DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

A Choice Prediction Competition: Choices From Experience and From Description

Citation
Erev, Ido, Eyal Ert, Alvin E. Roth, Ernan Haruvy, Stefan Herzog, Robin Hau, Ralph Hertwig, Terrence Stewart, Robert West, and Christian Lebiere. 2009. A choice prediction competition: choices from experience and from description. Journal of Behavioral Decision Making 23(1): 15-47.

Published Version
doi:10.1002/bdm.683

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:5343169

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Forthcoming, Journal of Behavioral Decision Making.

A choice prediction competition, for choices from experience and from description

Ido Erev -- Technion

Eyal Ert and Alvin E. Roth -- Harvard University

Ernan Haruvy -- University of Texas at Dallas

Stefan Herzog, Robin Hau, and Ralph Hertwig -- University of Basel

Terrence Stewart -- University of Waterloo, Robert West -- Carleton University, and

Christian Lebiere -- Carnegie Mellon University

September 1, 2009

Abstract: Erev, Ert, and Roth organized three choice prediction competitions focused on three related choice tasks: one shot decisions from description (decisions under risk), one shot decisions from experience, and repeated decisions from experience. Each competition was based on two experimental datasets: An estimation dataset, and a competition dataset. The studies that generated the two datasets used the same methods and subject pool, and examined decision problems randomly selected from the same distribution. After collecting the experimental data to be used for estimation, the organizers posted them on the Web, together with their fit with several baseline models, and challenged other researchers to compete to predict the results of the second (competition) set of experimental sessions. Fourteen teams responded to the challenge: the last seven authors of this paper are members of the winning teams. The results highlight the robustness of the difference between decisions from description and decisions from experience. The best predictions of decisions from descriptions were obtained with a stochastic variant of prospect theory assuming that the sensitivity to the weighted values decreases with the distance between the cumulative payoff functions. The best predictions of decisions from experience were obtained with models that assume reliance on small samples. Merits and limitations of the competition method are discussed.

Keywords: Fitting; generalization criteria; prospect theory; reinforcement learning; explorative sampler, equivalent number of observations (ENO), ACT-R, the 1-800 critique.

Correspondence to: Ido Erev, Max Wertheimer Minerva Center for Cognitive Studies, Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel.
E-mail: erev@tx.technion.ac.il. We thank the three editors, Frank Yates, Tim Rakow, and Ben Newell, and the reviewers for extremely useful suggestions. Part of this research was conducted when Ido Erev was a Marvin Bower Fellow at the Harvard Business School. Ralph Hertwig and Robin Hau were supported by Swiss National Science Foundation Grant 100014-118283.
Competition website: http://tx.technion.ac.il/~erev/Comp/Comp.html.

A major focus of mainstream behavioral decision research has been on finding and studying counter-examples to rational decision theory, and specifically examples in which expected utility theory can be shown to make a false prediction. This has led to a concentration of attention on situations in which utility theory makes a clear, falsifiable prediction; hence situations in which all outcomes and their probabilities are precisely described, so that there is no room for ambiguity about subjects' beliefs. Alternative theories, such as prospect theory (Kahneman & Tversky, 1979), have been formulated to explain and generalize the deviations from utility theory observed in this way.

The focus on counterexamples and their explanations has many attractive features. It has led to important observations, and theoretical insights. Nevertheless, behavioral decision research may benefit from broadening this focus. The main goal of the current research is to facilitate and explore one such direction: The study of *quantitative predictions*. We share a certain hesitation about proceeding to quantitative predictions prematurely, before the groundwork has been laid for a deep understanding that could motivate fundamental models. But our interest comes in part from the observation that the quest for accurate quantitative predictions can often be an inspiration for precise theory. Indeed, it appears that many important scientific discoveries were triggered by an initial documentation of quantitative regularities that allow useful predictions.[1]

A second motivation for the present study comes from the "1-800 critique" of behavioral research. According to this critique, the description of many popular models, and of the conditions under which they are expected to apply, is not clear. Thus, the authors who publish these models should add 1-800 toll free phone numbers and be ready to help potential users in deriving the predictions of their models. The significance of the 1-800 problem is clarified by a comparison of exams used to evaluate college students in

[1] One of the earlier examples is the Pythagorean theorem. Archeological evidence suggests that the underlying regularity (the useful quantitative predictions) were known and used in Babylon 1300 years before Pythagoras (Neugebauer & Sachs, 1945). Pythagoras' main contribution was the clarification of the theoretical explanation of this rule and its implications. Another important example is provided by Kepler's laws. As suggested by Klahr and Simon (1999) it seems that these laws were discovered based on data mining techniques. The major theoretical insights were provided by Newton, almost 100 years after Kepler's contributions. A similar sequence characterizes one of the earliest and most important discoveries in Psychology. Weber's law was discovered before Fechner provided an elegant theoretical explanation of this quantitative regularity. These successes of research that starts with a focus on quantitative regularities suggest that a similar approach can be useful in behavioral decision research too.

the exact and behavioral sciences. Typical questions in the exact sciences ask the examinees to predict the outcome of a particular experiment, while typical questions in the behavioral sciences ask the examinees to exhibit understanding of a particular theoretical construct (see Erev & Livne-Tarandach's, 2005 analysis of the GRE exams). This gap appears to reflect the belief that the leading models of human behavior do not lead to clear predictions. A more careful study of quantitative predictions may help change this situation.

A third motivating observation comes from the discovery of important boundaries of the behavioral tendencies that best explain famous counterexamples. For example, one of the most important contributions of prospect theory (Kahneman & Tversky, 1979) is the demonstration that two of the best-known counterexamples to expected utility theory, the Allais paradox (Allais, 1953) and the observation that people buy lotteries but also insurance (Friedman & Savage, 1948), can be a product of a tendency to overweight rare events. While this tendency is robust, it is not general. The recent studies of decisions from experience demonstrate that in many settings people exhibit the opposite bias: They behave as if they underweight rare events (see Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004; Hau, Pleskac, Kiefer, & Hertwig, 2008; Erev, Glozman, & Hertwig, 2008; Rakow, Demes, & Newell, 2008; Ungemach, Chater & Stewart, 2009). A focus on quantitative predictions may help identify the boundaries of the different tendencies.

Finally, moving away from a focus on choices that provide counterexamples to expected utility theory invites the study of situations in which expected utility theory may not provide clear predictions. There are many interesting environments that fall into this category, including decisions from experience. The reason is that, when participants are free to form their own beliefs based on their experience, almost any decisions can be consistent with utility theory under certain assumptions concerning these beliefs.

The present competition (which is of course a *collaboration* among many researchers) is designed in part to address the fact that evaluating quantitative predictions offers individual researchers different incentives than those for finding counterexamples to expected utility theory. The best presentations of counterexamples typically start with the presentation of a few interesting phenomena, and conclude with the presentation of an

elegant and insightful model to explain them. The evaluation of quantitative predictions, on the other hand, tends to focus on many examples of a choice task. The researcher then has to estimate models, and run another large (random sample) study to compare the different models. In addition, readers of papers on quantitative prediction might be worried that the probability a particular paper will be written increases if it supports the model proposed by the authors.

To address this problematic incentive structure, the current research uses a choice prediction competition that can reduce the cost per investigator, and can increase the probability of insightful outcomes. The first three authors of the paper (Erev, Ert, & Roth, hereafter EER) organized three choice prediction competitions. They ran the necessary costly studies of randomly selected problems, and challenged other researchers to predict the results.[2] One competition focused on predicting decisions from description, and two competitions focused on predicting decisions from experience. The participants' goal in each of the competitions was to predict the results of a specific experiment.

Notice that this design extends the classical study of counterexamples along two dimensions. The first dimension is the parameters of the choice problems (the possible outcomes and their probabilities). The current focus on randomly selected parameters is expected to facilitate the evaluation of the robustness of the relevant tendencies. The second dimension is the source of the information available to the decision makers (description or experience). The comparison of the different sources and the different models that best fit behavior in the different conditions was expected to shed light on the gap between decisions from description and decisions from experience. It could be that the differences in observed behavior are more like differences in degree than differences in kind, and that both kinds of behavior might be predicted best by similar models, with different parameters. Or, it could be that decisions from description will be predicted best by very different sorts of models than those that predict decisions from experience well,

[2] A similar approach was taken by Arifovic, McKelvey, and Pevnitskaya (2006) and Lebiere & Bothell (2004) who organized Turing tournaments. Arifovic et al. challenged participants to submit models that emulate human behavior (in 2-person games) and sniffers (models that try to distinguish between human and emulators). The models were ranked based on an interaction between the two types of submissions. As explained below, the current competitions are simpler: The sniffers are replaced with a pre-determined criterion to rank models. Note that to the extent that competitions ameliorate counterincentives to conducting certain kinds of research, they can be viewed as a solution to a market design problem (Roth, 2008).

in which case the differences between the models may suggest ways in which the differences in behavior may be further explored.

1. Methods

The current research involved three related but independent choice prediction competitions. All three competitions focused on the prediction of binary choices between a safe prospect that provides a Medium payoff (referred to as M) with certainty, and a risky prospect that yields a High payoff (H) with probability Ph, and a Low payoff (L) otherwise. Thus, the basic choice problem is:

Safe: M with certainty

Risky: H with probability Ph; L otherwise (with probability 1-Ph)

Table 1a presents 60 problems of this type that will be considered below. Each of the three competitions focused on a distinct experimental condition, with the object being to predict the behavior of the experimental subjects in that condition. In Condition "Description," the participants in the experiment were asked to make a single choice based on a description of the prospects (as in the decisions under risk paradigm considered by Kahneman & Tversky, 1979). In Condition "Experience-Sampling" (E-Sampling) subjects made one-shot decisions from experience (as in Hertwig et al., 2004), and in Condition "Experience-Repeated" (E-Repeated) subjects made repeated decisions from experience (as in Barron & Erev, 2003).

<Insert Table 1>

The three competitions were each based on the data from two experimental sessions, an estimation session, and a competition session. The two sessions for each condition used the same method and examine similar, but not identical, decision problems and decision makers as described below. The estimation sessions were run in March 2008. After the completion of these experimental sessions EER posted the data (described in Table 1a) on the Web (see EER, 2008) and challenged researchers to participate in three competitions that focused on the prediction of the data of the second

(competition) sessions.[3] The call to participate in the competition was published in the *Journal of Behavioral Decision Making* and in the e-mail lists of the leading scientific organizations that focus on decision-making and behavioral economics. The competition was open to all; there were no prior requirements. The predictions submission deadline was September 1st 2008. The competition sessions were run in May 2008, but we did not look at the results until September 2nd 2008.

Researchers participating in the competitions were allowed to study the results of the estimation study. Their goal was to develop a model that would predict the results of the competition study. The model had to be implemented in a computer program that reads the payoff distributions of the relevant gambles as an input and predicts the proportion of risky choices as an output. Thus, the competitions used the generalization criterion methodology (see Busemeyer & Wang, 2000).[4]

1.1 The problem selection algorithm.

Each study focused on 60 problems. The exact problems were determined with a random selection of the parameters (prizes and probabilities) L, M, H and Ph using the algorithm described in Appendix 1. Notice that the algorithm generates a random distribution of problems such that about 1/3 of the problems involve rare (low probability) High outcomes (Ph < .1), and about 1/3 involve rare Low outcomes (Ph > .9). In addition 1/3 of the problems are in the gain domain (all outcomes are positive), 1/3 are in the loss domain (all outcomes are negative), and the rest are mixed problems (at least one positive and one negative outcome). The medium prize M is chosen from a distribution with a mean equal to the expected value of the risky lottery.

Table 1a presents the 60 problems that were selected for the estimation study. The same algorithm was used to select the 60 problems in the competition study. Thus, the two studies focused on choice problems that were randomly sampled from the same space of problems.

---

[3] The main prize for the winners was an invitation to co-author the current manuscript; the last seven co-authors are the members of the three winning teams.

[4] This constraint implies that the submissions could not use any information concerning the observed behavior in the competition set. Specifically, each model was submitted with fixed parameters that were used to predict the data of the competition set.

1.2 The estimation study

One hundred and sixty Technion students participated in the estimation study. Participants were paid 40 Sheqels ($11.40) for showing up, and could earn more money or lose part of the show-up fee during the experiment. Each participant was randomly assigned to one of the three experimental conditions.

Each participant was seated in front of a personal computer and was presented with a sequence of choice tasks. The exact tasks depended on the experimental condition as explained below. The procedure lasted about 40 minutes on average in all three conditions.

The payoffs on the experimental screen in all conditions referred to Israeli Sheqels. At the end of the experiment one choice was randomly selected and the participant's payoff for this choice determined his/her final payoff.

