10, its prediction of the probability of risky choice in a particular problem is expected to be as accurate as a prediction that is based on the observed proportion of risky choices in an experimental study of that problem with 10 participants.[10]

The third row in Table 3 presents the estimated ENO. It shows that all three models provide useful predictions of the unconditional choice proportions. The ENOs for this statistic are above 10 for all three models.

Examination of the conditional choice proportions reveals a large advantage of the 4-mode contingent sampler model. This model outperforms the other models in all three cases. In addition, the results suggest the 4-mode model is less accurate in predicting risky choice after low payoff. The ENO in this case is around 7, while the ENOs of the other predictions of this model are above 10. Examination of this issue reveals that the model tends to under-predict the probability of repeated risky choice after high payoff. One explanation of this bias involves the assertion that our estimate of the WSLS parameter (ω) is too low.

To improve our understanding of the joint effect of inertia, recency, and contingent sampling we considered a fourth model that assumes naïve sampling with inertia and recency. This "naïve 4-mode" model was also clearly outperformed by the contingent 4-mode model.

Fig. 6(a) and (b) presents the observed results and the predictions of the contingent 4-mode model graphically. The curves are fairly flat and demonstrate that the accuracy of the predictions is relatively stable over time.

To further examine how the 4-mode model predicts unconditional and conditional risky choices across all 20 problems, we regressed the observed choice probabilities on the predicted choices and on an intercept. The intercept was introduced to investigate if the model systematically over- or underestimates choice probabilities. Table 4 depicts the results of the regressions and shows a good general agreement between model and participants that is reflected in the high R^2 values. The small but significant coefficient for the intercept when predicting $P(\text{risky})$ and $P(\text{risky after low})$ suggests further that the model underestimates the probability of choosing risky and risky after low for at least some problems. The extent to which the probability of choosing risky after low was underestimated was largest in problems 2, 4 and 20. In all three problems the strategy WSLS impaired expected return. Under one explanation of this bias, the participants exhibited stronger WSLS tendency than predicted by the model. Under this explanation, the model's bias after high outcomes (in these problems) was weaker because inertia and WSLS make the same prediction after high payoffs.

To examine participants' sensitivity to the problems' transition probabilities we regressed the predicted and observed choices after high and low outcomes from the risky option on the

[9] MSE is the mean squared distance between the models' predicted proportions and the choice proportions of the individual participants. MSE is highly correlated with the MSD score used in Experiment 1 and is preferred here because it allows the derivation of the ENO statistic described below.

[10] This score is computed as $\text{ENO} = S^2/(\text{MSE}-S^2)$ where S^2 is the pooled variance over problems, and MSE is the mean squared distance between the model and each of the participants. That is, S^2 is the estimate of the MSE score of an accurate model, and ENO increases when the MSE decreases toward its minimal value.

Fig. 6. The observed and predicted results in the 20 problems studied in Experiment 2 using Fig. 2's format. Each block summarizes 25 trials. See caption for Fig. 5 for a more detailed explanation of abbreviations and the detection of recency and inertia effects.

Table 5
Regression of risky after high on p and of risky after low on q, for observed and predicted choices of the 4-mode contingent sampler model

Dep. variable: Choice proportions		Coefficients		p values		R^2	$F(1)$
		Intercept	p/q	Intercept	Predicted probability		
Observed	Risky after high	.497	.381	$< .0001$	$< .0001$	.862	52.01
	Risky after low	.355	.431	$< .0001$	$< .0001$	.872	122.71
Predicted	Risky after high	.575	.350	$< .0001$	$< .0001$	.546	21.67
	Risky after low	.436	.271	$< .0001$	.028	.241	5.71

respective problem parameters p and q. Table 5 shows the results. The first observation is that in all cases the problem parameters and an intercept significantly contributed to the prediction of choices. Interestingly, the regression results for risky after high are very similar for the model and for participants, whereas they differ clearly for risky after low. Specifically, participants were more sensitive to changes in the problem parameter q than was the 4-mode contingent sampler model. This might be due to the way inertia is implemented in the model, which assumes a minimum choice proportion for risky after low, independent of problem parameters.

Our favorite interpretation of the current results is based on the assertion that the contingencies assumed by the contingent sampler models are approximations of the actual contingencies. Human decision makers learned the "relevant contingencies" during the experiment, and the learned contingencies are similar (but not identical) to the contingencies assumed by the model. The relatively high ENO of the 4-mode contingent sampler model suggests that the approximation is useful. Nevertheless, it is clear that our understanding of the processes that affect a decision maker's sensitivity to relevant contingencies is limited. We hope to address these processes in future research.