In Proceedings of ICCM - 2007- Eighth International Conference on Cognitive Modeling. 67 - 72. Oxford, UK: Taylor & Francis/Psychology Press.

Instance-Based Decision Making Model of Repeated Binary Choice

Christian Lebiere (cl@cmu.edu)
Psychology Department
Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Cleotilde Gonzalez (coty@cmu.edu) and Michael Martin (mkmartin@andrew.cmu.edu)
Dynamic Decision Making Laboratory,
Social and Decision Sciences Department
Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

We describe an instance-based model of decision-making for repeated binary choice. The model provides an accurate account of existing data of aggregate choice probabilities and individual differences, as well as newly collected data on learning and choice interdependency. In particular, the model provides a general emergent account of the risk aversion effect that does not require any metacognitive assumptions. Advantages of the model include its simplicity, its compatibility with previous models of choice and dynamic control, and the strong constraints it inherits from the underlying cognitive architecture.

Keywords: Learning; dynamic decision making; RELACS; memory; cognitive architectures; ACT-R.

Introduction

Erev and Barron (2005) have discussed the tradeoffs of adaptation and maximization in repeated choice tasks. A main demonstration from their studies is that extended practice with a binary choice problem with immediate feedback does not always lead to payoff maximization.

The deviations from maximization may be due to different effects. One of them, the *payoff variability effect*, refers to a tendency to increase exploration in a noisy environment (Erev & Barron, 2005). That is, when payoff variability is associated with an alternative of higher expected value compared to the other alternative, choice behavior moves toward random choice. This payoff variability effect has been found in one-shot decisions (Busemeyer & Townsend, 1993) but it is more robust in repeated choice (Erev & Barron, 2005).

Erev and Barron (2005) proposed a model of Reinforcement Learning Among Cognitive Strategies (RELACS) to account for the payoff variability effect and other deviations from maximization. RELACS assumes that a decision maker follows one of three cognitive strategies in each choice, and that the probability of using a strategy is determined by previous experiences with the strategy.

The *fast best reply* strategy involves selecting the alternative with the highest recent payoff. The *case-based reasoning* strategy involves moving from a random selection of alternatives initially to a two-stage process in which a belief is first determined and then verified as not being associated with large losses. The *slow best reply* strategy involves choosing to explore the two alternatives initially and moving gradually toward preferring the alternative more likely to maximize earnings. According to RELACS, the three strategies are reinforced with their frequency of use and are updated according to the observed payoffs.

In their analyses and comparisons to other models, Erev and Barron determine that the slow best reply strategy is the one that best captures the payoff variability effect. They also found that the assumption of learning among the different strategies is not important because a random selection among strategies fits the data as well as RELACS does.

In our past research we have proposed a framework and computational model that characterize decision makers' preferences and utilities in terms of action-outcome links. This theory called Instance-Based Learning Theory (IBLT) (Gonzalez, Lerch & Lebiere, 2003), implemented in ACT-R (Anderson & Lebiere, 1998; Anderson et al, 2004), proposes learning (i.e., increasing maximization) occurs through a progressive accumulation of *decision instances*. Instances are discrete units of knowledge (action-outcome links) which are constructed, upgraded, and reused through experiential learning in a decision making situation. Better decision policies emerge gradually as decision makers move from using explicit rules of action to implicit recognition of familiar patterns (cf. Dienes & Fahey, 1995), similar to the gradual process proposed in Logan's (1988) instance theory of automaticity. Many decision making tasks have successfully been implemented in ACT-R using this process, including dynamic control tasks (Wallach & Lebiere, 2003), supply chain management (Gonzalez & Lebiere, 2005; Martin, Gonzalez & Lebiere, 2004), backgammon (Sanner et al, 2000) and simple 2x2 games like the Prisoner's Dilemma (Lebiere, Wallach & West, 2000).

Our main contention in this paper is that the experiential accumulation, activation, retrieval and generalization of action-outcome decision instances is a general decision making strategy applicable to multiple decision making tasks, including the simple repeated choice effects posed by Erev and Barron (2005). Accordingly we describe an instance-based decision making model that captures the

learning effects and the tradeoffs of adaptation and maximization reported by Erev and Baron (2005). Our instance-based decision making model works in ways similar to the slow best reply strategy proposed by Erev and Barron (2005). The results from our ACT-R model support Erev and Barron's arguments that the slow best reply strategy is the one that best captures the payoff variability effect and that learning among different cognitive strategies is unnecessary. Thus, deviations from maximization in repeated binary choice problems can be reproduced without pre-defining a set of cognitive strategies and positing reinforcement learning as a mechanism for selecting among them.