The 60 choice problems listed in Table 1a (the estimation set) were studied under all three conditions. The main difference between the three conditions was the information source (description, sampling or feedback). But the manipulation of this factor necessitated other differences as well (because the choice from experience conditions are more time consuming). The specific experimental methods in each of the three conditions are described below:

Condition Description (One-shot decisions under risk):

Twenty Technion students were assigned to this condition. Each participant was seated in front of a personal computer screen and was then presented with the prizes and probabilities for each of the 60 problems. Participants were asked to choose once between the sure payoff and the risky gamble in each of the 60 problems that were randomly ordered. A typical screen and the instructions are presented in Appendix 2.

Condition Experience-Sampling (E-sampling, one shot decisions from experience)

Forty Technion students participated in this condition. They were randomly assigned to two different sub-groups. Each sub-group contained 20 participants who were presented with a representative sample of 30 problems from the estimation set (each problem appeared in only one of the samples, and each sample included 10 problems

from each payoff domain). The participants were told that the experiment includes several games, and in each game they were asked to choose once between two decks of cards (represented by two buttons on the screen). It was explained that before making this choice they will be able to sample the two decks. Each game was started with the sampling stage, and the participants were asked to press the "choice stage" key when they felt they had sampled enough (but not before sampling at least once from each deck).

The outcomes of the sampling were determined by the relevant problem. One deck corresponded to the safe alternative: All the (virtual) cards in this deck provided the medium payoff. The second deck corresponded to the payoff distribution of the risky option; e.g., sampling the risky deck in problem 21 resulted with the payoff "+2 Sheqels" in 10% of the cases, and outcome "-5.7 Sheqels" in the other cases.

At the choice stage participants were asked to select once between the two virtual decks of cards. Their choice yielded a (covert) random draw of one card from the selected deck and was considered at the end of the experiment to determine the final payoff. A typical screen and the instructions are presented in Appendix 2.

Condition Experience-repeated (E-repeated, repeated decisions from experience):

One-hundred Technion students participated in this condition. They were randomly assigned to five different sub-groups. Each sub-group contained 20 participants who were presented with 12 problems (each problem appeared in only one of the samples, and each sample included an equal proportion of problems from each payoff domain). Each participant was seated in front of a personal computer and was presented with each of the problems for a block of 100 trials. Participants were told that the experiment would include several independent sections (each section included a repeated play of one of the 12 problems), in each of which they would be asked to select between two unmarked buttons that appeared on the screen (one button was associated with the safe alternative and the other button corresponded to the risky gamble of the relevant problem) in each of an unspecified number of trials. Each selection was followed by a presentation of its outcome in Sheqels (a draw from the distribution associated with that button, e.g., selecting the risky button in problem 21 resulted in a gain of 2 Sheqels with probability 0.1 and a loss of 5.7 Sheqels otherwise). Thus, the feedback was limited to

the obtained payoff; the forgone payoff (the payoff from the unselected button) was not presented. A typical screen and the instructions are presented in Appendix 2.

1.3 The competition study

The competition session in each condition was identical to the estimation session with two exceptions: Different problems were randomly selected, and different subjects participated. Table 1b presents the 60 problems which were selected by the same algorithm used to draw the problems in the estimation sessions. The 160 participants were drawn from the same population used in Study 1 (Technion students) without replacement. That is, the participants in the competition study did not participate in the estimation study, and the choice problems were new problems randomly drawn from the same distribution.

1.4 The competition criterion: Mean Squared Distance (MSD), interpreted as the Equivalent Number of Observations (ENO).

The competitions used a Mean Squared Distance (MSD) criterion. Specifically, the winner in each competition is the model that minimizes the average squared distance between the prediction and the observed choice proportion in the relevant condition (the mean over the 20 participants in Conditions Description and E-sampling, and over the 20 participants and 100 trials in Condition E-repeated). This measure has several attractive features. Two of these features are well known: The MSD score underlies traditional statistical methods (like regression and the t-test) and is a proper scoring rule (see Brier, 1950; Selten, 1998; and a discussion of the conditions under which the properness is likely to be important in Yates, 1990). Two additional attractive features emerge from the computation of the ENO (Equivalent Number of Observations), an order-preserving transformation of the MSD scores (Erev, Roth, Slonim, & Barron, 2007). The ENO of a model is an estimation of the size of the experiment that has to be run to obtain predictions that are more accurate than the model's prediction. For example, if a model has an ENO of 10, its prediction of the probability of the R choice in a particular problem is expected to be as accurate as the prediction based on the observed proportion of R choices in an experimental study of that problem with 10 participants. Erev et al. show

that this score can be estimated as ENO = $S^2/(MSE - S^2)$ where S^2 is the pooled estimated variance over problems, and MSE is the mean squared distance between the prediction and the choices of the individual subjects (0 or 1 in the current case).[5] When the sample size is n = 20, MSE = MSD + $S^2(20/19)$.

One advantage of the ENO statistics is its intuitive interpretation as the size of an experiment rather than an abstract score. Another advantage is the observation that the ENO of the model can be used to facilitate optimal combination of the models' prediction with new data; in this case the ENO is interpreted as the weight of the model's prediction in a regression that also includes the mean results of an experiment (see a related observation in Carnap, 1953).

2. The results of the estimation study

The right hand columns in Table 1a present the aggregate results of the estimation study. They show the mean choice proportions of the risky prospect (the R-rate) and the mean number of samples that participants took in condition E-sampling over the two prospects (60% of the samples were from the risky prospect).

2.1 Correlation analysis and the weighting of rare events

The left hand side of Table 2 presents the correlations between the risky choices (R-rates) in the three conditions using problem as a unit of analysis. The results over the 58 problems without dominant[6] alternatives reveal a high correlation between the two experience conditions (r[E-Sampling, E-Repeated] = 0.83, $p < .0001$), and a large difference between these conditions and the description condition (r[Description, E-Sampling) = -0.53, p = .0004; and r[Description, E-Repeated] = -0.37, p = .004). The lower panel in Table 2 distinguishes between problems with and without rare events. These analyses demonstrate that only with rare events does the difference between experience and description emerge.

---

[5] A reliable estimation of ENO requires a prior estimation of the parameter of the models, and a random draw of the experimental tasks. Thus, the translation of MSD scores to ENO is meaningful in an experiment such as this one in which parameters are estimated from a random sample of problems, and predictions are over another random sample from the same distribution of problems.

[6] There were two problems that included a dominant alternative in the estimation set (problems 1 and 43) and 4 such problems in the competition set (problems 15, 22, 31, 36).

Additional clarification of this difference between the three conditions is provided in Figure 1a, which presents the R-rate as a function of Ph by condition. The results reveal an increase in the R-rates with Ph in the two experience conditions, and a decrease in the description condition. Since for each value of Ph the riskless payoff M is on average equal to the expected value of the risky lottery, this pattern is consistent with the assertion that people exhibit overweighting of rare events in decisions from description, and underweighting of rare events in decisions from experience (see Barron & Erev, 2003).

<Insert Table 2, Insert Figure 1>

3. Baseline models

The results of the estimation study were posted on the competition Website on April 1st 2008 (a month before the beginning of the competition study). At the same time EER posted several baseline models. Each model was implemented as a computer program that satisfies the requirements for submission to the competition. The baseline models were selected to achieve two main goals. The first goal was technical: The programs of the baseline models were part of the "instructions to participants." They served as examples of feasible submissions.

The second goal was to illustrate the range of MSD scores that can be obtained. One of the baseline models for each condition was the best model that EER could find (in terms of fitting the results of the estimation study). The presentation of these "strong baselines" was designed to reduce the number of submissions that were not likely to win the competition.

The following sections describe some of the baseline models. We present the strongest baseline for each competition (the one that minimized the MSD on the estimation set). To clarify the relationships of strongest baselines to previous research, we start each subsection with the presentation of one predecessor of the strongest baseline.

3.1 Baseline Models for Condition Description (One-shot decisions under risk)

3.1.1 Original (5-parameter) Cumulative prospect theory (CPT)

According to cumulative prospect theory (Tversky & Kahneman, 1992), decision-makers are assumed to select the prospect with the highest weighted value. The weighted value of Prospect X that pays x_1 with probability p_1, and x_2 otherwise (probability p_2=1-p_1) is:

$$(1) \quad WV(X) = V(x_1)\pi(p_1) + V(x_2)\pi(p_2)$$

where $V(x_i)$ is the subjective value of outcome x_i, and $\pi(p_i)$ is the subjective weight of outcome x_i. The subjective values are given by a value function that can be described as follows:

$$(2) \qquad V(x_i) = \begin{cases} x_i^{\alpha} & if \quad x_i \geq 0 \\ -\lambda |x_i|^{\beta} & if \quad x_i < 0 \end{cases}$$

The parameters $0 < \alpha < 1$ and $0 < \beta < 1$ reflect diminishing sensitivity to increases in the absolute payoffs in the gain and the loss domain respectively. According to the diminishing sensitivity assumption the subjective impact of a change in the absolute payoff decreases with the distance from zero (see Tversky & Kahneman, 1992, and motivating observations in Stevens, 1957). The parameter $\lambda > 1$ captures the loss aversion assertion suggesting that losses loom larger than equivalent gains.

The subjective weights are assumed to depend on the outcomes' rank and sign, and on a cumulative weighting function. When the two outcomes are of different signs, the weight of outcome *i* is:

$$(3) \qquad \pi(p_i) = \begin{cases} \dfrac{p_i^{\gamma}}{(p_i^{\gamma} + (1-p_i)^{\gamma})^{1/\gamma}} & if \quad x_i \geq 0 \\ \dfrac{p_i^{\delta}}{(p_i^{\delta} + (1-p_i)^{\delta})^{1/\delta}} & if \quad x_i < 0 \end{cases}$$

The parameters $0 < \gamma < 1$ and $0 < \delta < 1$ capture the tendency to overweight low-probability extreme outcomes.

When the outcomes are of the same sign, the weight of the most extreme outcome (largest absolute value) is computed with equation (3) (as if it is the sole outcome of that

sign), and the weight of the less extreme outcome is the difference between that value and 1.

The competition Website (EER, 2008) presents the predictions of CPT with the parameters that best fit the current data. The top left-hand side of Table 3a presents the estimated parameters and three measures of the accuracy (fit) of the model with these parameters. The first two measures are the proportion of agreement between modal choice and the prediction (Pagree = 1 if the observed and predicted R-rates fall in the same side of 0.5; it equals 0.5 if one of the two equals 0.5; and 0 otherwise) and the correlation between the observed and the predicted results across the 60 problems. These measures show high agreement (95%) and high correlation (0.85). The third measure, and the focus of the current competition, is a Mean Square Distance (MSD) score. It reflects the mean of the squared distance of the prediction from the mean results (over participants) in each problem. Thus it is the mean of 60 squared distance scores.

< Insert Table 3 >

3.1.2 Stochastic cumulative prospect theory (SCPT).

The second model considered here was found to be the best baseline model in Condition Description. It provided the best fit for the estimation data. This model is a stochastic variant of cumulative prospect theory proposed by Erev, Roth, Slonim and Barron (2002; and see a similar idea in Busemeyer, 1985). The model assumes that the probability of selecting the risky prospect (R) over the safe prospect (S) increases with the relative advantage of that prospect. Specifically, this probability is:

$$\Pr(R) = \frac{e^{WV(R)(\mu/D)}}{e^{WV(R)(\mu/D)} + e^{WV(S)(\mu/D)}} = \frac{1}{1 + e^{[WV(S)-WV(R)](\mu/D)}} \tag{4}$$

The parameter μ captures the sensitivity to the differences between the two prospects, and D is the absolute distance between the two value distributions (under CPT). In the current context D = |H-M|[π(Ph)]+|M-L|[π(1-Ph)].

Table 3a presents the scores of SCPT with the parameters that best fit the estimation data set (α=.89, β=.98, λ=1.5,μ=2.15 γ=δ=.7). Comparison with the CPT row shows that the stochastic response rule (added in SCPT) dramatically reduces the MSD

score (from .093 to .012). To clarify the intuition behind this advantage, consider two problems in which the observed R-rates are 0.75 and 1.0. Deterministic models like CPT cannot distinguish between the two problems. Their MSD score is minimized by predicting R-rates of 1.0 in both problems. Thus the minimal MSD score is $[(1- .75)^2 -(1 - 1)^2]/2 = 0.03125$. Stochastic models like SCPT can distinguish between these problems and their minimal MSD score is 0. Notice that when parameter μ is large, SCPT approximates the predictions of CPT. The advantage of SCPT highlights the importance of this parameter.

3.1.3 Other baseline models for Condition Description

The other baseline models considered by EER for Condition Description include restricted variants of SCPT, and the priority heuristic (Brandstätter, Gigerenzer, & Hertwig, 2006). The analysis of the restricted variants of SCPT highlights the robustness of this model: It provides useful predictions even when it is used with the parameters estimated in previous research. The analysis of the priority rule shows that its fit of the current data is comparable to the fit of the original variant of CPT.

