Over-precise predictions cannot identify good choice models

Anjali Sifar[1] and Nisheeth Srivastava[1,2*]

[1*]Cognitive Science, IIT Kanpur, Kanpur, 208016, Uttar Pradesh, India.

[2*]Computer Science, IIT Kanpur, Kanpur, 208016, Uttar Pradesh, India.

*Corresponding author(s). E-mail(s): nsrivast@iitk.ac.in;
Contributing authors: sanjali@iitk.ac.in;

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 illusory, because human preferences are irreducibly noisy, so that overly precise predictions are unlikely to reflect reality. We support this argument by measuring 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 the level 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 also found that models of decisions from description generalize poorly across sessions of choice responses from the same participants, whereas models of decisions from experience generalize quite well. A historical model comparison of 16 influential risky choice models on our dataset reveals that it is very difficult to go beyond simple expected utility models in predictive ability for aggregate measures of risky choices from description. We discuss some theoretical and practical implications of our results.

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

1 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, Russell, & Griffiths, 2019; He, Analytis, & Bhatia, 2021; Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021).

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; Peterson et al., 2021).

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, 1965). Then, if someone uses one of these elicitations to construct a dataset to fit a theory of decisions under

risk, the theory would be unable to account well for one out of the two behavioral instances. In plain language, given the intuitively fickle nature of human choices, can we actually expect choice models to predict them as well as they presently seem to be doing? This is the question we ask and try to answer in this paper.

While it has been historically evident intuitively that there is considerable variability in peoples' choices, recent research has begun to quantify the extent of this variability in risky choices. Fudenberg, Kleinberg, Liang, and Mullainathan (2019) use certainty equivalent judgment data on a 100 point scale for risky choices from Bruhin, Fehr-Duda, and Epper (2010) to show that cumulative prospect theory (CPT) is able to accommodate almost all the reducible error in this dataset, such that any predictive model that is operating using the same data features as CPT, viz. each individual problem's description, should not be able to reduce the MSE more than about 5% below CPT results. In contrast, state-of-the-art predictive choice models obtain an MSE reduction of nearly 50% compared to CPT (He et al., 2021; Peterson et al., 2021).

This paradox of excess predictability is the target of our examination in this paper. Since Fudenberg et al. (2019) work with certainty equivalent judgments while predictive models tend to work with aggregates of binary choices, their findings do not directly constrain the high levels of model-data correlation seen in recent choice models, which tend to target revealed preferences for options, averaged across individuals. For a more direct comparison, we design and implement a test-retest paradigm of risky choice behavior, following an experimental paradigm identical to the one used in Erev et al. (2010).

Fudenberg et al. (2019) demonstrate that no possible model that sees certain problem parameters could perform better than simply looking up entries from a table containing past certainty equivalent responses for a particular problem (described using the same parameters) for each participant, and use this prediction limit to argue that cumulative prospect theory is already near-optimal in predicting certainty equivalent judgments. Since we analyze aggregate choices rather than certainty equivalent judgments, we interpret the correspondence between human choice proportions made for problems in one session and choices made for the same problems in a different session as a similar benchmark for aggregate measures of risky choice behavior. By treating the choices elicited in the first session as an ideal model for themselves, we measure the correspondence such a perfect model would have with the data, if it had been collected on a different day.

2 Methods

2.1 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.

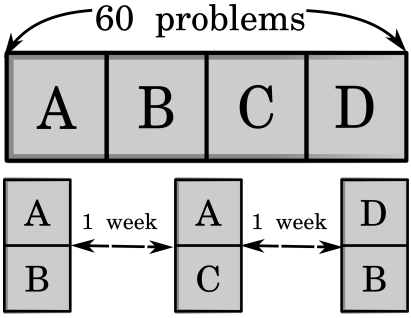

Fig. 1 Experiment Design For each participant, problem space is randomly divided into four equal subsets.Half of the problems presented in Week 1 are repeated in Week 2 (subset A), and the other half in Week 3 (subset B). Week 2 and 3 repeated problems are interspersed with remaining problems (subsets 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 will refer to data from our own experiment as Repeated Risk (RR), in the specified context of decisions from description or experience. We hereby specify labels for different subsets of our data for easier navigation. A subset with only fresh presentation of all problems is RR-F $\in \{A_1,B_1,\text{C},\text{D}\}$, a subset of problems with repeated elicitations is RR-R $\in \{A_1,B_1,A_2,B_2\}$. Further, the first presentation of repeated problems is called RR1 $\in \{A_1,B_1\}$ while the second presentation of repeated problems as RR2 $\in \{A_2,B_2\}$. The subset of problems which were not repeated, are RR0 $\in \{\text{C},\text{D}\}$.