In what follows, we discuss the example problems we have taken from Erev and Barron (2005), and discuss how we replicated their behavioral results. Next, we discuss our instance-based decision making ACT-R model and the results from our model as compared to RELACS results. Finally, we discuss some predictions of our model and possibilities for unification with models of other tasks.

The Payoff Variability Effect

We replicated, with human participants, the payoff variability effect using the following three key problems from Erev and Barron (2005):

Problem 1.	H	11 points with certainty
	L	10 points with certainty
Problem 2.	H	11 points with certainty
	L	19 points with probability 0.5
		1 otherwise
Problem 3.	H	21 points with probability 0.5
		1 otherwise
	L	10 points with certainty

All three problems required participants to choose between a high payoff alternative H (with an expected value of 11 points) and a low payoff alternative L (with an expected value of 10 points). The problems differed only on the variance but not the mean of the two payoff distributions.

We randomly assigned 60 participants to one of the three problems. The undergraduate and graduate students at Carnegie Mellon University were paid a flat fee for performing the repeated choice task for 400 trials.

We followed almost identical instructions as in Erev and Barron's experiments: individuals did not receive any information about the payoff structure. They were told their task was to select one of the alternatives by clicking on one of two unmarked and masked buttons. They were provided with the payoff value of the button they clicked on. Individuals were not informed of the trial number. Payoffs were drawn from the distribution associated with the selected button.

There are two differences between our methods and Erev and Barron's: (1) we did not use a performance-based incentive structure and (2) we ran 400 rather than 200 trials to better explore learning effects.

Figure 1 shows the proportion of maximization (Pmax) (H) choices during the 400 trials. The average proportions of maximization are very similar to those reported in the original experiments: average Pmax for the second 100-problem block (a.k.a. Pmax2) was 0.82, 0.61 and 0.50 for Problem 1, 2, and 3 respectively (compared to .90, .71, .57 in Erev and Barron's data).

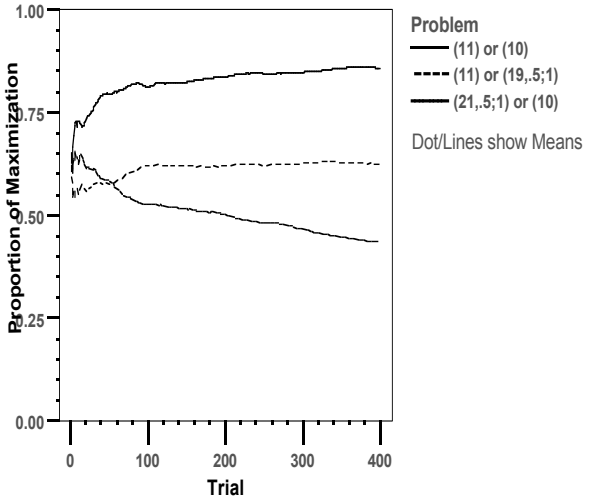


Figure 1: Proportion of maximization over practice

The learning curves shown in Figure 1 demonstrate that, as expected, an increase in payoff variability impairs maximization. In contrast to data reported by Erev and Barron (2005) the Problem 3 learning curve, where the alternative with the maximum payoff is risky, shows a *decrease* in the proportion of maximization over time.

As suggested by Erev and Barron, the difference between problems 1 and 3 demonstrates the risk aversion effect and the difference between problems 1 and 2 the risk seeking effect. We investigated the risk aversion or certainty effect (Kahneman & Tversky, 1979) in this repeated choice task by collecting data in the following problem:

Problem 4.	Certain	11 points with certainty
	Risky	21 points with probability 0.5
		1 otherwise

Problem 4 presents participants with a tie, i.e. alternatives have the same expected value, but one is risky and the other is certain. Using the same methods as in the first 3 problems, we collected data from 20 participants. We will report detailed findings on that condition in the model comparison section.

One of the challenges for Erev and Barron's RELACS model is that it consistently underpredicts individual differences. A particular problem in RELACS seems to be the management of memory, as it does not capture the interdependency of past experiences.

Memory management and learning is a strength in ACT-R and a particular strength of our instance-based decision-making models as there are strong constraints on the effect that particular past instances would have on a future choice (Gonzalez & Quesada, 2004). As we will demonstrate, our ACT-R based models of instance-based decision making predict observed individual differences quite accurately.

ACT-R instance-based decision making model

One advantage of instance-based learning models is that they reduce degrees of freedom in modeling. The modeler does not have to select and implement strategies, or decide upon arbitrary criteria on which a decision is made. Instead, the model represents the information immediately available to the subject in the most direct form possible, and uses that information directly to make its decisions.