3.2 Baseline models for Condition E-sampling (one-shot decisions from experience)

3.2.1 Primed sampler

The primed sampler model (Erev, Glozman & Hertwig, 2008) implies a simple choice rule in condition sampling: The participants are expected to take a sample of k draws from each alternative, and select the alternative with the higher sample mean. Table 3b shows that this simple model provides a good approximation of the current results. The value $k = 5$ minimizes the MSD score.

3.2.2 Primed sampler with variability.

Under a natural extension of the primed sampler model the exact value of the sample size differs between participants and decisions. The current model captures this idea with the assumption that the exact sample size (from each alternative) is uniformly drawn from the integers between 1 and k. Best fit is obtained with k =9. Table 3b shows that the added variability improves the fit.

3.3 Baseline models for Condition E-Repeated (repeated decisions from experience)

3.3.1 Explorative sampler

The predictions of the explorative sampler model (Erev, Ert & Yechiam, 2008) for the current task can be summarized with the following assumptions:

A1: Exploration and exploitation. The agents are assumed to consider two cognitive strategies: exploration and exploitation. Exploration implies a random choice. The probability of exploration is 1 in the very first trial, and it reduces toward an asymptote (at ε) with experience. The effect of experience on the probability of exploration depends on the expected number of trials in the experiment (*T*). Exploration diminishes quickly when *T* is small, and slowly when *T* is large (in the current study *T* = 100). This assumption is quantified as follows:

$$P(Explore_t) = \varepsilon^{\frac{t-1}{t+T^{\delta}}} \quad (5)$$

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

A2: Experiences. The experiences with each alternative include the set of observed outcomes yielded by this alternative in previous trials. In addition, the very first outcome is recalled as an experience with both alternatives.

A3: Naïve sampling from memory. Under exploitation the agent draws (from memory, with replacement) a sample of m_t past experiences with each alternative. All previous experiences are equally likely to be sampled. The value of m_t at trial t is assumed to be randomly selected from the set {1, 2,., *k*} where *k* is a free parameter.

A4: Regressiveness, diminishing sensitivity, and choice. The recalled subjective value of the outcome x (from selecting alternative j) at trial t are assumed to be affected by two factors: regression to the mean of all the experiences with the relevant alternative (in the first t-1 trials), and diminishing sensitivity. Regression is captured with the assumption that the regressed value is $R_x = (1-w)x + (w)A_j(t)$, where w is a free parameter and $A_j(t)$ is the average outcome from the relevant alternative.[7]

[7] Implicit in this regressiveness (the assumption W > 0) is the assumption that all the experiences are weighted (because all the experiences affect the mean). The value of this implicit assumption was demonstrated by Lebiere, Gonzalez and Martin (2007).

Diminishing sensitivity is captured with a variant of prospect theory's (Kahneman & Tversky, 1979) value function that assumes

$$(6) \qquad sv(x) = \begin{cases} R_x^{\alpha_t} & if \quad R_x \geq 0 \\ -(-R_x)^{\alpha_t} & if \quad R_x < 0 \end{cases}$$

where $\alpha_t = (1+V_t)^{(-\beta)}$, $\beta \geq 0$, is a free parameter, and V_t is a measure of payoff variability. V_t is computed as the average absolute difference between consecutive obtained payoffs in the first t-1 trials (with an initial value at 0). The parameter β captures the effect of diminishing sensitivity: large β implies a quick increase in diminishing sensitivity with payoff variability.

The estimated subjective value of each alternative at trial t is the mean of the subjective value of the alternative's sample in that trial. Under exploitation the agent selects the alternative with the highest estimated value.

3.3.2 Explorative sampler with recency

Evaluation of the fitting scores of the explorative sampler model reveals that this model over-predicts the tendency to select the risky prospect. The best baseline model for Condition E-Repeated is a refinement of the explorative sampler model that was developed to address this bias. Specifically, the refined model assumes that the most recent outcome with each alternative is always considered. This assumption triggers a hot stove effect (see Denrell & March, 2001): When the recent payoffs are considered, the effect of low outcomes last longer than the effect of high outcomes (because low outcomes reduce the probability of additional exploration and they remain the most recent outcome more trials). As a result, the refined model predicts lower R-rates. The change is implemented by replacing assumption A3 with the following assumption:

A3': Naïve sampling from memory with recency. Under exploitation the agent draws (from memory, with replacement) a sample of m_t past experiences with each alternative. The first draw is the most recent experience with each alternative. All previous experiences are equally likely to be sampled in the remaining m_t-1 draws.

Table 3c presents the scores of the refined model with the parameters that best fit the estimation data set (β=.10, w=.3, ε=.12, k=8). Additional analysis reveals that the added recency effect does not impair the predictions of the explorative sampler model in the experimental conditions reviewed by Erev and Haruvy (2009).

3.3.3 Other baseline models for Condition E-repeated

The other baseline models considered by EER for Condition E-Repeated include different variants of reinforcement-learning models. This analysis shows the advantage of the normalized reinforcement-learning model (see Erev & Barron, 2005; and a similar model in Erev, Bereby-Meyer & Roth, 1999), over basic reinforcement-learning models. In addition, it shows that it is not easy to find a reinforcement-learning model that outperforms the explorative sampler model with recency.

4. The competition sessions.

Table 1b presents the aggregate experimental data of the competition sessions. They show the mean choice proportions of the risky prospect (the R-rate) and the mean samples that participants took in condition E-sampling.

4.1 Correlation analysis and the weighting of rare events

The right-hand columns in Table 2 present the correlations between the R-rates in the different conditions in the competition study, and Figure 1b presents the R-rates by Ph. The results replicate the pattern documented in the estimation study. The two experience conditions were similar, and different from the description condition. The difference suggests that the R-rates increase with Ph in the two experience conditions, and decrease with Ph in the description condition.

5. Competition results

Twenty-three models were submitted to participate in the different competitions; eight to the description condition, seven to the E-sampling condition and eight to the E-repeated condition. The submitted models involved a large span of methods ranging from logistic regression, ACT-R based cognitive modeling, neural networks, production

rules, and basic mathematical models. In accordance with the competition rules, the ranking of the models was determined based on the mean squared distance (MSD) between the predicted and observed choice proportion in the competition data set.

5.1 Condition Description:

The lower panels in Table 3a present the three best submitted models for Condition Description. Two of these abstractions are variants of cumulative prospect theory (and some added assumptions) with a stochastic choice rule. The winner of this competition is a logit-regression model submitted by Ernan Haruvy, described in detail in the following section.

5.1.1 The winning model in Condition Description: Linear utility and logistic choice

The current model was motivated by the observation that leading models of decisions from description, like prospect theory, imply weighting of several variables (functions of the probabilities and the outcomes). That is, they can be described as regression models. Under this assertion, one can use regression techniques in order to facilitate the predictive accuracy of models of this type. Thus, Haruvy submitted the best regression-based model that he could find to the competition.

The model can be captured by two equations. The first defines T(R) -- the tendency to prefer the risky prospect:

(7) $$T(R) = \beta 0 + \beta 1*H + \beta 2*L + \beta 3*M + \gamma 1*Ph + \gamma 2*EV(R) + \gamma 3*(Dummy1)$$

The values H, L, M and Ph are the parameters of the choice problem as defined above. EV(R) is the expected payoff of the risky prospect, and Dummy1 is a dummy variable that assumes the value 1 if the risky choice has higher expected value than the safe choice and 0 otherwise.

The second equation assumes a logistic choice rule that defines the predicted proportion of risky choices, P(R), based on the relevant tendency:

(8) $$P(R) = \frac{1}{1 + e^{-T(R)}}$$

The estimated parameters are: $\beta 0 = 1.004$, $\beta 1 = 0.012$, $\beta 2 = 0.066$, $\beta 3 = -0.410$, $\gamma 1$ 1.417, $\gamma 2 = 0.317$, $\gamma 3 = -0.621$.

Table 3a shows that the current model provides a better fit (lower MSD score) for the estimation data than the best baseline model. The model has slightly higher MSD score in the competition set (0.0126 vs. 0.0099 in the estimation set). The implied ENO is 56.4. This value implies that the model's accuracy (in predicting the population mean) is similar to the expected accuracy of the observed R-rate in an experiment with 56 participants.

5.1.2 Comparison to other models

Comparison of the winner to the best baseline (SCPT) reveals that SCPT provides more useful predictions. Its ENO was 80.99. Analysis of the differences between the two models suggests that the linear utility and logistic choice predictions tend to be more conservative than SCPT. That is, the former's predictions are somewhat biased towards 50%. This observation suggests that the normalized stochastic response rule assumed by SCPT may be a better approximation to behavioral data than the logistic response rule used in the regression model.

Evaluation of the deterministic models shows that CPT outperformed the priority heuristic, but has relatively low ENO (2.32). As noted above, the low ENO is a reflection of the fact that deterministic models, like CPT, cannot discriminate between problems in which almost all the participants select the modal choice, and problems in which only small majority select the modal choice.

5.1.3 Intuition

Another interesting analysis of the models' predictions involves the comparison between their accuracy and the accuracy of intuitive predictions. To evaluate this relationship we asked 32 Harvard students to predict the proportion of R choices in each of the problems of the competition set. The 32 "predictors" played each of the problems

themselves for real money (just like the participants of the competition set) before making their predictions. To motivate the predictors to be accurate, they were also compensated based on the accuracy of their prediction via a proper scoring rule; this compensation decreased linearly with their MSD score in a randomly selected problem. Table 3a shows that the students' intuition was not very useful for predicting the competition data. The intuitive predictions of the typical predictors were outperformed by the predictions of most models (the intuition MSD was 0.01149, and the median ENO was only 1.88). Additional analysis reveals that in 97% of the problems the mean estimations were conservative (closer to 50% than the actual results). For example, in Problem 15 (-3.3, 0.97; -10.5) or (-3.2), the observed R-rate was 0.1, and the mean intuitive prediction was 0.34. This conservatism of the mean judgments can be a product of a stochastic judgment process (see Erev, Wallsten, & Budescu, 1994).

5.2 Condition E-Sampling:

Table 3.b presents the three best submitted models for condition E-Sampling. The winner in this competition is the ensemble model submitted by Stefan Herzog, Robin Hau, and Ralph Hertwig. This model assumes four equally likely choice rules.

5.2.1. The winning model in Condition E-Sampling: Ensemble

The ensemble model is motivated by three observations. First, different people appear to use different mental tools when making decisions from experience and simple, robust models predict these decisions well (Hau, Pleskac, Kiefer, & Hertwig, 2008). Second, several variants of the models considered above perform well above chance in predicting the estimation data, and equally important, the correlations between the models' errors are relatively low. Third, research on forecast combination has demonstrated that averaging predictions from different models is a powerful tool for boosting accuracy (e.g., Armstrong, 2001; Hibon & Evgeniou, 2005; Timmermann, 2006). To the extent that individual models predict decisions well above chance, and errors are uncorrelated between models[8], the average across models may even outperform

[8] How strongly the errors of two models are correlated can be summarized by their bracketing rate (Larrick & Soll, 2006), which is the proportion of predictions where the two models err on different sides of the

the best individual model.

The ensemble model assumes that each choice is made based on one of four equally likely rules; thus, the predicted choice rate is the average across the predictions of the four models (rules), using equal weights[9]. The first two rules in the ensemble are variants of the natural-mean heuristic (see Hertwig & Pleskac, 2008). The first rule is similar to the primed sampler model with variability described in Section 3.2.2. The decision makers are assumed to sample each option m times, and select the option with the highest sample mean. The value of m is uniformly drawn from the set {1,2….9}. Predictions below 5% or above 95% were curbed to these valued, as more extreme proportions were not observed in the estimation set. The second rule is identical to the first, but m is drawn from the distribution of sample sizes observed in the estimation set, with samples larger than 20 treated as 20 (mean = 6.2; median = 5).

The third rule in the ensemble is a stochastic variant of *cumulative prospect theory* (Tversky & Kahneman, 1992). Its functions are identical to the functions assumed by the SCPT model presented in Section 3.1.2 with the exception that D is set to equal 1. However, the current implementation rests on quite different parameter values (and implied processes), namely, values fitted to the estimation set: $\alpha = 1.19$, $\beta = 1.35$, $\gamma = 1.42$, $\delta = 1.54$, $\lambda = 1.19$, $\mu = 0.41$. These values imply underweighting of rare events and a reversed S-shape value function (a mirror image of the functions that Tversky & Kahneman estimated for decisions from descriptions).

The final rule is a stochastic version of the lexicographic *priority heuristic* (Brandstätter et al., 2006). The stochastic version was adapted from the *priority model* proposed by Rieskamp (2008). Up to three comparisons are made in one of two orders of search. The first order begins by comparing minimum outcomes (i.e., minimum gain or minimum losses depending on the domain of gambles), then their associated probabilities, and finally the maximum outcomes. The second order begins with probabilities of the minimum outcomes, then proceeds to check minimum outcomes, and

---

truth (i.e., one model over- and the other underestimates the true value). In the long run, the average prediction of several models will necessarily be at least as accurate as the prediction of a randomly selected model. The former will outperform the latter when the bracketing rate is larger than zero, and therefore, some errors will cancel each other out.