The problem space used in our experiments is the Estimation set in the first Technion Choice Prediction Tournament (TCPT2010) 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 university campus where the study was conducted. Email reminders were sent every week to all participants. The study protocol was reviewed and approved by a university IRB.

2.2 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 et al., 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 (Bourgin et al., 2019; Erev et al., 2017; Plonsky et al., 2019).

2.2.1 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). This model 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).

2.3 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. All participants were nominally paid a base participation fee of INR 100. However, a random game's outcome was selected at the end of three sessions which was scaled such that the final payoff was bracketed between INR 0 to 200 for each participant.

2.3.1 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). However, since the winning model was not significantly better than the baseline model in decisions from experience in the Technion tournament (Erev et al., 2010), we used only the baseline model in our evaluations.

2.4 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 paradigms.

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 (SD) greater than 2.

2.5 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 Pearson correlation coefficient 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).

However, using any one measure of consistency is likely inadequate, since such measures make hidden assumptions about the distributions of the variables being compared, e.g. that they are approximately normally distributed. Therefore, we report two additional measures.

We report Mean squared distance (MSE) to gauge the drift between the observed and predicted R-rates across problems. 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).

It is quite possible that more robust measurements of test-retest consistency may be possible using a more sophisticated generative model of the task (Wall et al., 2021). However, the robustness of such measures depend on the soundness of the generative model to the task, which is itself unknown. In conjunction with the fact that the choice modelling literature has historically focused on reporting the three statistical measures mentioned above Erev et al. (2017, 2010); He et al. (2021), this paper is better served by reporting choice consistency using the same measures.

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.

3 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). 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 as EEP.

Table 1 DFD : retest reliability. Retest reliability (data-data) agreement statistics presented alongside model-data statistics reported from different data sources. Scatterplot between two elicitations of R-rate corresponding to first row in Figure B2

Observed	Predictor	Correlation	MSE	κ	P_{Agree}
RR1	RR2	0.85	0.0116	0.32	82%
TE	BM	0.85[1]	0.09[1]	0.22	95%[1]
	WM	0.92[1]	0.0099[1]	0.23	88%[1]
TC	BM	0.86[1]	0.08[1]		93%[1]
	WM	0.94[1]	0.012[1]		90%[1]
EEP	BEAST	0.95[2]	0.0098[1]	-	

Note: Data sources and models are specific to DFD. Label notations in text.

[1]claimed in Erev et al. (2010).

[2]claimed in Erev et al. (2017).

3.1 Predicting choice proportions

3.1.1 Decisions from description

Table 1 summarizes consistency metrics calculated for our DFD experiment's data, presented alongside human-model consistency metrics reported previously on existing datasets.

Table 2 DFD : generalizability analysis Choice agreement indicators with best fit predictions from baseline (BM = CPT) and winning (WM = Logistic Regression) choice models (TCPT2010) using training dataset predicting on test dataset. Observed versus predicted R-rate for each set of observations is presented in Appendix B1

Train	Test	Model	Correlation	MSE	κ	P_{Agree}
RR-F	RR-F	BM	0.82	0.01	0.09	84%
		WM	0.93	0.004	0.12	89%
TE	RR-F	BM	0.74	0.025	0.12	84%
		WM	0.83	0.023	0.14	81%
RR1	RR2	BM	0.71	0.018	0.13	82%
		WM	0.84	0.011	0.10	88%
RR2	RR1	BM	0.78	0.017	0.08	80%
		WM	0.87	0.011	0.14	80%

Note: Data sources and models are specific to DFD. Label notations in text.

