Limits on Predictability of Risky Choice Behavior

Anjali Sifar (sanjali@iitk.ac.in)
Cognitive Science
IIT Kanpur, UP 208016 India

Nisheeth Srivastava (nsrivast@cse.iitk.ac.in)
Computer Science
IIT Kanpur, UP 208016 India

Abstract

Research in decision-making has recently begun to emphasize predictive accuracy as the dominant principle for designing and evaluating choice models. This emphasis has led to the development of increasingly more precise models of humans' risk preferences, as measured in certain experimental paradigms built upon certainty equivalence testing. In this paper, we argue that the level of precision attained by recent choice models is unexpected, because human preferences are irreducibly noisy. We support this argument by conducting experiments to measure intra-observer consistency in choice behavior in two common risk preference paradigms: decisions from description and experience. We find that while current choice models of decisions from experience align fairly well with upper limits of choice consistency seen in our experimental data, choice models for decisions from description are significantly more consistent with humans' choices than humans themselves are consistent with their own choices. We discuss some theoretical and practical implications of our results.

Keywords: risk preferences; predictability; decisions from experience; choice modelling

Introduction

The certainty equivalence paradigm for measuring risk preferences is one of the workhorses of behavioral economics research (Farquhar,1984). A typical certainty equivalence task seeks to elicit the lowest certain amount that someone might prefer over a given risky gamble. Beginning with Erev et al. (2010), variants of this task have been developed and studied using recurring choice prediction *tournaments*. The primary ambition of these tournaments is to potentiate the development of models that can make accurate quantitative predictions for risky choice behavior, including the reproduction of classic anomalies previously reported in the behavioral decision theory literature (Erev, Ert, Plonsky, Cohen, & Cohen,2017).

While models of human decisions have historically been assessed using a mix of qualitative insights and quantitative tests, prediction tournaments have focused on making quantitatively precise predictions to the exclusion of other possible criteria for assessing the feasibility of models (Erev et al.,2017). Tournaments are conducted by allowing teams to fit choice models to human choices made on some certainty equivalence problems, and winning models are identified as the ones that most accurately predict human choices for a different set of problems. This paradigm aligns quite well with how supervised classification algorithms are trained from data (Bishop,2006). Perhaps as a consequence, machine learning models are now both competing and collaborating with theory-driven models in more recent prediction tournaments with excellent empirical success (Bourgin, Peterson, Reichman, Griffiths, & Russell,2019).

The empirical success of this research program, given its pure predictive emphasis, is measured in terms of the correlation of model predictions with human choices. Choice models developed through these tournaments have gone from explaining about 70% of the variance in human choices, as in the baseline models used in Erev et al. (2010) to explaining more than 90% of the variance in human choices, as in the BEAST model presented in (Erev et al.,2017). Machine learning models built using features identified as important by BEAST are able to approach test set values even more closely (Bourgin et al.,2019).

However, this empirical success is more than a little surprising, given the irreducibly stochastic nature of risky choices (Bhatia & Loomes,2017). If someone asks you to either pick 20 tokens of cash for certain or a gamble that will pay 100 tokens 20% of the time, it is very likely that your response may vary across multiple elicitations (Luce, Suppes, et al.,1965). Thus, if someone uses one of these elicitations to construct a dataset to fit a theory of decisions under risk, one would expect that the theory would not be complete. In plain language, given the intuitively fickle nature of human choices, can we actually expect choice models to predict them so well? This is the question we ask and try to answer in this paper.

How predictable are risky choices?

To obtain a direct empirical upper bound for predictability of risky choices, we ask how well observers' behavior when presented with a particular choice problem predicts the same observers' behavior when presented with the same choice problem again, controlling for memory-based consistency effects. In effect, we have experiment participants act as models of choice for their own behavior and see how well this model does, as measured by standard metrics used in choice prediction tournaments.

Method

Design

A set of expectation-matched risky choice problems were presented to each participant, at a gap of at least a week over the course of three weeks, following the protocol schematized in Figure 1. Two experiments were conducted, testing for choice consistency in decisions from description and experience respectively.