[9] Equal weighting is robust and can outperform more elaborate weighting schemes (Clemen, 1989; Einhorn & Hogarth, 1975; Timmermann, 2006).

ends with the maximum outcomes (the probabilities with which both search orders are implemented were determined from the estimation set: $p_{order\ 1} = 0.38$; $p_{order\ 2} = 0.62$). The difference between the values being compared is transformed into a subjective difference, normally distributed around an objective difference.[10] The variance of the distribution is a free parameter estimated to equal $\sigma = .037$. If the subjective difference involving the first comparison in each search order exceeds a threshold *t*, the more attractive option is selected based on this comparison; otherwise the next comparison is executed. The values of the thresholds are free parameters. The estimated values are T_o= 0.0001 for the minimum- and maximum-based comparisons, and T_p=0.11 for the probability-based comparison.

The priority rule as implemented here differs in several respects from the original priority heuristic, which was originally proposed to model decisions from description (Brandstätter et al., 2006). Most importantly, the heuristic assumed only one search order, namely, the first order described above. The fitted parameters suggest that in the current decisions from experience, most subjects (62%) follow the second order described above. This difference is important because the correlation between the predicted behaviors assuming the two orders is negative (-0.66).

5.2.2 Comparison to other models

Comparison of the winner to the best baseline (primed sampler with variability) reveals that the ensemble model provides more useful predictions. Although both models have larger MSD in the competition data, the ensemble model gets to an ENO of 25.92, higher than the primed sampler (15.23).

The advantage of the ensemble model highlights the value of the assumption that several decision rules are used. The success of this assumption can be a product of a within-subject variability (the use of different rules at different points in time), a

[10] The exact means of these subjective distributions depend of the sign of the payoff H and on the maximal absolute payoff (MaxAbs= Max[Abs(L),Abs(M),Abs(H)]). In the minimum-based comparison the mean is (Min-M)/MaxAbs where Min =L if H>0, and H otherwise. In the probability-based comparison the mean is Ph when H > 0, and Ph-1 otherwise. In the maximum-based comparison the mean is (Maxi-M)/MaxAbs where Maxi =H if H>0, and L otherwise.

between-subject variability (different people use different rules), and/or between-problem variability (the different problems trigger the usage of different rules).

5.3 Condition E -Repeated

The lower panel in Table 3.c presents the three best submitted models for Condition E-Repeated. All models succeed in capturing the main behavioral trends observed in the data: the underweighting of rare events, and the hot stove effect. The models differed in their ways of capturing these trends. Two of the models were based on contingent sampling, and the third focused on normalized reinforcement learning that assumed inertia. The winner of this competition is the model submitted by Terrence Stewart, Robert West, and Christian Lebiere. This model uses the ACT-R architecture and assumes similarity-based inference.

5.3.1 The winning model in Condition E-Repeated: ACT-R, blending, and sequential dependencies

The current model rests on the assumption that the effect of experience in Condition E-Repeated is similar to the effect of experience in other settings. Thus, it can be captured by the general abstraction of the declarative memory system provided by the ACT-R model (Anderson & Lebiere, 1998). The model can be summarized as follows:

Declarative memory with sequential dependencies. Each experience is coded into a chunk that includes the context, choice, and obtained outcome. The context is abstracted here by the two previous consecutive choices (see related ideas by Lebiere & West, 1999; West et al., 2005). At each trial, the decision maker considers all her experiences under the relevant context, and recalls all the experiences with activation levels that exceeded the activation cutoff (captured by the parameter τ).

The activation level of experience i is calculated using Equation 1, where t_k is the amount of time (number of trials) since the kth appearance of this item, d is the decay rate, and $\varepsilon(s)$ is a random value chosen from a logistic distribution with variance $\pi^2 s^2/3$.

(9) $$A_i = \ln \sum_{k=1}^{n} t_k^{-d} + \varepsilon(s)$$

The learning term of the equation captures the power law of practice and forgetting (Anderson & Schooler, 1991), while the random term implements a stochastic "softmax" (a.k.a. Boltzmann) retrieval process where the probability P_i of retrieving i is given by:

$$P_i = \frac{e^{A_i/t}}{\sum_j e^{A_j/t}}$$

where $t = \sqrt{2}s$ and the summation is over all experiences over the retrieval threshold.

Choice rule - blending memories. When the model attempts to recall an experience that matches the current context, multiple experiences (chunks) may be found. For example, when recalling previous risky choices there are two chunks in memory – one for cases that resulted in the high reward and another for cases associated with the low reward. In such cases the chunks are blended such that the mean recall value of each alternative at trial t is the weighted (by P_i) mean over all the recalled experiences. The alternative with the larger mean is selected (see related ideas in Gonzalez et al., 2003).

Parameters. The value for parameter d in Equation 1 was set to 0.5, as this is the value used in almost all ACT-R models. The other two parameters were estimated based on the estimation set, using the relativized equivalence methodology (Stewart & West, 2007). The estimated values are s = 0.35 and τ = -1.6. It should be noted that these values are very close to the default settings for ACT-R, and there are only minor differences in predictions between this model and a purely default standard ACT-R model with no parameter fitting at all.

5.3.2. Comparison to other models

Implicit in the ACT-R model is the assumption of high sensitivity to a small set of previous experiences in situations that are perceived to be similar to the current choice task. The best baseline model (explorative sampler with recency) can be described as a different abstraction of the same idea. The baseline model provided slightly better predictions. Its ENO was 47.22 (compared to the ACT-R ENO of 32.5).

5.4 The relationships among the three competitions

Comparisons of the models submitted to the three competitions show large differences among the description and the two experience competitions. Whereas all the models in the description competitions assumed that outcomes are weighted by probabilities, the concept "probability" did not play an important role in the models submitted to the experience-repeated competition. Another indication of the large differences between the competitions comes from an attempt to use the best models in one competition to predict the results in a second competition. This analysis reveals that all the models developed to capture behavior in the description condition have ENO below 3 in the two experience conditions. For example, the best model in Condition Description (SCPT with ENO of 80.99) has ENO of 2.34 and 2.19 in Condition E-Sampling and E-Repeated, respectively. These values are lower than the ENO of a model that predicts random choice (the ENO of the random choice model that predicts an R-rate of 50 in all problems is 5.42 in Condition E-Sampling and 6.75 in Condition E-Repeated). Similarly, the best model in the E-sampling condition (Ensemble with an ENO of 25.92) has an ENO of 1.13 in Condition Description. Again, this value is lower than for a model that predicts random choice (1.89 in Condition Description).

6. Additional descriptive analyses

6.1 Learning curves

Figures 2a and 2b present the observed R-rates in Condition E-repeated in 5 blocks of 20 trials. The learning curves documented in the 60 problems in each study were plotted in 12 graphs. The classification of the problems to the 12 graphs was based on two properties: the probability of high payoff (Ph) and the relative value of the risky prospect. The most common pattern is a decrease in risky choices with experience. This

pattern is predicted by the hot stove effect (Denrell & March, 2001). Comparison of the three rows suggests an interesting nonlinear relationship between the probability of high payoff (Ph) and the magnitude of the hot stove effect. A decrease in R rate with experience is clearer for high Ph and low Ph, but not for medium Ph level. This nonlinear relationship explains why previous studies that focus on gambles with equally likely outcomes (like Biele, Erev & Ert, 2009) found no evidence for the hot stove effect. The learning curves in the medium Ph problem show higher sensitivity to the expected values. This pattern can be a product of the joint effect of underweighting of rare events and the hot stove effect.

< Insert Figure 2>

7. Discussion

The current project was motivated by the hope that a careful study of quantitative predictions could contribute to behavioral decision research by facilitating theoretical insights and clarifying the models and the boundaries of the different phenomena. In addition, we suggested that the organization of prediction competitions can facilitate the study of quantitative models. We can re-evaluate these potential contributions in light of the findings and experience described above.

7.1 On predictions and explanations

The current results shed some light on the ways in which quantitative models can facilitate the development of theoretical explanations of choice behavior. One example is the success of the ACT-R model (and the similar baseline explorative sampler model) in condition E-Repeated. This success suggests that the processes that underlie repeated decisions are likely to be close to the processes that underlie retrieval from memory.

A more developed example involves the explanation of gaming (the decision to buy lotteries and play Casino games). The most popular explanation of gaming involves the assertion that rare events are overweighted (Kahneman & Tversky, 1979). The current analysis highlights an important shortcoming of this explanation, and suggests an alternative. The shortcoming is revealed by the observation of underweighting of rare

events in decisions from experience. Since people game even when they base decisions on experience, the overweighting explanation seems insufficient.

Another interesting shortcoming of the common explanations of gaming involves the observation that most people do not game frequently. For example, in a survey conducted in New Zealand, Amey (2001) found that 87% of the 1500 respondents exhibit this behavior at least once in the last 12 months, but only 10% of the respondents game more than 6 times at that period.[11] The alternative explanation suggested by the current analysis involves the possibility that gaming, by non-professionals, is a reflection of the stochastic nature of choice behavior. It is possible that like the decisions analyzed here, the decisions to game are best described by a stochastic model. The fact that people game, under this account, is not the product of a consistent bias. Rather, it is a product of the fact that people are inconsistent: Most people choose to avoid gaming in most cases, but noisy processes lead them to game in some rare cases. The leading models presented above suggest that this inconsistency can result from the tendency to rely on small samples from experience.

7.2 The 1-800 critique

Recall the 1-800 critique from the introduction. Under that critique, leading models of choice behavior do not apply in the same way in all situations, and therefore a toll-free help line may be needed to assist users who wish to apply such models. The prediction competition procedure remedies the problem identified by the 1-800 critique. The competition required the development of models that produce quantitative predictions in a well-defined space of situations. We did not have to call the participants to derive the predictions of their models. The success of the submitted models (their high ENO) suggests that the 1-800 problem is not a reflection of a deep shortcoming of behavioral decision research. The knowledge accumulated in behavioral decision research can be used to derive clear and very useful predictions of behavior. Moreover, the large advantage of the best models over intuition implies that exams in behavioral decision research need not always be dissimilar to exams in the exact sciences.

---

[11] We thank Robin Hogarth for this example.

7.3 The boundaries of the different regularities

The experimental part of the current project clarifies the boundaries of the tendencies to overweight and underweight rare events. Even on randomly chosen problems, when the risky prospects involve rare outcomes (probability below 0.2), the results reveal significant negative correlation between the proportions of risky choices in decisions from description and decisions from experience. The observed deviations from maximization suggest that the participants behave as if they overweight these rare events in decisions from description, and underweight rare events in decisions from experience. This pattern replicates previous findings (e.g., Barron & Erev, 2003; Hertwig et al., 2004). The main contribution of the current replication is the demonstration of the generality of this gap: It is not limited to the particular problems used in the initial demonstrations, but reliably emerges in experiments that study 120 randomly selected choice problems.

The high ENO of the stochastic version of cumulative prospect theory (SCPT) in the description condition, but not in the two experience conditions, also underlines the difference between decisions from description and from experience. It highlights the limitations of trying to build general theories of decision making by focusing only on decision making in environments in which clear counterexamples to the predictions of expected utility theory can be constructed. The competition results suggest that the differences between decisions from description and from experience may be differences in kind, more than just differences in parameters. This point of view gains some support from the observation that no participant in the competition chose to use parametric variations of a single model in the different settings.

So we are claiming not just that SCPT does not win in the experience conditions, but that it does not do very well. Similarly for the other models; our conclusions would be quite different if some model was a close second choice for every condition, for example. But that was not the case; the models that predicted well for decisions from description predicted poorly for decisions from experience, and vice versa.

7.4 Partial effectiveness and future research

The choice prediction competition attempted to achieve two related goals. The first was to facilitate the development of clear and useful quantitative models of choice behavior. We believe this goal was only partly achieved. First, only in one of the three competitions did the winning model outperform the best baseline model. A pessimistic interpretation is that the competition procedure is not a very effective way to produce useful models. We favor, however, a more optimistic interpretation: The best baselines were minimal modifications of models that were found to have high ENO in previous research on similar problems. The submitted models, on the other hand, reflected more creative generalizations. The success of the baseline models suggests that the knowledge accumulated in previous studies of quantitative models is very useful. It seems likely that the creative approach taken by the participants (and by other researchers) will eventually lead to the development of more accurate predictions that will outperform the baseline models in all three conditions. And the development of simpler, more principled models will enable to their integration in increasingly broader task contexts.