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 a correlation of about 0.85 which is consistent with the baseline model predictions on both tested and trained datasets. That is, the baseline model predicts peoples' choices about as well as their own responses to the same problems during retesting predict their responses in the first session. However, this correspondence is not seen in individual level statistics, with the empirical test-retest κ being considerably larger than the baseline model's κ.

Thus, we see that peoples' own choices for the same choice problems vary considerably across sessions. If people respond differently to the same problem when they see it again, with all other factors remaining constant, it is impossible for a model that uses only individual problem characteristics to predict both responses correctly (Fudenberg et al., 2019). The test-retest statistics we report in Table 1 imply limits on the predictability of risky choice behavior, in a manner that we make more precise below in Section 3.2.

Table 2 summarizes results from a model generalizability analysis we used to assess test-retest consistency mediated by parametric model fits [1] [2] [3]. Through this analysis, we seek to identify how well models fit to responses from participants fit responses to the same problems by the same participants on a different day.

We note that the models, when fit to the entirety of the RR dataset, evince goodness-of-fit statistics nearly identical to the ones reported in the Technion

[1] Please note that when we implemented the models from (Erev et al., 2010), the parameter values given in the original paper could not reproduce the results for TE and TC. But when we fitted our own parameter values for TE dataset (which are remarkably different), we could reproduce the results for the Winning model but not for the Baseline model. Additionally, the κ statistic for the TE dataset mentioned in the table is estimated using our model implementation and fits found for TE dataset.

[2] {Reported, Estimated} fits for Baseline model on TE dataset : α = {0.7,0.846}, β = {1, 0.995}, λ = {1, 1.11}, γ = {0.65, 0.68}, δ = {0.65, 0.38}. MSE = {0.09, 0.0197}

[3] {Reported, Estimated} fits for Winning model on TE dataset : β_0 = {1.004,0.415}, β_1 = {0.012, 0.0097}, β_2 = {0.066, 0.058}, β_3 = {0.410, -0.4288}, γ_1 = {1.417, -1.508}, γ_2 = {0.317, 0.349}, γ_3 = {0.621, 0.596}. MSE = {0.0099, 0.0099}

Tournament for the TE dataset Erev et al. (2010). However, these goodness-of-fit statistics deteriorate considerably when we measure them on one session's data after having fit the models on the other session's data. For instance, the winning model's data-model correlation drops by about 10% and MSE rises by more than 100%. Similar deterioration in fit is seen when we try to transfer model fits from TE to the whole RR dataset.

Most interestingly, we see that baseline as well as winning models from the Technion 2010 tournament regress to goodness-of-fit values not exceeding the test-retest benchmark value when we train them on one session's data and try to predict the other session's choices. Thus, at least empirically, there appears to be a limit to how well data fit to one session predicts the other session's behavior for the same set of participants, one that closely matches the test-retest correlation between the two sessions' data.

This observation illuminates the central paradox of modelling risky choices we highlight in this paper: they are highly variable even within individuals, but individual choice elicitations cannot represent this variability. Thus, models that seek to optimize fits to individual choice elicitations can end up being more precise than the very phenomenon they are seeking to explain, and thus generalize poorly to other elicitations of the same phenomenon.

3.1.2 Decisions from experience

Table 3 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 estimation set in Erev et al. (2010). 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 3, we also 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 $\kappa \in \{0.33, 0.33\}$ 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.

Unlike the case for decisions-by-description, we find that models fit to one session's data are able to generalize very well to the other session's data,

Table 3 DFE : retest reliability Choice agreement indicators to estimate test-retest reliability between repeated problems in the first row presented alongside model-data statistics reported from different data sources. For each indicator, we also separately report values for observations with sampling duration greater than 2. The last two columns present the retest reliability of sampling duration.