Each decision-making instance in the repeated choice paradigm is composed of two elements: the choice being made and the payoff immediately received as a result. Those two elements of a decision-making instance are consciously available to the subject and thus will be represented together in declarative form.

The basic unit of declarative representation in ACT-R is the chunk. A chunk is a typed structure composed of a number of named fields, also called slots. Each slot usually contains another chunk (although it can also be empty or contain special values). Our model contains only a single chunk type, **choice**, with only two slots: **decision**, which holds the decision made by the model, and **payoff**, which holds the payoff awarded after the decision. For example, a chunk encoding the experience that pressing the left button resulted in a payoff of 10 would have the following form:

```
Decision1
     isa decision
     choice Left
     payoff 10
```

That chunk type serves both as the only type of goal for the model, and as the repository of the problem-solving experience in long-term declarative memory. The learning of that symbolic information is thus automatically accomplished by the architecture as it stores past goals into long-term memory.

The experimental paradigm covered by Erev and Barron (2005) includes three feedback conditions. In the first one called *minimal information*, payoff feedback is given only for the choice being made, and encoded as described above. In the second condition, called *complete feedback*, payoff is given for the choice made as before, but the payoff that would have resulted if the other choice had been made is also given. In that case, the model generates two chunks, one for each potential choice and its feedback. In the third condition, called *probability learning*, no numerical payoff feedback is given directly but instead the payoff is translated into a relative probability of correct choice, which is then relayed to the subject as a correct/incorrect binary feedback. In the model, that binary feedback is simply encoded as a 0/1 payoff and the same modeling approach can then apply.

How does the model use this information about choices and payoffs? The basic decision-making procedure is the same as that used in the model prisoner's dilemma and other 2x2 games (Lebiere, Wallach & West, 2000). The model evaluates each option by retrieving its expected payoff from memory, selects the one with the highest value, then registers the feedback as described above. This procedure is implemented in half-a-dozen generic production rules.

As in some past instance-based models (e.g. the Paper Rocks Scissors model of West & Lebiere, 2001), the possible combinations of symbolic information are so few (less than a handful in the payoff functions studied here) that the key knowledge of the task does not reside at the symbolic level but instead in its statistical properties. Specifically, the key information is the frequency (and recency) of each combination of decision and payoff. While subjects (and the model) could potentially keep track of those frequencies explicitly, there is no evidence that they do so. Instead, the architecture automatically learns such information in the activation values of the various chunks. Specifically, the base-level activation A_i of chunk i is determined by the following Bayesian learning formula:

$$A_i = \ln \sum_{j=1}^{n} t_j^{-d} \qquad \textit{Base Level Learning}$$

Each t_j is the lag of time since the jth occurrence of chunk i. The architectural parameter d is the decay rate of each occurrence, which is set to 0.5 as is (almost) always the case in ACT-R models. The power law of practice emerges from the log-summation over all references whereas the power law of forgetting results from the decay of each reference. Just as for chunks, this learning of the statistical properties of the symbolic knowledge is accomplished automatically by the architecture. Activation determines the probability of retrieving each qualifying chunk according to the following equation:

$$P_i = \frac{e^{A_i/t}}{\sum_j e^{A_j/t}} \qquad \textit{Boltzmann Equation}$$

This equation, also known as the softmax equation, defines retrieval as a noisy process where the probability of retrieving a given chunk is proportional to the ratio of its activation and a retrieval noise level t. The noise level determines the degree of stochasticity of the retrieval process and similar to the decay rate parameter it is left at its default value of 0.25.

However, the retrieval process described above has one problem. If it only retrieves one chunk associated with a given choice, it will usually not be sensitive to the magnitude of the payoff values. If one alternative has a

certain payoff of 11, it will not matter whether the other has equally likely payoffs of 1 and 19 (averaging 10) or 1 and 199 (averaging 100). It will choose each about half the time in both cases, which is clearly not right. What we want is a retrieval procedure that takes into account both the frequency (and recency) of each payoff as reflected in its activation and the magnitude of the payoff itself. To that effect, Lebiere (1999) introduced a variation of the retrieval process called *blending* that has since been used in many instance-based models (e.g. Gonzalez et al., 2003; Wallach & Lebiere, 2003). The key equation controlling blended retrieval is the following:

$$V = \min \sum_i P_i \cdot \left(1 - Sim\left(V, V_i\right)\right)^2 \quad \textbf{\textit{Blending Equation}}$$