A more problematic observation involves the complexity of the submitted models. The competition's requirement to submit a model (in terms of a computer program) that enables unambiguous predictions and its emphasis on predictive accuracy came at the expense of other desirable modeling properties such as simplicity. For example, the ensemble model that won the E-sampling competition includes four sub-models and encompasses numerous parameters. Clearly, the model is not easy to handle. However, its predictive success suggests that its key psychological motivation, namely the observation that in making choices different people recruit different psychological processes, is important. Future research could use this insight to develop a simpler version of the ensemble model while retaining its high ENO. In brief, the choice competition as implemented here is not a magical method that stimulates the discovery of simple models. Nevertheless, it is a powerful tool that spurs the development of benchmark models, which tell us how good or bad our established models are. Moreover, the psychological gist of the benchmark models can then be used to develop better descriptive models or to improve on the existing ones. In addition, future research that will broaden the competition to increasingly wider sets of task conditions might have the

effect of favoring simpler, more general models over those optimized and engineered to relatively narrow task conditions.

The second goal of the competition was to clarify the meaning of the term "predictions." Many studies use the same word to refer to fitted values. Moreover, papers that try to distinguish between fitted values and predictions are subject to a selection bias. That is, researchers are more likely to complete papers if the results are clear to them, and subjective clarity is correlated with the success of the researchers' favorite model. Thus, potential readers of papers that focus on quantitative predictions often treat them with distrust and/or ignore them. We believe that the choice prediction competition procedure has addressed these important problems. It provides a clear definition of the term "prediction," and implies mechanism to minimize selection biases and enhance trust.

7.5 Summary

The current project addresses a decision problem faced by behavioral decision scientists: the decision between "a focus on counterexamples" or "a focus on quantitative models." This decision is made from experience. The decision makers (scientists) do not receive a description of the incentive structure. They have to rely on their personal experience, and on the experience of other scientists. Our analysis suggests that a focus on counterexamples is likely to lead to better outcomes most of the time. The evaluation of quantitative predictions tends to be more expensive and less interesting than the evaluation of counterexamples. Nevertheless, it is not clear that the "focus on counterexamples" choice always maximizes expected return. There are reasons to believe that in certain (perhaps rare) cases a focus on quantitative models can lead to extremely important results. Some of the most important breakthroughs in science were based on prior development of useful quantitative models of the relevant phenomena. We believe that behavioral decision scientists tend to underweight these rare cases, and hope that the current project clarifies the value of quantitative predictions and will help change this situation.

Appendix 1: The problem selection algorithm.

The 60 problems in each set are determined according to the following algorithm:

- The probability p is drawn (with equal probability) from one of the following sets (.01-.09), (.1-.9), (.91-.99) . (Each interval is chosen with probability 1/3, and points within the interval are then chosen with equal probability from a grid with interval .01.
- Two random draws are generated for the risky option (Xmax, Xmin):
- Xmin is drawn (with equal probability) from (-10, 0); Xmax is drawn from (0, +10).
- H'= Round(Xmax, .1)[12]
- L'= Round(Xmin, .1)
- The Expected Value of the risky option is determined and an error term is added to create the value of the safe option:
- m = Round(H'*p+L'*(1-p), .1);
- sd = min(abs(m-L')/2,abs(m-H')/2,2); e=rannor(0)*sd; m=m+e; Notice that the addition of e creates some problens with a dominant strategy (see Problem 1 in Table 1a).
- Finally, the dataset is balanced to include equal proportions of problems that include nonpositive payoffs (loss domain), nonnegative payoffs (gain domain) and both positive and negative payoffs (mixed domain), by adding a constant (con) to H and L and M.
- If problem <21 then con = -max+min;
- If 20 < problem < 41 then con = 0;
- If problem > 40 then con = +max-min;
- L = L'+con; M = round(m+con,.1); H = H'+con;

[12] The function Round(x, .1) rounds x to the nearest decimal. The function abs(x) returns the absolute value of x. the function rannor(0) returns a randomly selected value from a normal distribution with a mean of 0 and standard deviation of 1.

Appendix 2: Translation of the instructions and typical experimental screens of each of the three conditions (Description, Experience-Sampling, and Experience-Repeated).

Condition Description:

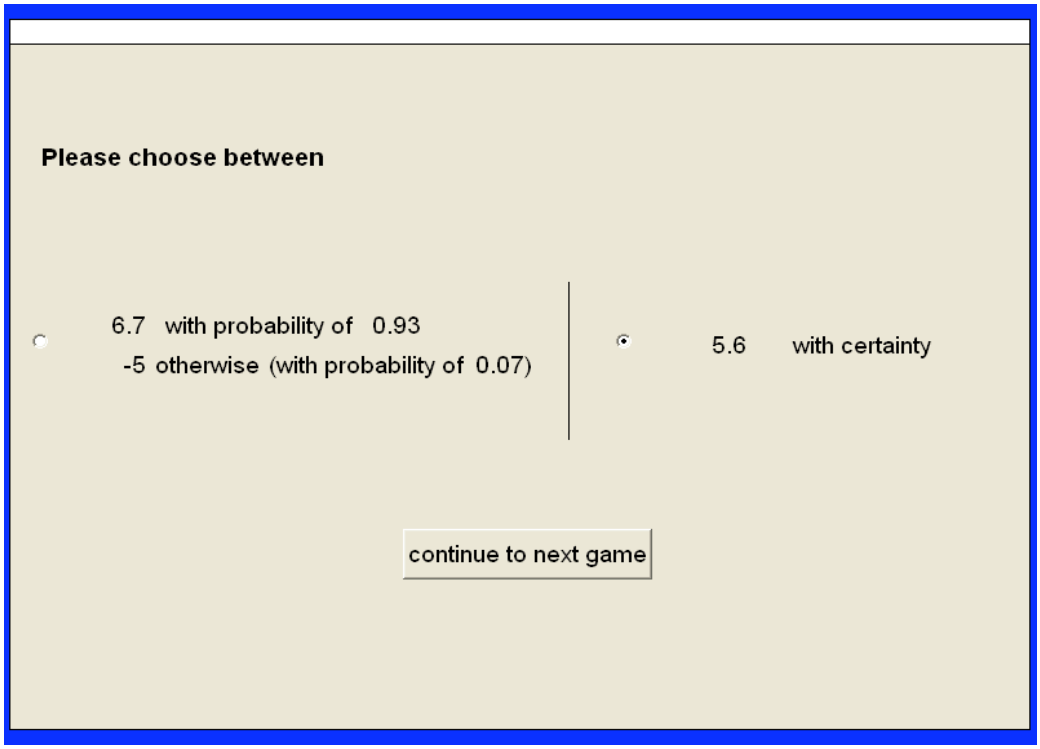
This experiment includes several games. In each game you will be asked to select one of two alternatives.

At the end of the experiment one of the games will be randomly drawn (all the games are equally likely to be drawn), and the alternative selected in this game will be realized.

Your payoff for the experiment will be the outcome (in Sheqels) of this game.

Good luck!

Experimental Screen after selecting the safer option in Problem 32:



Condition E-Sampling:

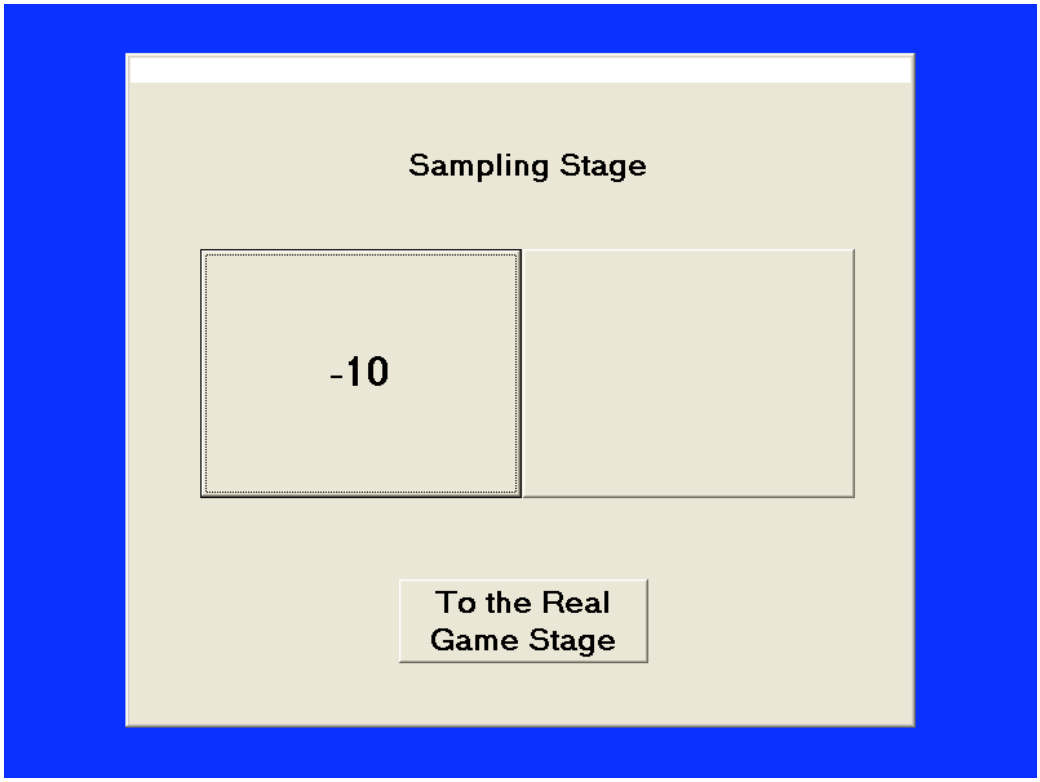
This experiment includes several games. Each game includes two stages: The sampling stage and the choice stage.

At the choice stage (the second stage) you will be asked to select once between two virtual decks cards (two buttons). Your choice will lead to a random draw of one card from this deck, and the number written on the card will be the "game's outcome." During the sampling stage (the first stage) you will be able to sample the two decks. When you feel that you have sampled enough press the "choice stage" key to move to the choice stage.

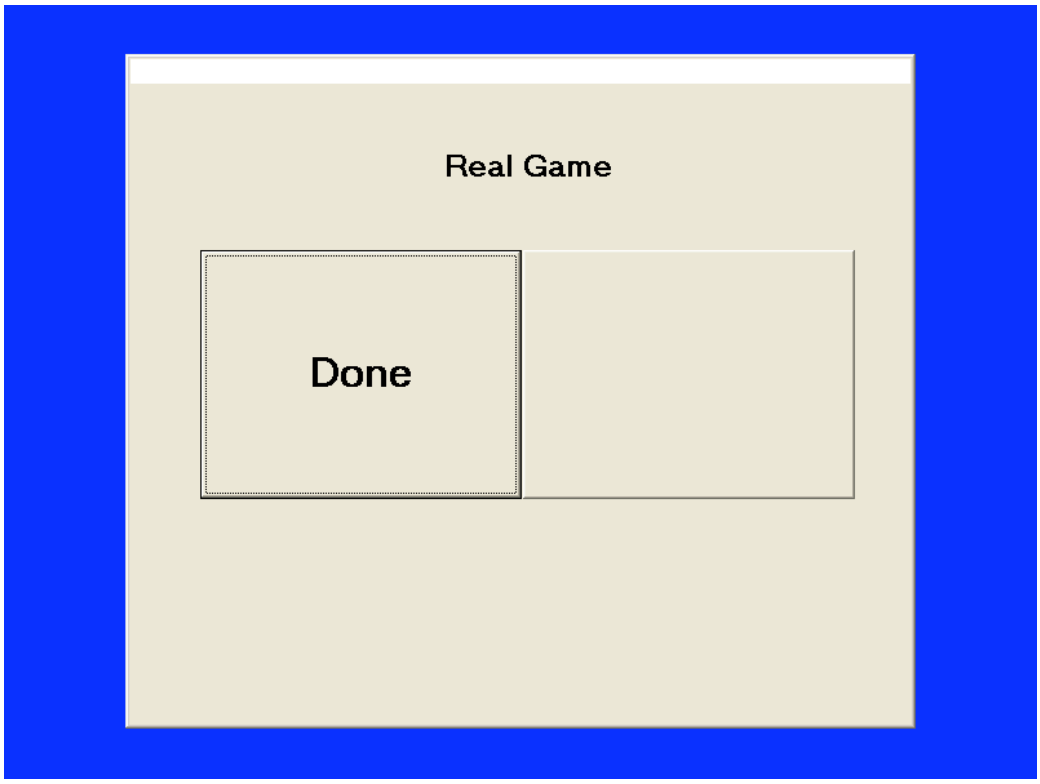
At the end of the experiment one of the games will be randomly drawn (all the games are equally likely to be drawn). Your payoff for the experiment will be the outcome (in Sheqels) of this game.

Good luck!