		Correlations		MSE		κ		P_{Agree}		Correlation SD	
Observed	Predictor	All	SD >2	All	SD >2	All	SD >2	All	SD >2	All	SD >2
RR1	RR2	0.89	0.88	.0137	0.019	0.33	0.39	93%	94%	0.61	0.56
TE	BM	0.88[1]	-	0.017[1]	-	0.25	0.25	95%[1]	-	-	-
	WM	0.92[1]	-	0.0099[1]	-	-	-	95%[1]	-	-	-
TC	BM	0.80[1]	-	0.0244[1]	-	0.19	0.2	82%[1]	-	-	-
	WM	0.80[1]	-	0.0187[1]	-	-	-	83%[1]	-	-	-

Note: Data sources and models are specific to DFE. Label notations in text. WM = Winning Model (ensemble), BM = Baseline Model (primed sampler) in Erev et al. (2010).

[1]claimed in Erev et al. (2010).

Table 4 DFE : generalizability analysis Choice agreement indicators with best fit predictions from baseline model (TCPT2010) on RR subsets presented alongside similar indicators where the best fit parameters from training dataset are used to predict on test dataset.

Train	Test	Model	Best fit parameter	Correlations	MSE	κ	P_{Agree}
RR-F	RR-F	BM	3	0.89	0.0234	0.284	88%
TE	RR-F	BM	4	0.82	0.0266	0.253	89%
RR1	RR2	BM	3	0.86	0.0254	0.272	92%
RR2	RR1	BM	3	0.84	0.0297	0.237	87%

Note : Data sources and models are specific to DFE. Label notations in text. BM = Baseline Model (simple primed sampler) in Erev et al. (2010).

showing correlation drops of about 5% and MSE increases of about 2% from values seen when we fit the model to both sessions' data. This large difference in generalization ability across the two paradigms is particularly noteworthy. Whereas decisions by description have been used prolifically by experimentalists for their simplicity, psychologists have consistently championed the use of decisions-by-experience as being more consistent with humans' natural information ecosystem (Hertwig & Erev, 2009), and thus more likely to yield valid psychological explanations for choice behavior.

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 last column of Table 3, human-human correlations for sampling duration in repeated problems for all observations is 0.61, dropping to 0.56 when only observations with sampling duration greater than two are considered. These values indicate reasonable benchmarks for the predictability of sampling duration in decisions from experience.

Interestingly, this value 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).

3.2 Interpretation of Results

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), or estimated indirectly as in Fudenberg et al. (2019). Two, we identify an interesting benchmark for how precisely cognitively realistic models of humans' risky choices could be expected to match empirical choice data.

For decisions from description, we found that participants' own future choices predicted at most about 70% of the variability in their previous choices on the same choice problems, and suggest this as a reasonable benchmark for prediction performance for realistic choice models of one-shot decisions from description. While model-data correlations greater than this value may be technically possible, as is evident from the results of multiple prediction tournaments (Erev et al., 2017, 2010) and other recent studies He et al. (2021); Peterson et al. (2021), our results show that such out-performance cannot be treated as evidence of model superiority, since the model is likely to show worse fits if tested on data from the same participants collected on a different occasion, assuming the participants were just as inconsistent in their choices across sessions as participants in our experiment.

We presented empirical evidence for such worse fits across sessions in our results above. It is also possible to present a theoretical argument that such drops in correlations must necessarily occur for sufficiently high values of model-data correlations seen in any one session. In particular, assuming X is response data obtained during the first session of our experiment, Y is response data obtained during the second session of our experiment, and Z are a model's predictions for the second session's data, a well-known statistical identity shows that,

$$\rho_{XY}^2 + \rho_{YZ}^2 + \rho_{ZX}^2 \leq 2\rho_{XY}\rho_{YZ}\rho_{ZX} + 1,$$

which, when viewed as a quadratic inequality in ρ_{ZX}, implies an upper bound on ρ_{ZX} as

$$\rho_{ZX} \leq \rho_{XY}\rho_{YZ} + \sqrt{(1-\rho_{XY}^2)(1-\rho_{YZ}^2)}.$$

For example, as illustrated in Figure 2, the correlation between the model's predictions and the first session's data cannot be greater than this upper bound. For $\rho_{XY} = 0.85$ (as measured in our experiment) and an over-precise $\rho_{YZ} = 0.98$, the greatest possible $\rho_{ZX} = 0.94$. In general, as we can see from Figure 2, given the estimate of ρ_{XY} obtained in our experiment, model-data correlations above 0.95 become unreliable as goodness-of-fit indicators, as seeing such a large correlation for one session's data renders it statistically impossible to see equally high correlations for a different session's data. Thus, model-data correlations larger than this value cannot be reliably used for model selection. Note that this theoretical limit would apply irrespective of normality or i.i.d sampling assumptions on the random variables X, Y, Z.