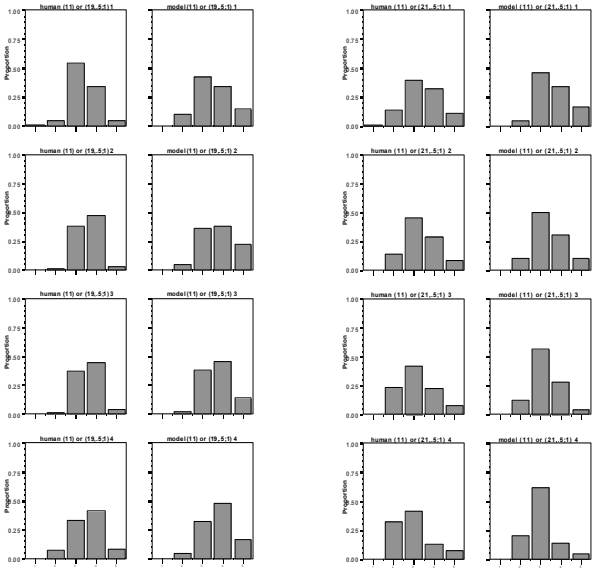
The equation states that the value V returned by retrieval is the one that best satisfies the constraints offered by all matching chunks i weighted by their probability of retrieval P_i as computed in the Boltzmann equation above. Satisfying chunk constraints is defined in terms of minimizing the dissimilarity (i.e. maximizing the similarity) between the consensus answer V and the actual answer V_i contained in chunk i. This process is applicable to all domains, discrete and continuous, as long as a similarity metric is defined over those values. As such it can be seen as an implementation of the generalized Bayesian framework of Tenenbaum & Griffiths (2001) or an approximation of the generalization capabilities of connectionist architectures based on distributed representations (e.g. O'Reilly & Munakata, 2000). In practice, we define linear similarity values over payoffs, which result in the retrieval process averaging their values weighted by activation.

A final point concerns the initialization of the model. If the model started with no expectations of the payoffs, it would start by deciding randomly, but then as soon as one payoff had been experienced for each choice, it would happily take the best indefinitely. To trigger exploration at the start, we initialized the model with a single chunk for each decision encoding high initial expectations (payoff of 1000). That initial value will quickly get overwhelmed by actual experience as it decays and is never reinforced, but it results in an initial period of exploration that corresponds well to human subjects without the need to arbitrarily define a specific strategy to that effect.

Results and Comparison

Our model fits the data quite well using Erev and Barron's (2005) primary measure of performance, namely the probability of maximization in the second block of 100 problems (Pmax2).. That measure for problems 1, 2 and 3 is 0.91, 0.65 and 0.53 respectively for our model, as compared to 0.90, 0.71 and 0.57 for Erev and Barron's data and 0.82, 0.61 and 0.50 for our data. The variation between data and model is substantially smaller than the variations between data sets, suggesting the substantial role played by individual differences.

To examine individual differences, we have plotted in Figure 2 the distribution of the probability of maximization for each 100-trial block for individual subjects (and model runs) within each of five intervals: 0-20%, 20-40%, 40-60%, 60-80% and 80-100%. For reasons of space, we have selected problems 2 (Figure 2a) and 4 (Figure 2b) as the most interesting for display. Focusing for now on problem 2, one can see that the distribution of probabilities ranges across the highest 4 categories, a range well-reproduced by our model. One could argue that it is too well reproduced, with the highest category over-represented compared to the data. However, in Erev & Barron's data (which only report distribution figures for the second block), the two highest categories (60-80% and 80-100%) dominate with many fewer subjects in the 40-60% category than for our data. This would seem to explain the discrepancy between the values of Pmax2 observed by Erev and Barron and us (0.71 vs. 0.61). In this respect, our model fits comfortably between the two data sets, but it is again a reminder to be careful when comparing to aggregate data across subjects.



2a 2b

Figure 2: Individual differences for problems 2 (left) and problem 4 (right). Within each panel human data (left) and model data (right) distribution of maximization probability are shown, aggregated in 20% increments.

We now consider how deviations from maximization emerge from our model and in particular the source of the payoff variability effect. For problem 1, with deterministic payoffs, maximization is simply a matter of quickly overcoming the high initial expectations through the exploration phase. Alternative H consistently returns the highest payoff. Thus blending consistently produces a higher expected value for alternative H. For problem 2, with variable payoffs for alternative L, blending combines

the distinct payoffs of 1 and 19 for alternative L to produce random fluctuations in expected value. Although blending will tend to average the two payoffs of alternative L to 10 if they are of equal activation, the noise of the activation process and the random distribution of L payoffs will tend to make activations unequal, pushing the average on either side of 10, and sometimes higher than 11, which leads to lapses in maximization. The same happens with Problem 3, except that alternative H averaging 11 is now the one with the noisy distribution, and the model does reproduce the tendency to select it less frequently than it does in Problem 2, indicating risk aversion. This brings us to problem 4, the risk aversion problem, where this symmetry argument would suggest that both options would be chosen equally often on average. However, on average both subjects and model tend to prefer the certain option, roughly 55% of the time. This risk aversion effect (and the difference in Pmax2 between Problems 2 and 3) arises from a subtle interaction in the dynamics of the task illustrated in Figure 3.