Experimental screen (a) after sampling the deck associated with the safer option in Problem 4 during the sampling stage:



Experimental screen (b) – After choosing the deck associated with the safer option in Problem 4 during the real game stage:



Condition E-Repeated:

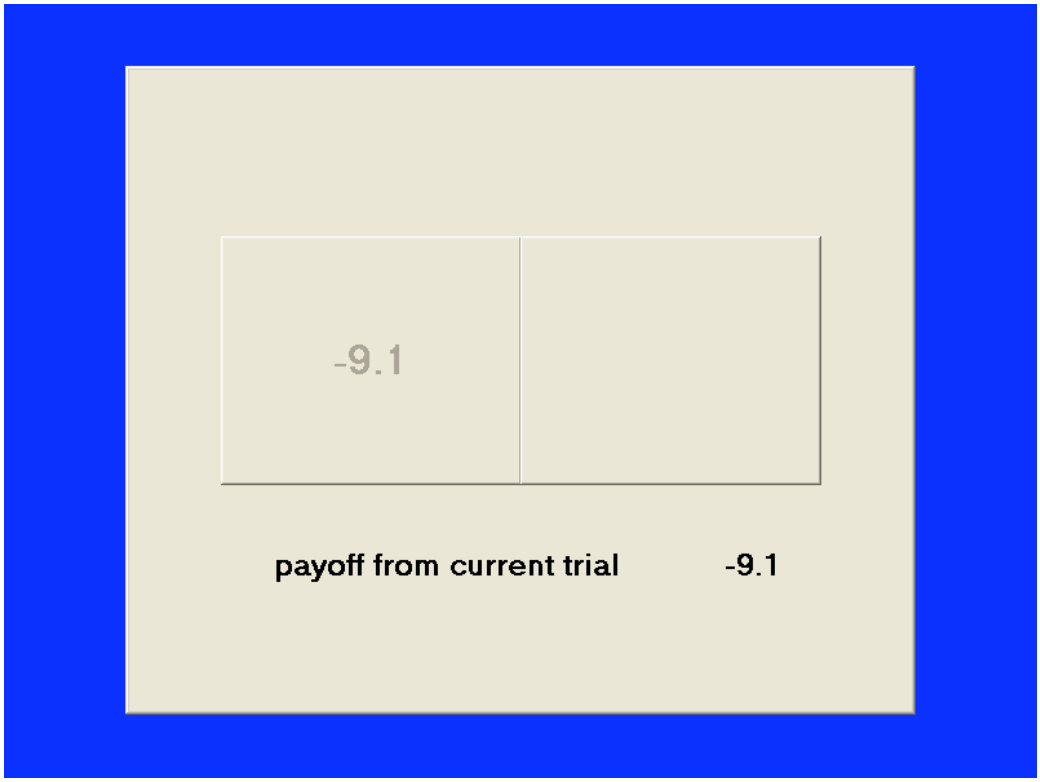
This experiment includes several games. Each game includes several trials. You will receive a message before the beginning of each game.

In each trial you will be asked to select one of two buttons. Each press will result in a payoff that will be presented on the selected button.

At the end of the experiment one of the trials will be randomly drawn (all the trials are equally likely to be drawn). Your payoff for the experiment will be the outcome (in Sheqels) of this trial.

Good luck!

Experimental Screen after choosing the risky alternative in Problem 36:



References

Amey, B. 2001. People's participation in and attitudes to gaming, 1985-2000: Final results of the 2000 survey. Welling-ton, New Zealand: Department of Internal Affairs. http://dia.govt.nz/Pubforms.nsf/URL/Titlecontentspt1.pdf

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Arifovic, J., McKelvey, R. D., & Pevnitskaya, S. (2006). An initial implementation of the Turing tournament to learning in repeated two-person games. *Games and Economic Behavior*, *57*, 93-122.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). New York: Kluwer.

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description based decisions. *Journal of Behavioral Decision Making, 16,* 215-233.

Biele, G., Erev I., & Ert, E. (2009). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of Mathematical Psychology, 53,* 155-167.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without tradeoffs. *Psychological Review*, *113*, 409-432.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, 78, 1–3.

Busemeyer, J. R., (1985). Decision making under uncertainty: A comparison of simple scalability, fixed sample, and sequential sampling models. *Journal of Experimental Psychology*, *11*, 538-564.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on the generalization criterion methodology. *Journal of Mathematical Psychology*, *44*, 171-189.

Carnap, R. (1953). On the comparative concept of confirmation. *British Journal for the Philosophy of Science, 3*, 311-318.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559–583.

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*, 523-538.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13*, 171-192.

Erev, I., & Barron, G**.** (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*, 912-931.

Erev, I., Bereby-Meyer, Y., & Roth A. E. (1999). The effect of adding a constant to all payoffs: Experimental investigation, and implications for reinforcement learning models. *Journal of Economic Behavior and Organizations*, *39*, 111-128.

Erev, I., Ert, E., & Roth, A. E. (2008). The Technion 1st prediction tournament. http://tx.technion.ac.il/~erev/Comp/Comp.html.

Erev I., Ert, E., & Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making, 21,* 575-597.

Erev, I., Glozman, I., & Hertwig, R. (2008). Context, mere presentation and the impact of rare events. *Journal of Risk and Uncertainty, 36*, 153-177.

Erev, I., & Haruvy, E., (2009). Learning and the economics of small decisions. forthcoming in J. H. Kagel & A. E. Roth (Eds.) *The Handbook of Experimental Economics*.http://www.utdallas.edu/~eeh017200/papers/LearningChapter.pdf

Erev, I., & Livne-Tarandach, R. (2005). Experiment-based exams and the difference between the behavioral and the natural sciences. In R. Zwick & A. Rapoport (Eds.), *Experimental business research, Vol 3*. (pp. 297-308). Dordrecht, The Netherlands: Springer.

Erev, I., Roth, A. E., Slonim, R. L., & Barron, G. (2002). Combining a theoretical prediction with experimental evidence**.** *http://papers.ssrn.com/abstract_id=1111712*.

Erev, I., Roth, A. E., Slonim, R. L., & Barron, G. (2007). Learning and equilibrium as useful approximations: Accuracy of prediction on randomly selected constant sum games. *Economic Theory*, *33*, 29-51.

Erev, I., Wallsten, T.S. & Budescu, D.V. (1994), Simultaneous over- and underconfidence: the role of error in judgment processes. *Psychological Review*, 101, 519-527.

Ert, E., & Erev, I. (2007). *Loss aversion in decisions under risk and the value of a symmetric simplification of prospect theory*. (Working Paper). Haifa, Israel: The Technion, Faculty of Industrial Engineering and Management.

Friedman, M., & Savage, L. (1948). The utility analysis of choices involving risk. *Journal of Political Economy, 56,* 279-304.

Gonzalez, C., Lerch, F. J., & Lebiere, C. (2003). Instance-based learning in real-time dynamic decision making. *Cognitive Science, 27*, 591-635.

Haruvy, E., Erev, I., & Sonsino, D. (2001). The medium prize paradox: Evidence from a simulated casino. *Journal of Risk and Uncertainty, 22*, 251-261.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008) The description-experience gap in risky choice: the role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21,* 493-518.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*, 534-539.

Hertwig, R., & Pleskac, T. J. (2008). The game of life: How small samples render choice simpler. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 209-236). Oxford, England: Oxford University Press.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: Selecting among forecasts and their combinations. *International Journal of Forecasting, 21*, 15–24.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263-291.

Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin, 125,* 524-543.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52*, 111–127.

Lebiere, C., & Bothell, D. (2004). Competitive modeling symposium: Pokerbot World Series. In *Proceedings of the 2004 International Conference on Cognitive Modeling*. Mahwah, NJ: Erlbaum.

Lebiere, C., Gonzalez, C., & Martin, M. (2007). Instance-based decision making model of repeated binary choice. *Proceedings of the 8th International Conference on Cognitive Modeling*. Ann Arbor, Michigan, USA

Lebiere, C., & West, R. L. (1999). *A dynamic ACT-R model of simple games*. In Proceedings of the Twenty-first Conference of the Cognitive Science Society, (pp. 296-301). Mahwah, NJ: Erlbaum.

March, J. G. (1996). Learning to become risk averse. *Psychological Review, 103*, 309-319.

Neugebauer, O., & Sachs, A. J. (1945). *Mathematical cuneiform texts.* New Haven.

Quiggin, J. (1991). On the optimal design of lotteries. *Economica, 58,* 1–16.

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes, 106*, 168-179.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 34,* 1446-1465.

Roth, A. E. (2008). What have we learned from market design? Hahn Lecture, *Economic Journal, 118*, 285-310.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics, 1*, 43-62.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153-81.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297-323.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science, 4*, 473-479.

Weber, E. U., Shafir, S., & Blais, A. R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*, *111*, 430-445.

West, R. L., Stewart, T. C., Lebiere, C., & Chandrasekharan, S. (2005). Stochastic resonance in human cognition: ACT-R vs. game theory, associative neural networks,

recursive neural networks, q-learning, and humans. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 2353-58). Mahwah, NJ: Lawrence Erlbaum Associates.

Yates, J. F. (1990). *Judgment and Decision Making*._Englewood Cliffs, NJ: Prentice Hall.

Authors' biographies:

Ido Erev is the ATS' Women's Division Professor of Industrial Engineering and Management at the Technion. His current research focuses on decisions from experience and the economics of small decisions.

Eyal Ert is a faculty fellow at the Harvard Business School.
His current research interests focus on models of learning and decision making, and their implications for everyday life, consumer behavior, and social interactions.

Alvin E. Roth is the Gund Professor of Economics and Business Administration at Harvard University. His research is in game theory, experimental economics, and market design. (See his home page at http://kuznets.fas.harvard.edu/~aroth/alroth.html).

Ernan Haruvy is an Associate Professor of Marketing at the University of Texas at Dallas. He received his PhD in Economics from the University of Texas at Austin. His research interests are in the application of models of human behavior to markets.

Stefan Herzog is a research scientist of Cognitive and Decision Sciences in the Department of Psychology at the University of Basel, Switzerland. His research focuses on bounded rationality and "The Wisdom of Crowds".

Robin Hau is a post-doctoral researcher of Cognitive and Decision Sciences in the Department of Psychology at the University of Basel, Switzerland. His research focuses on experience-based decisions and cognitive modeling.

Ralph Hertwig is a Professor of Cognitive and Decision Sciences in the Department of Psychology at the University of Basel, Switzerland. His research focuses on models of bounded rationality, social intelligence, and methodology of the social sciences.

Terrence Stewart is a post-doctoral researcher in the Centre for Theoretical Neuroscience at the University of Waterloo. His research involves the methodological issues surrounding cognitive modelling, and he currently applies this work towards developing neural models of high-level reasoning.

Robert West is an Associate Professor in the Institute of Cognitive Science and the Department of Psychology at Carleton University. His main research interest is computational cognitive architectures and their applications to psychology, human game playing, cognitive engineering, and work in sociotechnical systems.

Christian Lebiere is a research faculty in the Psychology Department at Carnegie Mellon University. His main research interest is computational cognitive architectures and their applications to psychology, artificial intelligence, human-computer interaction, decision-making, intelligent agents, robotics and neuromorphic engineering.

Table 1a: The 60 estimation set problems and the aggregate proportion of choices in risk in each of the experimental conditions.

	Risky			Safe	Proportion of Risky Choices (R - rate)			Average Number of
Problem	H	Ph	L	M	Description	E-Sampling	E-Repeated	Samples per Problem
1*	-0.3	0.96	-2.1	-0.3	0.20	0.25	0.33	10.35
2	-0.9	0.95	-4.2	-1.0	0.20	0.55	0.50	9.70
3	-6.3	0.30	-15.2	-12.2	0.60	0.50	0.24	13.85
4	-10.0	0.20	-29.2	-25.6	0.85	0.30	0.32	10.70
5	-1.7	0.90	-3.9	-1.9	0.30	0.80	0.45	9.85
6	-6.3	0.99	-15.7	-6.4	0.35	0.75	0.68	9.85
7	-5.6	0.70	-20.2	-11.7	0.50	0.60	0.37	11.10
8	-0.7	0.10	-6.5	-6.0	0.75	0.20	0.27	13.90
9	-5.7	0.95	-16.3	-6.1	0.30	0.60	0.43	10.95
10	-1.5	0.92	-6.4	-1.8	0.15	0.90	0.44	11.75
11	-1.2	0.02	-12.3	-12.1	0.90	0.15	0.26	11.90
12	-5.4	0.94	-16.8	-6.4	0.10	0.65	0.55	11.15
13	-2.0	0.05	-10.4	-9.4	0.50	0.20	0.11	10.35
14	-8.8	0.60	-19.5	-15.5	0.70	0.80	0.66	12.10
15	-8.9	0.08	-26.3	-25.4	0.60	0.30	0.19	11.60
16	-7.1	0.07	-19.6	-18.7	0.55	0.25	0.34	11.00
17	-9.7	0.10	-24.7	-23.8	0.90	0.55	0.37	15.10
18	-4.0	0.20	-9.3	-8.1	0.65	0.40	0.34	11.15
19	-6.5	0.90	-17.5	-8.4	0.55	0.80	0.49	14.90
20	-4.3	0.60	-16.1	-4.5	0.05	0.20	0.08	10.85
21	2.0	0.10	-5.7	-4.6	0.65	0.20	0.11	8.75
22	9.6	0.91	-6.4	8.7	0.05	0.70	0.41	9.15
23	7.3	0.80	-3.6	5.6	0.15	0.70	0.39	10.70
24	9.2	0.05	-9.5	-7.5	0.50	0.05	0.08	14.60
25	7.4	0.02	-6.6	-6.4	0.90	0.10	0.19	8.90
26	6.4	0.05	-5.3	-4.9	0.65	0.15	0.20	13.35
27	1.6	0.93	-8.3	1.2	0.15	0.70	0.50	8.90
28	5.9	0.80	-0.8	4.6	0.35	0.65	0.58	10.60
29	7.9	0.92	-2.3	7.0	0.40	0.65	0.51	10.60
30	3.0	0.91	-7.7	1.4	0.40	0.70	0.41	9.95
31	6.7	0.95	-1.8	6.4	0.10	0.70	0.52	11.00
32	6.7	0.93	-5.0	5.6	0.25	0.55	0.49	10.95
33	7.3	0.96	-8.5	6.8	0.15	0.75	0.65	11.10
34	1.3	0.05	-4.3	-4.1	0.75	0.10	0.30	11.35
35	3.0	0.93	-7.2	2.2	0.25	0.55	0.44	12.80
36	5.0	0.08	-9.1	-7.9	0.40	0.2	0.09	14.60
37	2.1	0.80	-8.4	1.3	0.10	0.35	0.28	10.90