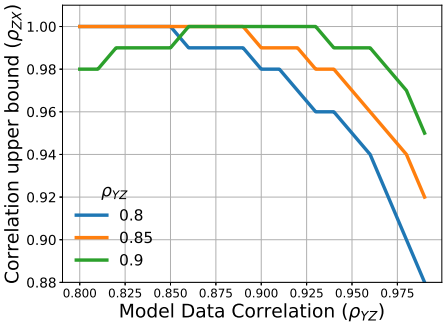

Fig. 2 Upper bound on other session's correlation with model. Given less than perfect correlation between responses measured across two different sessions, this plot shows the theoretical upper limit of model-data correlation possible for the second session's data (y-axis) for a model that shows different levels of correlation with one session's responses (x-axis). Different line plots show how this relationship changes as a function of the data-data correlation itself.

Another perspective from which we can view this limit is to appreciate that a good model should yield good MSE estimates across multiple elicitations of the R-rate across all problems. If the R-rate for the same problem is 0.2 when I ask participants about it on Monday, and 0.8 if I ask them about it on Friday, then the best model would predict an R-rate of 0.5, thereby incurring a minimum *irreducible* MSE of 0.09 for that problem in both datasets.

For average R-rate across data elicitations to differ by an MSE of 0.012 as in our dataset, under normality assumptions on the nature of the shift, the average absolute shift size per problem would be approximately 0.11. An omniscient predictive model would perfectly identify the direction of this shift per problem, and minimize the cumulative MSE across the shifts by placing its prediction halfway between them (mean predictive shift = 0.055). If this predictive model had an MSE of approximately 0.02 originally, as is the case for state-of-the-art theory-guided neural models (Peterson et al., 2021), and hence an average absolute deviation of about 0.14 from the real value, adding equi-probable shifts of size 0.055 in either direction will increase the MSE to approximately 0.023. Thus, our MSE estimate of test-retest variability suggests that for models in the 0.02 MSE range, MSE differences of up to 15% (and perhaps even larger) are unreliable differentiators of predictive ability.

From either perspective, we note that our results place constraints on the ability to interpret goodness-of-fit statistics, whether mediated by considerations of parameter complexity or not, as evidence for model superiority. At sufficiently high levels of correspondence between choice models and choice data, model predictions become too precise to be trustworthy as indicators of a model capturing something specific about the phenomenon being modeled rather than the dataset being fitted. Thus, extremely high goodness-of-fit values in decisions-by-description may be fundamentally illusory - they are derived correctly in a purely technical sense, but better reflect a model's ability to fit choice data than it's ability to actually predict choices.

Our results for decisions-from-experience stand out in clear contrast to the paradoxical results seen in our DFD analysis. Not only do the best models for DFE not report goodness-of-fit statistics significantly better than our test-retest benchmarks, they also generalize well from one session to another. The superior robustness of DFE models in our analysis is doubly reassuring. On the one hand, it supports the case that these models are indeed explaining real aspects of the choice process in DFE paradigms. On the other, it also suggests that our experiment design and analytic tools are not fundamentally biased against the possibility of finding consistency in choice models.

4 Historical model performance analysis

Our test-retest analysis revealed a close correspondence between the data-data correlation seen in our experiments and model-data correlations shown by fairly simple models of risky choice for both decisions from description and experience. In conjunction with recent indications that CPT is close to optimal in explaining certainty equivalence judgments (Fudenberg et al., 2019), these observations suggest that these test-retest consistency measures are reasonable benchmark values for choice models to target.