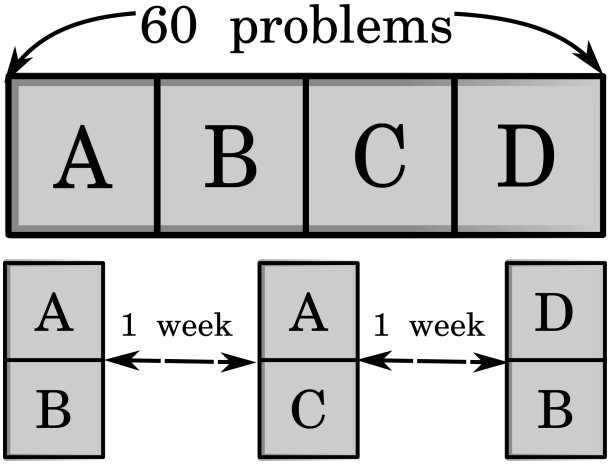

Figure 1: **Experiment Design** For each participant, problem space is randomly divided into four equal subsets.Half of the problems presented in Week 1 are repeated in Week 2 (subset A), and the other half in Week 3 (subset B). Week 2 and 3 repeated problems are interspersed with remaining fresh problems (subsets C and D).

In each experiment, as shown in Figure 1, each participant solves 30 problems per week, with half the problems seen in Week 1 being repeated during Week 2, and the other half repeating in Week 3. Throughout this paper, we refer to the first instances of problem presentation for a given participant as fresh problems, while second presentations will be called repeated problems.

The problem space used in our experiments is the Estimation set in the first Technion Choice Prediction Tournament consisting of 60 problems (Erev et al.,2010). We hosted the experiments online and participants were able to participate at their convenience. Participants were recruited from the IIT Kanpur campus where the study was conducted. Email reminders were sent every week to all participants. The study protocol was reviewed and approved by an IRB.

Decisions from Description

In decisions from description (DFD), for each problem, participants were asked to choose between risky and safe choices, given explicit payoff and probability descriptions. Participants responded to 30 unique problems each day, with each problem presented only once on any given day. Problem order was randomized across participants, and within participants for repeat presentations as indicated in Figure 1.

A total of 58(19 female ,39 male) participants completed the experiment, without compensation. Following the protocol used in Erev et al. (2010) we presented no outcome feedback following choice selections. However, feedback was provided for one of the randomly selected problems at the end of each day of the experiment as a notional payoff.

The one-shot DFD paradigm used is identical to the one used in Erev et al. (2010). This also parallels other large-scale risk preference elicitation protocols (Bruhin, Fehr-Duda, & Epper,2010). However, subsequent prediction tournaments have used a modified version of this paradigm. In these tournaments, participants respond to a choice problem multiple times in the same sitting, with the first few trials conducted without feedback, and the remaining trials conducted with feedback about both payoffs received and foregone after each choice (Erev et al.,2017;Plonsky et al.,2019;Bourgin et al.,2019).

Choice Models In the Technion competition, the interesting baseline model is Cumulative Prospect Theory (CPT) given by Tversky and Kahneman (1992) suggesting that the decision makers choose the prospect with the highest subjective probability weighted value.

As we mentioned above, the winner of the first DFD tournament was a logistic regression model (Erev et al.,2010). The logistic choice rule predicts the proportion of risky choices based on a linear relationship with the predictor variables - which in this case were the parameters of the problem and the expected value difference.

A special model has been designed for the special paradigm of decisions from description with feedback (Erev et al.,2017). This complex model attempts to computationally unite dynamic expected utility estimation with stochastic implementations of four cognitive biases. The resulting Best Estimate and Sampling Tools (BEAST) model was used as a baseline model for the fourth and fifth prediction tournaments, and has proved extremely difficult to beat, with tournament winners being mostly minor variants of BEAST, and performing statistically identically (Erev et al.,2017;Plonsky et al.,2019).

Decisions from Experience

In decisions from experience (DFE), we instantiated the sampling condition of the Technion tournament (Erev et al.,2010). A total of 25 male and 22 female participants completed the experiment. The experiment was presented to the participant as a series of games representing each problem - the parameters of which were derived from the problem space as described above. Instructions were followed by two practice games which were played under the guidance of the experimenter to ensure that the participant understood the game. The actual experiment started after the participant consented to continue the experiment, after playing the practice games.

For each game, the participant was able to view two buttons corresponding to safe and risky choices respectively. In the sampling stage, clicking on any one of the buttons, one at a

time, revealed one outcome for that option, sampled from a Bernoulli trial corresponding to the conditions of the gamble. These sampling trials were inconsequential, and participants were free to sample as many times as they wanted. Once they had sampled sufficiently many outcomes, they explicitly indicated a desire to make a final consequential selection with a button press. In this selection stage, they clicked on any of the outcomes once, and this outcome was considered the final outcome of the game. The participation fee was INR 100. However, a random game's outcome was selected at the end of three sessions which was scaled such that the final payoff bracketed in INR 0 to 200.