38	6.7	0.07	-6.2	-5.1	0.65	0.20	0.29	10.90
39	7.4	0.30	-8.2	-6.9	0.85	0.70	0.58	12.65
40	6.0	0.98	-1.3	5.9	0.10	0.70	0.61	13.50
41	18.8	0.80	7.6	15.5	0.35	0.60	0.52	9.00
42	17.9	0.92	7.2	17.1	0.15	0.80	0.48	10.80
43*	22.9	0.06	9.6	9.2	0.75	0.90	0.88	9.90
44	10.0	0.96	1.7	9.9	0.20	0.70	0.56	10.05
45	2.8	0.80	1.0	2.2	0.55	0.70	0.48	19.40
46	17.1	0.10	6.9	8.0	0.45	0.20	0.32	9.15
47	24.3	0.04	9.7	10.6	0.65	0.20	0.25	11.80
48	18.2	0.98	6.9	18.1	0.10	0.75	0.59	9.00
49	13.4	0.50	3.8	9.9	0.05	0.45	0.13	8.85
50	5.8	0.04	2.7	2.8	0.70	0.20	0.35	9.95
51	13.1	0.94	3.8	12.8	0.15	0.65	0.52	8.95
52	3.5	0.09	0.1	0.5	0.35	0.25	0.26	11.85
53	25.7	0.10	8.1	11.5	0.40	0.25	0.11	9.00
54	16.5	0.01	6.9	7.0	0.85	0.25	0.18	13.40
55	11.4	0.97	1.9	11.0	0.15	0.70	0.66	9.55
56	26.5	0.94	8.3	25.2	0.20	0.50	0.53	14.25
57	11.5	0.6	3.7	7.9	0.35	0.45	0.45	10.00
58	20.8	0.99	8.9	20.7	0.25	0.65	0.63	12.90
59	10.1	0.30	4.2	6.0	0.45	0.45	0.32	10.10
60	8.0	0.92	0.8	7.7	0.20	0.55	0.44	10.20

Note - All problems involve binary choice between a sure payoff (M) and a risky option with two possible outcomes (H with probability Ph, L otherwise). For example, Problem 60 describes a choice between a gain of 7.7 Sheqels for sure, and a gamble that yields a gain of 8.0 Sheqels with probability of 0.92 and a gain of 0.8 Sheqels otherwise. The proportions of choices are over all 20 participants, and (in Condition E-Repeated) over the 100 trials. Problems with a dominant strategy (1 and 43) are marked with a star.

Table 1b: The 60 competition problems and the aggregated risky choices per problem.

	Risk			Safe	Proportion of Risky Choices (R - rate)			Average Number of
Problem	H	Ph	L	M	Description	E-Sampling	E-Repeated	Samples per Problem
1	-8.7	0.06	-22.8	-21.4	0.70	0.45	0.25	16.35
2	-2.2	0.09	-9.6	-8.7	0.60	0.15	0.27	15.65
3	-2.0	0.10	-11.2	-9.5	0.45	0.10	0.25	15.60
4	-1.4	0.02	-9.1	-9.0	0.85	0.20	0.33	15.90
5	-0.9	0.07	-4.8	-4.7	0.80	0.35	0.37	15.55
6	-4.7	0.91	-18.1	-6.8	0.50	0.75	0.63	14.75
7	-9.7	0.06	-24.8	-24.2	0.95	0.50	0.30	20.95
8	-5.7	0.96	-20.6	-6.4	0.35	0.65	0.66	15.85
9	-5.6	0.10	-19.4	-18.1	0.75	0.20	0.31	15.50
10	-2.5	0.60	-5.5	-3.6	0.45	0.50	0.34	17.15
11	-5.8	0.97	-16.4	-6.6	0.40	0.65	0.61	17.35
12	-7.2	0.05	-16.1	-15.6	0.75	0.40	0.25	16.85
13	-1.8	0.93	-6.7	-2.0	0.25	0.55	0.44	11.85
14	-6.4	0.20	-22.4	-18.0	0.70	0.15	0.21	12.05
15*	-3.3	0.97	-10.5	-3.2	0.10	0.10	0.16	18.20
16	-9.5	0.10	-24.5	-23.5	0.90	0.70	0.39	15.70
17	-2.2	0.92	-11.5	-3.4	0.25	0.65	0.47	14.70
18	-1.4	0.93	-4.7	-1.7	0.30	0.55	0.41	16.50
19	-8.6	0.10	-26.5	-26.3	0.90	0.60	0.49	16.25
20	-6.9	0.06	-20.5	-20.3	1.00	0.60	0.25	15.95
21	1.8	0.60	-4.1	1.7	0.05	0.10	0.08	10.80
22*	9.0	0.97	-6.7	9.1	0.00	0.15	0.14	14.85
23	5.5	0.06	-3.4	-2.6	0.40	0.20	0.28	18.05
24	1.0	0.93	-7.1	0.6	0.25	0.65	0.46	14.05
25	3.0	0.20	-1.3	-0.1	0.35	0.25	0.21	14.50
26	8.9	0.10	-1.4	-0.9	0.70	0.25	0.23	17.65
27	9.4	0.95	-6.3	8.5	0.20	0.55	0.67	13.25
28	3.3	0.91	-3.5	2.7	0.25	0.65	0.58	12.95
29	5.0	0.40	-6.9	-3.8	0.75	0.70	0.39	15.10
30	2.1	0.06	-9.4	-8.4	0.50	0.30	0.33	18.10
31*	0.9	0.20	-5.0	-5.3	1.00	0.95	0.88	14.80
32	9.9	0.05	-8.7	-7.6	0.65	0.30	0.21	19.70
33	7.7	0.02	-3.1	-3	0.90	0.35	0.28	15.95
34	2.5	0.96	-2.0	2.3	0.20	0.50	0.52	15.85
35	9.2	0.91	-0.7	8.2	0.15	0.60	0.56	14.70
36*	2.9	0.98	-9.4	2.9	0.00	0.35	0.34	18.15
37	2.9	0.05	-6.5	-5.7	0.60	0.35	0.30	15.30
38	7.8	0.99	-9.3	7.6	0.20	0.75	0.62	15.25
39	6.5	0.80	-4.8	6.2	0.00	0.35	0.32	11.00
40	5.0	0.90	-3.8	4.1	0.10	0.50	0.46	13.40
41	20.1	0.95	6.5	19.6	0.15	0.65	0.50	13.70
42	5.2	0.50	1.4	5.1	0.05	0.05	0.08	12.00
43	12.0	0.50	2.4	9.0	0.00	0.25	0.17	14.35
44	20.7	0.90	9.1	19.8	0.15	0.55	0.44	11.85

45	8.4	0.07	1.2	1.6	0.90	0.25	0.20	14.80
46	22.6	0.40	7.2	12.4	0.75	0.30	0.41	15.30
47	23.4	0.93	7.6	22.1	0.35	0.65	0.72	13.20
48	17.2	0.09	5.0	5.9	0.85	0.50	0.24	14.00
49	18.9	0.90	6.7	17.7	0.15	0.45	0.57	11.60
50	12.8	0.04	4.7	4.9	0.65	0.30	0.26	15.45
51	19.1	0.03	4.8	5.2	0.70	0.25	0.22	18.75
52	12.3	0.91	1.3	12.1	0.10	0.35	0.41	10.50
53	6.8	0.90	3.0	6.7	0.20	0.40	0.41	11.60
54	22.6	0.30	9.2	11.0	0.85	0.85	0.60	10.55
55	6.4	0.09	0.5	1.5	0.35	0.40	0.28	10.55
56	15.3	0.06	5.9	7.1	0.40	0.25	0.17	17.75
57	5.3	0.90	1.5	4.7	0.30	0.65	0.66	15.60
58	21.9	0.50	8.1	12.6	0.85	0.80	0.47	11.35
59	27.5	0.70	9.2	21.9	0.35	0.25	0.42	15.40
60	4.4	0.20	0.7	1.1	0.75	0.70	0.38	12.60

Problems with a dominant strategy (15, 22, 31 and 36) are marked with a star.

Table 2: The correlations between the R-rates (proportion of risky choices) in the different conditions using problem as a unit of analysis over the problems without dominant strategies in the estimation study (p-values in parentheses).

		Estimation set		Competition set	
		E-sampling	E-repeated	E- Sampling	E- Repeated
Problems without dominant choices	Description	-.53	-.37	.04	-.24
		(<.001)	(.004)	(.782)	(.081)
	E-sampling		.83		.76
			(<.001)		(<.001)
Problems with rare events (ph <.2 or ph >.8)	Description	-74	-66	-.33	-.60
		(<.001)	(<.001)	(.030)	(< .001)
	E-sampling		.84		.76
			(<.001)		(<.001)
Problems without rare events (.2 ≤ ph ≤ .8)	Description	.30	.51	.72	.74
		(.271)	(.051)	(.006)	(.004)
	E-sampling		.84		.83
			(<.001)		(<.001)

Table 3: Summary of the fit and prediction scores of the top three submitted models, and the most interesting baseline models in each competition. Pagree is the proportion of agreement between modal prediction and the modal choice, r is the Pearson correlation, MSD is mean squared deviation, ENO is the equivalent number of observations.

3a: Condition Description:

			Fitness Scores based on the Estimation Set (S^2=.1860)			Prediction Scores based on the Competition Set (S^2=.1636)			
Title	Team and idea	Parameters	Pagree	r	MSD	Pagree	r	**MSD**	ENO
Interesting Baselines	CPT	α=.7, β=1, λ=1, γ=δ=.65	95%	0.85	0.0930	93%	0.86	**0.0837**	2.16
	Priority	s=0.1	91%	0.76	0.1158	81%	0.65	**0.1437**	1.21
Best Baseline	SCPT with Normalization	α=.89,β=.98, λ=1.5,μ=2.15 γ=δ=.7	89%	0.92	0.0116	95%	0.95	**0.0102**	80.99
Winner	Haruvy: Logistic Regression	β0=1,β1=.01, β2=.07,β3=.41, γ1=1.42,γ2=.32 , γ3=-.621	88%	0.92	0.0099	90%	0.94	**0.0126**	56.36
Runner up	Yechiam: Version of SCPT with no Diminishing Sensitivity	w=.46, υ=.05, λ1=.61,λ2=.73, θ=1.14, ε=.76, Δ=.15, τ,=.8 α=.-10	92%	0.89	0.0141	91%	0.93	**0.0133**	31.95
Second Runner up	Ann & Picard: CPT with Aspiration Levels	β0=18.58,α=.61 β=.52 λ=1 wp=.99,wn=.99 γ=.25, γ1=.30 τ=.90	87%	0.93	0.0088	90%	0.92	**0.0165**	19.66
Intuition	Harvard students					83%	0.86	**0.1149**	1.88

			Fitness scores based on the estimation set (S^2=.2023)			Prediction scores based on the competition set (S^2 =.2111)			
Title	Team and Idea	Parameters	Pagree	r	MSD	Pagree	r	**MSD**	ENO
Interesting Baseline	Primed Sampler	k = 5	90%	0.81	0.0270	82%	0.79	**0.0251**	14.51
Best Baseline	Primed Sampler with Variability	k = 9	95%	0.88	0.0170	82%	0.80	**0.0244**	15.23
Winner	Herzog, Hau, Hertwig: Linear combination	α=1.19,β=1.35 γ=1.42,δ=1.54 λ=1.19; μ=.41, σ =.037, T_o= .0001,T_p=0.11,p_{order1} =.38, k = 9, and N*	95%	0.92	0.0099	83%	0.80	**0.0187**	25.92
Runner Up	Ann, Picard: Sample by CPT and Aspiration Levels	x=2.07,y=1.31 z=0.71,v=7.53 r=12.64,m=0.02	92%	0.90	0.0115	82%	0.82	**0.0203**	21.66
Second	Hau, Hertwig:	N*	95%	0.89	0.01548	82%	0.79	**0.0250**	14.61

3b: Condition E-Sampling.