On this account, simple baseline models in prediction tournaments are already close to optimal in explaining choice in experimental risky choice paradigms; more complex models possibly improve empirically upon these baselines by having been over-fit to specific datasets. We decided to explore this possibility by means of a historical model comparison. We followed the methodology presented recently in He et al. (2021), selecting 16 models of risky choice in chronological order from the literature for their prominence in the field and seeing how well they predict our data.

4.1 Methods

We implemented a total of 16 prominent risky models of choice for decisions from description paradigm. The names and reference paper for each model is given in Table A1. We followed the model implementation provided by He et al. (2021) and performed model optimisation to recover the parameter fits.

Deterministic predictions from models were passed through a stochastic logistic function to emit the choice probability per problem. We define choice probability as the probability that second option of the two given is chosen. For all the results reported below, we trained the model parameters on the Technion Prediction Tournament Estimation (TE) dataset, and calculated performance metrics using our dataset RR-F as a held-out test set.

We used max likelihood estimation to fit all parameters. For each model optimization, the optimisation routine was run 10 times using different starting point parameters and we picked the one with the minimum negative log likelihood function value, where log likelihood is defined by $[\sum_i^N \sum^M y_i \log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)]$, where y_i is the observed choice, $\hat{y}_i$ is the predicted choice

(0 if first option was chosen, 1 otherwise) for i_{th} problem of N problems and M is the number of subjects.

Finally, we used the predictions for each model to calculate the four response variables described in the first section, three cohort level statistics namely MSE, correlation and P_{Agree}; and one individual binary choice statistic κ. If baseline models are already near-optimal in predicting risky choice behavior, we expect to see model performance asymptote close to the empirical test-retest benchmark value fairly early in the historical sequence.

4.2 Results

Each plot in Figure 3 shows the historical trend of the four response variables. For every response variable, the test-retest benchmark is calculated using our dataset RR where the observed choice comes from the first elicitation of the problem (RR1) while the predicted choice comes from the second elicitation of the same problem (RR2). The models are laid out in chronological order from left to right in Figure 3, starting with a random heuristic corresponding to a flip of a coin, representing the worst possible performance.

Broadly speaking, Figure 3 bears out our expectation: we see performance closely approaching the test-retest benchmark (TRB) levels on all three aggregate measures for even the simplest baseline model of risky choice - expected value maximization, with subsequent models surprisingly unable to improve on this performance level. Thus, our historical analysis supports the view that simple models of risky choice already explain nearly all available interesting aggregate-level trends in participant behavior in decisions from description, and further supports the case for test-retest benchmark values across these performance measures as interesting targets for model-data correlation.

We note, however, that there appears to be considerable scope for improvement in predicting individual-level performance in risky choice tasks. As is evident from Figure 3, κ is considerably below the test-retest benchmark value for nearly all models, with more recent models showing better agreement with human judgments than EV and EU models, which are close to random in predicting individual-level behavior. Thus, while classic expected utility models are adequate to explain aggregate economic behavior in risky choice, more complex choice models like prospect theory are considerably better in predicting individual choices (Glöckner & Pachur, 2012).

4.3 Discussion

We see that historically ancient EV and EU models of risky choice are able to make large improvements in all consistency measures on our dataset over a random predictor. More interestingly, this excellent performance for these models leaves minimal room for genuine improvement in predictability, such that all models' predictions show approximately the same fit with our data as these classic models. We also note with interest that, in our evaluation,