Choice models The best baseline model for decisions from experience in the first prediction tournament was a primed sampler that draws ν samples from the gamble, where ν is uniformly distributed from 1 to 9, and selects the option which has the greater average value based on these sampled values (Erev et al.,2010).

The winning model in this competition was an ensemble model which makes decisions by sampling one of four equally weighted decision rules (Erev et al.,2010). Of these, the first decision rule is the baseline model as described above. The second decision rule is a variant of the first rule where ν is drawn from the observed distribution of sample sizes in the observed data, upper-bounded at 20. The third decision rule is a stochastic cumulative prospect theory model. The final rule is a stochastic implementation of the priority heuristic (Brandstätter, Gigerenzer, & Hertwig,2006).

Response Variables

For every problem presented in both paradigms above, participants make a binary decision between a risky prospect and safe prospect. Each participant's response to each problem is recorded. Additionally, the proportion of participants taking the risky alternative for every problem is represented as the problem's Risky Choice Rate (R-rate). We record R-rates for all problems in both decisions from description and experience.

Decisions from experience, however, involve another latent decision of when to stop sampling. Measuring the consistency of this additional decision is also potentially of interest for informing models of information search within the context of decisions from experience (Hills & Hertwig,2010;Markant, Pleskac, Diederich, Pachur, & Hertwig,2015;Srivastava, Muller-Trede, Schrater, & Vul,2016). To this end, we also record the number of samples the decision maker takes before committing to a final choice (henceforth sampling duration).

Finally, in decisions from experience, an observer is presented with two choices that can altogether result in any of three unique payoffs. Observers that terminate information search before seeing each of the three possible outcomes at least once will make their final choice without actually understanding the problem structure. The minimum number of trials an observer would expect to make to see three unique outcomes is three. So, to obtain a clearer view of on-task behavior in DFE, we also separately report our metrics for all observations with sampling duration greater than 2.

Measuring choice consistency

If risk preferences have low inherent stochasticity at the cohort level, we expect the R-rate (relative number of times the risky option is selected by participants) for a problem to be consistent across repeated elicitations. To quantify this consistency, we compute the correlation between observed and predicted R-rates across all tested problems. In the special case of repeated problems, using the R-rates seen in the second elicitation as predictors for the first week's values yields a simple consistency measure. This measure is additionally attractive for offering a direct interpretation in terms of percentage of variance explained (Erev et al.,2010).

We also report the proportion of agreement P_{Agree}, as calculated in Erev et al. (2010), as an additional cohort-level measurement of consensus in choices. This is set to 1 for a problem if both predicted and observed R-rates are greater than or less than 0.5; otherwise it is set to zero. We report this value, averaged across all tested problems, in percentage terms, following convention (Erev et al.,2010).

If risk preferences have low inherent stochasticity at the individual level, we expect participant responses to the same problem to be consistent across repeated elicitations. We measure this individual-level intra-rater reliability using Cohen's κ (Landis & Koch,1977). For our context with agreement to be measured only for binary choices, this is simply

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the number of consistent choices made by an individual across all repeated problems divided by the total number of choices made by that individual during any one presentation of these problems and p_e is the probability of random agreement, calculated using the base risky choice proportions calculated within participants. We report median intra-rater reliability across participants, unless stated otherwise.

Results

In all analyses reported below, we refer to our data sources as follows. Data collected in our experiments are denoted as coming from RR (Repeated Risk). Data from the training datasets from the first Technion tournament will be referred to TE (Technion Estimation), and from the competition datasets from the same tournament as TC (Technion Competition). We use both DFD and DFE datasets from all these data sources. Data from the mixed paradigm used in Erev et al. (2017) will be referred to using EEP.

Predicting choice proportions

Decisions from description Table 1 summarizes consistency metrics calculated for our DFD experiment's data, presented alongside human-model consistency metrics calcu-

Table 1: **DFD Analysis** Choice agreement indicators to check the test-retest reliability between the repeated problems in the first row. Remaining rows indicate similar analysis using different models as predictors on the observed data sources.
WM = Winning Model (Logistic Regression), BM = Baseline Model (CPT) (Erev et al.,2010), BEAST = Baseline Model (Erev et al.,2017).
*claimed in (Erev et al.,2010). ** claimed in (Erev et al.,2017). ' Fresh Presentation of the problems.