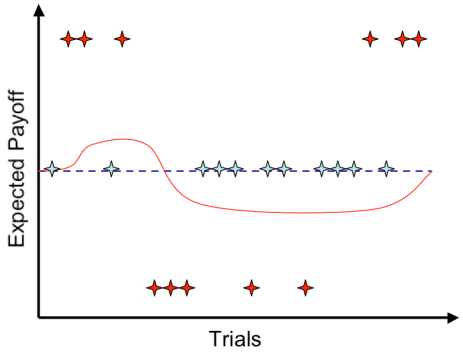


Figure 3: Emergence of risk aversion effect

The blue dashed line represents the (constant) expected value of the certain alternative while the red line represents the expected value of the risky alternative. On average, the expected values of the two alternatives are equal and they indeed start that way. Each star of a given color (red or blue) indicates an experienced payoff for the associated choice. After the start, the risky alternative provides some lucky payoffs (e.g. 21), which raises its expected value and leads to its selection more often. Luck even outs quickly however as a series of poor payoffs (e.g., 1) lowers its expected value to less than 11, which in turn leads to selection of the certain alternative most of the time. The key insight is that this bias toward certain payoffs leaves the risky alternative fewer opportunities to bring its average back to the level of the certain alternative, meaning that this interval where the certain alternative is selected most of the time is longer than the previous interval when the risky alternative was selected most often. This asymmetry is the source of the risk aversion effect in our model and its preference for certainty.

One prediction of this explanation arises from its origin in the base-level learning equation that reflects the occurrence of events into the activation of decision chunks and then into the expected outcomes of the respective choices. As experience accumulates, the impact of recent events in activation fluctuations will be gradually overcome by the increasingly long history. One would therefore expect risk aversion to disappear with practice, a prediction confirmed by Figure 4, which plots the probability of choosing the certain alternative with practice (in terms of blocks of 10 trials). In the initial exploration period, both model and subjects choose the certain alternative about 50% of the time. By around trial 50 the certain alternative is chosen over 60% of the time as the payoffs statistics are quickly learned, but the bias to select certain payoffs then gradually declines back to 50% as the increasingly long history overcomes short-term fluctuations. Figure 2b (right) illustrates this learning process across blocks of 100 trials as a quickly learned propensity to choose the certain alternative gradually reverts to the mean.

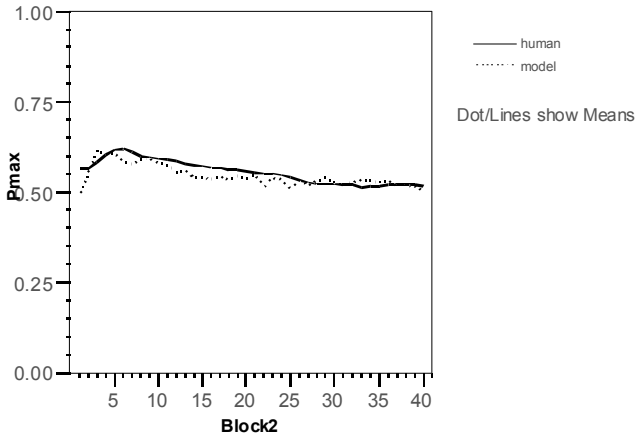


Figure 4: Time course of risk aversion effect

As we mentioned previously, one strong aspect of our model over RELACS is that it makes constrained predictions about the probability of making a given decision as a function of the recent history of choices and payoff outcomes. To study those probabilities, we used a methodology called model-tracing (Anderson et al, 1995) to force the model to make the same decisions as each human subject, thereby giving them the same context in which to make each decision. We can then directly compare each decision for model and subjects, as reported in Table 1. Columns 2 and 3 report the Pmax values for each subject and the model tracing its decisions. Columns 4 and 5 report the minimum and maximum probability of matching decisions given those base probabilities. Column 6 report the average prediction probability of agreement assuming that decisions are randomly distributed given those base probabilities while column 7 reports the actual probability of agreement. For all subjects but S8, the actual probability is higher than the predicted probability, establishing that the model is capturing some of the short-term factors used by the subjects in their decisions.