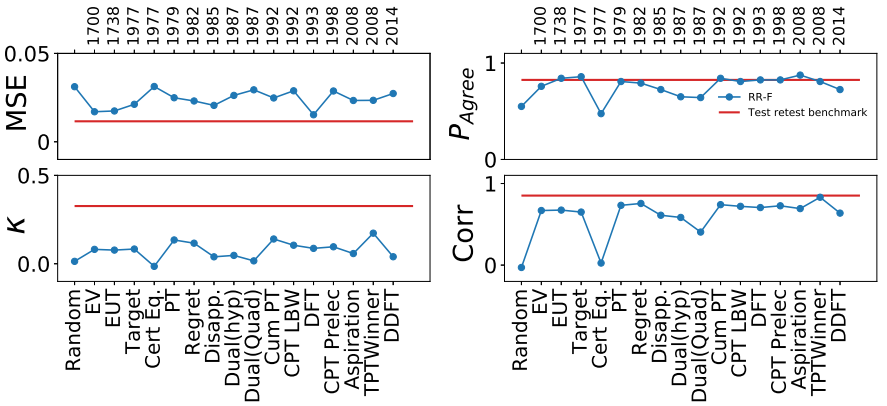

Fig. 3 Historical Analysis for DFD-RR dataset. In a given plot, every dot corresponds to the respective evaluation statistic [MSE, P_{Agree}, Kappa and Correlation] per model arranged in chronological order. Parameters corresponds to the best fit found for TE dataset. Horizontal red line is the test-retest benchmark value for each statistic.

the best fit models closely approached the test-retest benchmark values for all aggregate response variables, without (significantly) exceeding them.

These observations lead us to speculate that test-retest consistency measures may offer an interesting threshold for model-data correspondence, such that goodness-of-fit values up to this threshold may reliably indicate superior ability to explain behavior, but values beyond it may simply be artifacts of model complexity leading to over-fitting to dataset-specific properties.

As we document in Figure 4, MSE values lower than our benchmark value have been reported recently in the literature (He et al., 2021; Peterson et al., 2021; Plonsky et al., 2019). There are two possible explanations for such results, not necessarily mutually exclusive. One, the value we have calculated is a point estimate that does not express any information about its variability across datasets. It is possible that it may vary slightly across datasets, depending on the distribution of R-rates across problems within them, which in turn may depend on problem selection, population characteristics, and experiment protocol variations.

Two, it is also possible that complex models are able to *represent* subtle variations within specific datasets they are trained on, and even empirically generalize in a limited sense to other data points in the same dataset across cross-validation splits due to test set reuse, but are unable to *generalize* to truly unseen datasets, a phenomenon called adaptive overfitting (Roelofs et al., 2019). In fact, as recent work in statistical learning demonstrates, it is possible for sufficiently complex models to empirically generalize even systematic functional mappings from noise to random labels (Zhang, Bengio, Hardt, Recht, & Vinyals, 2021). Thus, it is becoming evident in the statistical learning literature that sufficiently complex models can fit labels to datasets with very high levels of precision, but are unable to generalize to repeat elicitations of the same data itself Recht, Roelofs, Schmidt, and Shankar (2019).

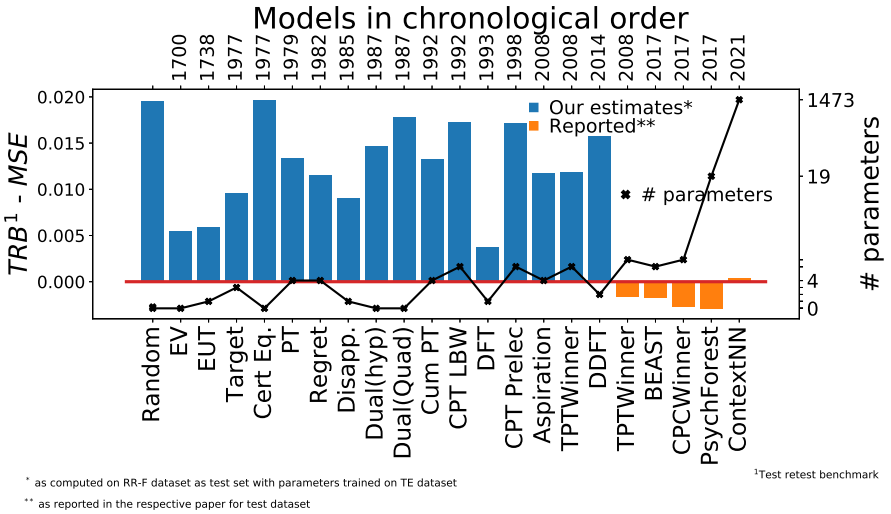

Fig. 4 Reported prediction trend in recent complex models of risky choice on different datasets compared with prediction estimates on DFD-RR (Fresh) in the backdrop of number of parameters required for fitting. Left ordinate = test retest benchmark - MSE, right ordinate = number of parameters corresponding to each model.