Data Sources	Subset	Observed	Predictor	Correlation	κ	P_{Agree}
RR	A	Week 1	Week 2	**0.78**	0.25	85%
	B	Week 1	Week 3	**0.67**	0.25	78%
RR	Fresh	A', B', C, D	BM	0.69	0.23	81%
TE	-	TE	BM	0.85*	0.43	95%*
			WM	0.92*		88%*
TC	-	TC	BM	0.86*	0.51	93%*
			WM	0.94*		90%*
EEP	-	EEP	BEAST	0.95**	-	-

lated on our data as well as reported previously on existing datasets. Three observations stand out as particularly salient.

First, when we use the cohort's R-rate calculated during repeated presentations of problems to predict their own R-rate during the first presentation of the same problems, we obtain correlations ranging between 0.67-0.78 and proportions of agreement ranging between 78%-85%. Notably, these values are much lower than corresponding values previously found for model-data comparisons, even for baseline models like cumulative prospect theory (Erev et al.,2010).

Second, the baseline model does not perform expectedly well when we tested it on our data, yielding correlation and agreement proportions of 0.69 and 81% respectively. The problems used in our DFD experiments were identical to the ones in the TE dataset, and we applied the model using parameters estimated on the TE dataset (reported in Erev et al. (2010)). Thus, the large difference between the model's performance on the TC dataset and ours is surprising, and warrants further investigation.

Third, very little individual level intra-level agreement is seen across participants (Landis & Koch,1977), with $\kappa = 0.25$ in both repeated sets. This is in contrast with the baseline model, which agrees with the human choice data consistently more while predicting the Technion data sets $\kappa \in \{0.43, 0.51\}$ than human choices in our dataset agrees with itself. To add to the puzzle, when applied to our dataset, the model agrees with the data about as much ($\kappa = 0.23$) as expected by the empirical measurements.

Decisions from experience Table 2 summarizes consistency metrics calculated for our DFE data, presented alongside human-model consistency metrics for all datasets.

The main observation here is that the range of human-human correlations and agreement proportions seen in our data includes the corresponding model-human measurements reported on the competition set in Erev et al. (2010), though not similar measurements seen on the estimation set. Since the winning ensemble model in this tournament was not significantly better than the simple primed sampler baseline, our observation is consistent with the possibility that a simple primed sampler model might be the best possible model for predicting R-rates in the decision from experience task.

This possibility is also supported by an additional analysis. Separate from the values reported in Table 2, we separately calculated the intra-rater reliability of our participants on the subset of problems where the expected value difference between the observed practice sequences was almost identical (within 5% of the smallest payoff outcome in the dataset) across the two presentations. In contrast with the moderate values $\kappa \in \{0.4, 0.32\}$ seen for the full set of problems, we find high reliability $\kappa \in \{0.73, 0.78\}$ on this subset of problems across observers. That is, when the same people observe the same expected value differences again, they make the same choices, consistent with the decision criteria of the simple primed sampler model. This finding also offers a possible explanation for the gap between the empirical and model κ seen in DFE. The baseline primed sampler does not take observation history into account and so performs worse than humans themselves in predicting their prior choices.

Finally, as noted above, the DFE paradigm actually involves two decisions per problem presentation - an overt risk preference, and a latent information search stopping decision governing when to stop sampling and make a final choice. As shown in Table 3, human-human correlations for sampling duration in repeated problems for all observations is 0.64, dropping to 0.54 when only observations with sampling duration greater than two are considered. These values indicate reasonable upper bounds on the predictability of sampling duration in decisions from experience.

Interestingly, this limit is approached by a recent trial-by-trial sampling duration model that incorporates the influence of expected value difference, order-dependent variability in observation sequences, and the expectation of seeing all three outcomes at least once before committing to a decision in predicting sampling duration in such decisions from experience (Srivastava et al.,2016).