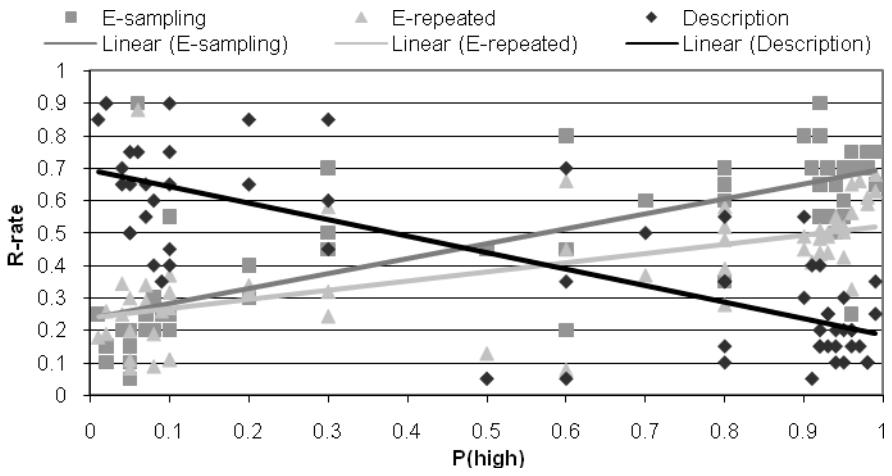
*N is a vector with 20 elements that determines the probability of the different sample sizes. It was estimated by the number of samples taken by participants in the estimation set.

3c: Condition E-Repeated:

			Fitness scores based on the estimation set(S^2 = .0875)			Prediction scores based on the competition set (S^2 = .0928)			
Title	Team and Idea	Parameters	Pagree	r	MSD	Pagree	r	**MSD**	ENO
Interesting Baselines	Normalized Reinforcement Learning	w=.15, λ=1.1	76%	0.83	0.0092	84%	0.84	**0.0087**	22.89
	Basic Reinforcement Learning	w=.15, λ=1	56%	0.67	0.0224	66%	0.51	**0.0263**	4.28
Best Baseline	Explorative Sampler with Recency	β=.10, w=.3, ε=.12, k=8	82%	0.88	0.0075	86%	0.89	**0.0066**	47.22
Winner	Stewart, West, & Lebiere: ACT – R with Sequential Dependencies and Blending Memory	s = .35, *τ*= -1.6	77%	0.88	0.0094	87%	0.89	**0.0075**	32.50
Runner Up	Hochman & Ayal: Two-Stage Sampler	w=1, δ=.55;β=.1 , f=.3;κ=6; ε=.11, r=.01;	80%	0.90	0.0065	83%	0.87	**0.0084**	24.71
Second Runner Up	Haruvy: NRL with Inertia	w=.14, λ=1.05 q=0.32 ,i=0.5;	75%	0.86	0.0080	86%	0.85	**0.0084**	24.71

Figure 1: R-rate (proportion of risky choices) as a function of Ph (the probability of getting the high outcome from the risky gamble) in each of the three experimental conditions.

1a: Estimation Study



1b: Competition Study

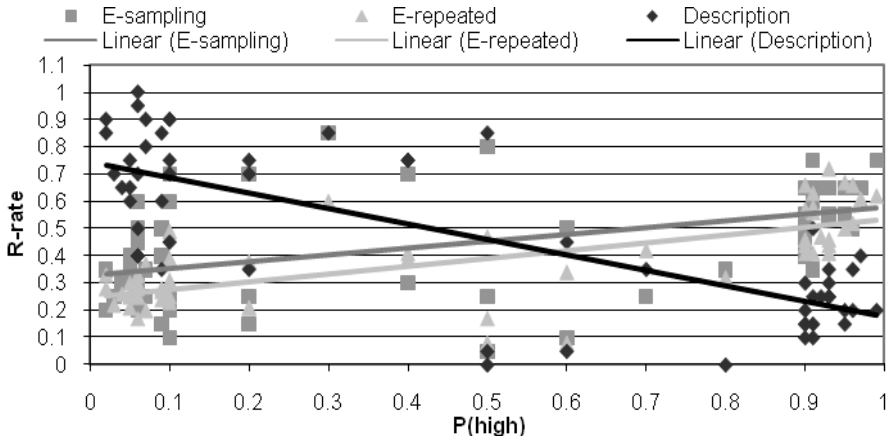
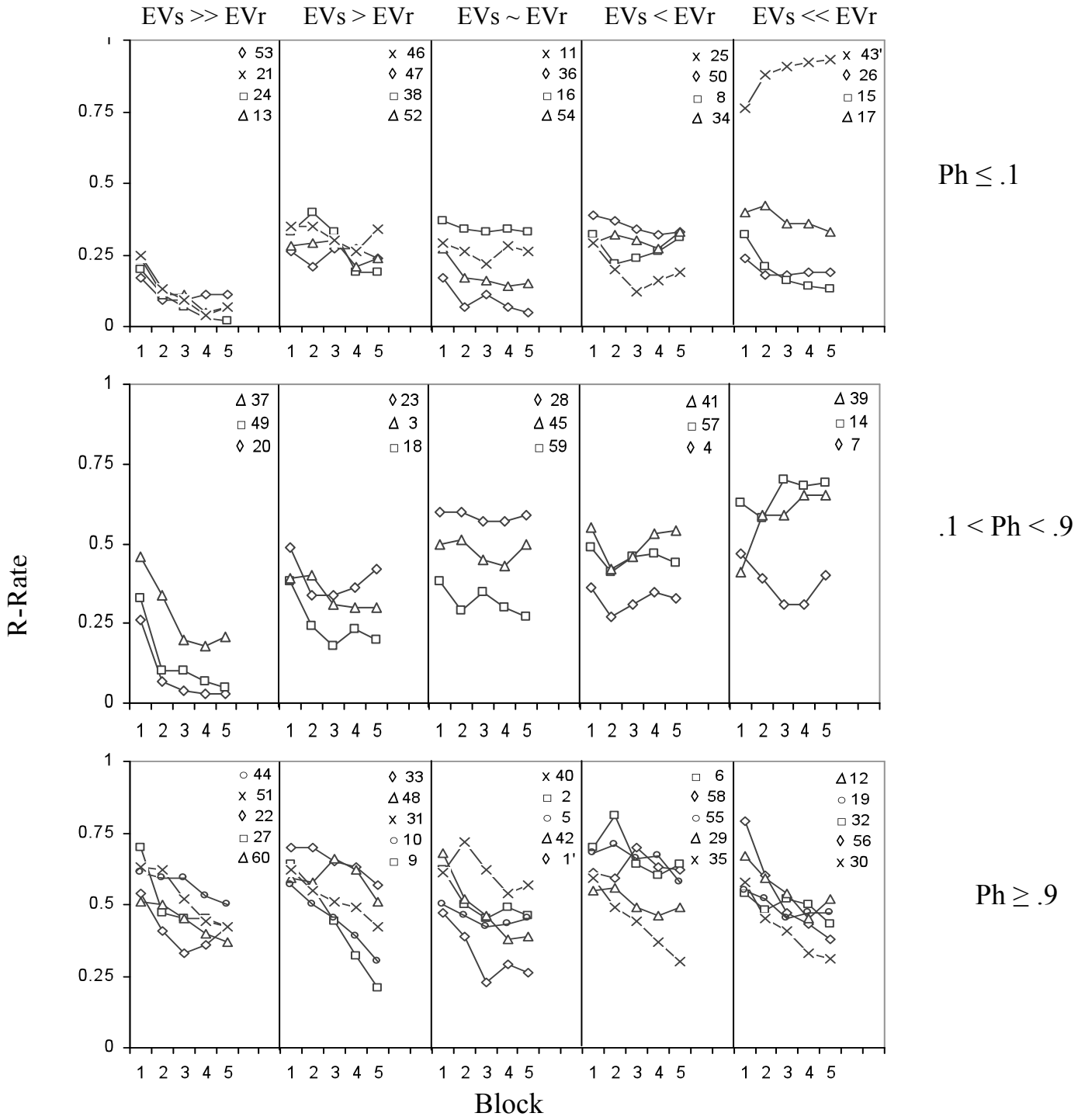
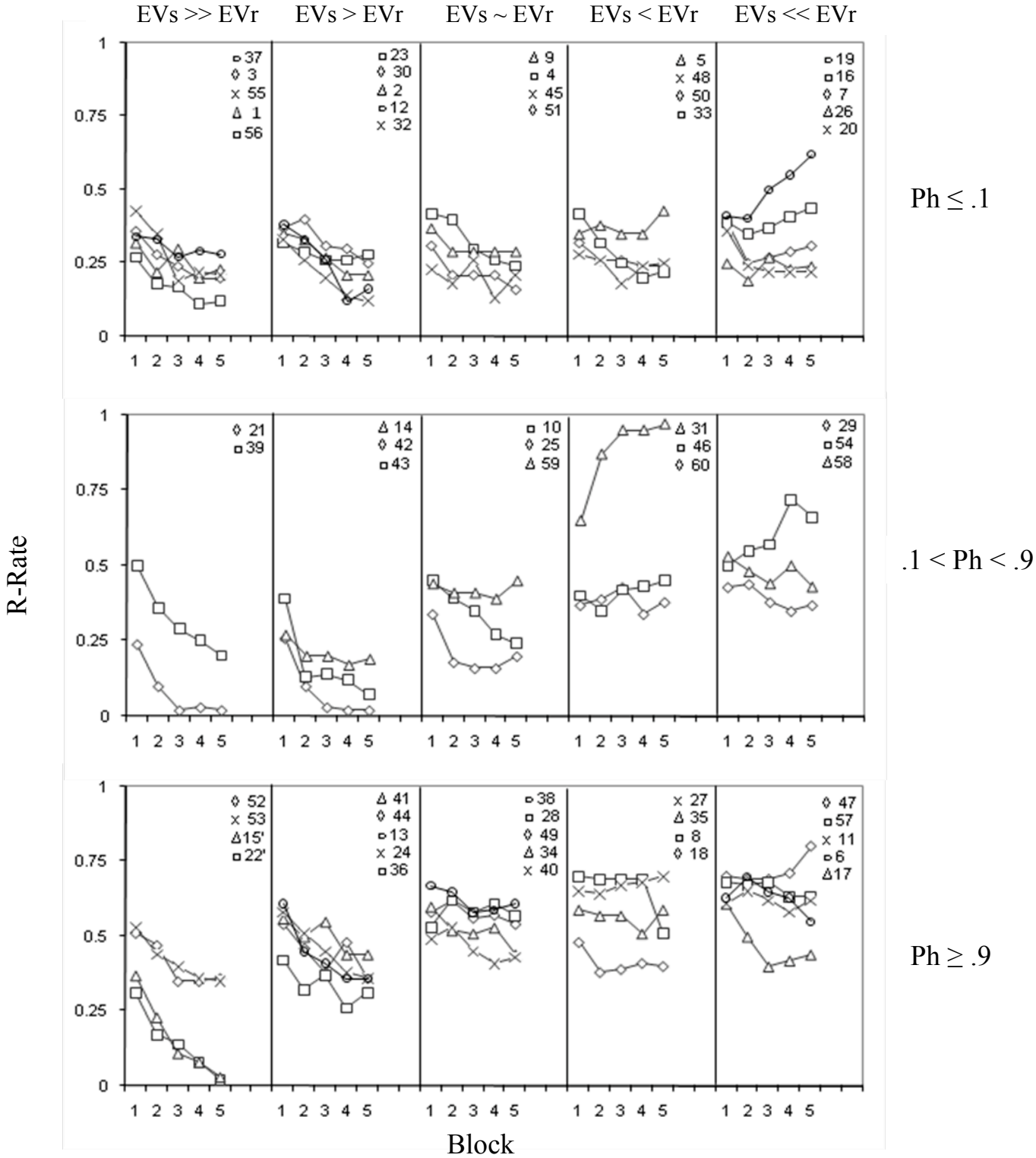


Figure 2: observed R-rates (proportion of risky choices) in Condition E-repeated in 5 blocks of 20 trials. The 60 problems were classified to 15 graphs according to (a) The probability of high payoff (Ph) and the relative expected value of the risky prospect (EVr-EVs).

2a: Estimation Study



2b: Competition Study



NOTE – The numbers in the legends are the problem id. In the tagged problems (e.g., 15 and 22 in the lower left cell of Figure 2a) one alternative dominates the other.