Table 2: **DFE Analysis** Choice agreement indicators to estimate test-retest reliability between repeated problems in the first row. Remaining rows indicate similar analysis using different models as predictors for observed choices. For each indicator, we also separately report values for observations with sampling duration greater than 2.
WM = Winning Model (ensemble) and BM = Baseline Model (primed sampler with variability) in Erev et al. (2010).
*claimed in (Erev et al.,2010). 'Fresh Presentation of the problems

				Correlations		κ		P_{Agree}	
Data Sources	Subset	Observed	Predictor	All	SD >2	All	SD >2	All	SD >2
RR	A	Week 1	Week 2	**0.83**	**0.80**	0.40	0.41	90%	85%
RR	B	Week 1	Week 3	**0.73**	**0.71**	0.32	0.27	88%	85%
RR	Fresh	A′, B′, C, D	BM	0.84	0.83	0.23	0.25	92%	92%
TE	-	TE	BM	0.88*	-	0.25	0.25	95%*	-
			WM	0.92*	-	-	-	95%*	-
TC	-	TC	BM	0.80*	-	0.19	0.2	82%*	-
			WM	0.80*	-	-	-	83%*	-

Table 3: **Sampling Duration Correlation for human model.** Correlation values are for each problem and participant pair.

Dataset	All observations	SD >2
A	0.64	0.54
B	0.65	0.56

Discussion

We measured the test-retest consistency of response choices in certainty equivalence experiments by correlating the decision-related behavior for the same problem by the same participant, separated by over a week in two standard risky choice paradigms.

By doing so, we fulfilled two inter-related goals. One, we obtained a direct characterization of the degree of natural variability in human observers' revealed preferences in certainty equivalence experiments as currently conducted, previously hinted at theoretically as in Bhatia and Loomes (2017). Two, we establish predictive upper bounds for the expected accuracy of cognitively realistic models of humans' risky choices.

For decisions from description, we found that participants' own future choices predicted at most about 60% of the variability in their previous choices on the same choice problems, and suggest this as a reasonable upper bound for achievable prediction performance for realistic choice models of one-shot decisions from description. We note that our measured intra-observer choice consistency values are considerably lower than model-human consistency achieved by contemporary models of risky choice behavior (Erev et al.,2010,2017). We find this unexpected excess predictability of choice models at both cohort and individual levels in tournament data sets, but crucially, not in our own data.

The pattern of results seen in our DFD analysis appears to be most consistent with the conclusion that earlier choice models for such tasks have been subtly over-fit to the validation set, a problem endemic to prediction tournaments with leaderboards (Dwork et al.,2015).

For decisions from experience, we found that participants' own future choices again predicted at most about 70% of the variability in their previous choices on the same choice problems. Unlike in the case of decisions from description, we found substantial agreement in the variance in responses captured by a primed sampler model presented in (Erev et al.,2010) with our upper bound estimates, suggesting that the primed sampler model is already close to optimal prediction performance on this task. We also found additional evidence supporting the use of an expected value difference criterion for deciding such decisions from experience, further supporting the plausibility of the primed sampler as a near optimal model of the DFE task.

How do these results affect the mixed paradigm of repeated DFD with feedback? We have shown that one-shot DFD responses greatly separated in time have high variability. It is unclear whether response variables collected in the mixed paradigm will have lower variability (Plonsky et al.,2019;Bourgin et al.,2019). It could, if the source of behavioral variability is short-term noise in the decision process, but wouldn't if the source of variability is longer timescale fluctuations in retrieval patterns from long-term memory (Beck, Ma, Pitkow, Latham, & Pouget,2012). We intend to investigate this question in future work.

References

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Bhatia, S., & Loomes, G. (2017). Noisy preferences in risky choice: A cautionary note. *Psychological review*, *124*(5), 678.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bourgin, D. D., Peterson, J. C., Reichman, D., Griffiths, T. L.,

& Russell, S. J. (2019). Cognitive model priors for predicting human decisions. *arXiv preprint arXiv:1905.09397*.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychological review*, *113*(2), 409.

Bruhin, A., Fehr-Duda, H., & Epper, T. (2010). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica*, *78*(4), 1375–1412.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). Generalization in adaptive data analysis and holdout reuse. In *Advances in neural information processing systems* (pp. 2350–2358).

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, *124*(4), 369.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., et al. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15–47.

Farquhar, P. H. (1984). State of the art—utility assessment methods. management science. *Management Science*, 1283-1300.

Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological science*, *21*(12), 1787–1792.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Luce, R. D., Suppes, P., et al. (1965). Preference, utility, and subjective probability.

Markant, D., Pleskac, T. J., Diederich, A., Pachur, T., & Hertwig, R. (2015). Modeling choice and search in decisions from experience: A sequential sampling approach. In *37th annual meeting of the cognitive science society* (pp. 1512–1517).

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., et al. (2019). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*.

Srivastava, N., Muller-Trede, J., Schrater, P., & Vul, E. (2016). Modeling sampling duration in decisions from experience. In *38th annual meeting of the Cognitive Science Society.*

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.