In this regard, it is interesting to note that whereas complex parametric models, extremely liable to overfitting to the test set under leaderboard conditions (Dwork et al., 2015), report MSE values much lower than the test-retest MSE benchmark value (He et al., 2021; Plonsky et al., 2019), a recent neural network model trained on a very large dataset, and potentially protected from overfitting by the use of dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), reports MSE values almost exactly coinciding with our benchmark value (Peterson et al., 2021).

In summary, the results of our historical analysis suggest that progress in modeling risky choices from description at an aggregate level has been somewhat illusory for the past several decades, that some improvements in capturing individual level variations in choice behavior have occurred, and that there still remains considerable scope for improvement in predicting individual level risky choices. These conclusions are broadly in concordance with earlier studies that have sought to measure the reliability of choice models (Glöckner & Pachur, 2012).

5 General Discussion

Given that human choices are irreducibly noisy, how well could we reasonably expect choice models to track human choices? In this paper, we tried to answer this question by measuring test-retest consistency in responses to the same risky choice problems asked one week apart to participants. If humans, looking at the same problem again, are unable to select the same response as they

selected previously, how could we expect models looking at the same problem to do so?

We found that, whereas models can be fitted to yield very high goodness-of-fit statistics, testing them on data collected from the same participants and for the same problems on a different day yields goodness-of-fit values that closely match test-retest consistency values. Thus, test-retest consistency measures appear to offer reasonable performance benchmarks for the predictability of risky choice behavior. Our results suggest that goodness-of-fit statistics significantly higher than these benchmarks have a high probability of regressing towards them if the same experiment is replicated on a different day. In fact, we demonstrate with a theoretical argument that, for sufficiently high model-data correlations, given our test-retest consistency values, it is guaranteed that a second session's data will show lower correlations with the model.

As in the analysis used by Fudenberg et al. (2019), it is important to note that these benchmarks apply only to models that use only problem-specific information to emit choices. In principle, both participants and models may use information from alternative sources, e.g. the order in which problems are seen, extremely high or low payoffs incurred during one session etc. that may not be replicated in retest sessions. Thus, it is possible for choice models that take such influences into account to fit data from both sessions better by taking these additional sources of information into account. Such models may be differentiated using goodness-of-fit statistics beyond the test-retest benchmarks we have identified in this paper. Separate experiments fixing problem order etc. are needed to establish similar benchmarks for such models.

Nonetheless, our benchmarks still apply to a large number of choice models, including all the ones covered in our historical analysis, choice tournament winners (Erev et al., 2017, 2010; Plonsky et al., 2019), and several state-of-the-art proposals (He et al., 2021; Peterson et al., 2021). Since problem descriptions remain unchanged across elicitations in our design, and all these models use only problem descriptions to produce choice predictions, the test-retest benchmarks identified in our experiments clearly apply for them.

The unexpectedly high precision seen in several recent models of decisions from description Erev et al. (2017, 2010); He et al. (2021); Peterson et al. (2021) appears to be most consistent with the possibility that they have been subtly over-fit to the test set, a problem endemic to prediction tournaments with leaderboards (Dwork et al., 2015). Such a conclusion would also resolve the mystery of over-performance beyond statistical expectation for risky decisions from description calculated in Fudenberg et al. (2019). Put together with the asymptotic trends seen in our historical model comparison, these results indicate that generalized expected utility maximization is a near-optimal model of decisions from description in certainty equivalence settings, and alternative models have very little scope for improvement in predicting aggregate choices before they begin to over-fit to training data.

Here, we don't mean over-fitting in the narrow technical sense of being able to generalize across train-test set splits from the same dataset. Rather,

we reference it in the more general sense of complex models being able to show good performance only in highly specific settings (Yarkoni, 2022). In contrast, simple expected value models generally yield reasonable correspondence with data across a wide variety of elicitation conditions and datasets in risky choice paradigms. If complex models cannot yield competitive performance on unseen datasets without having to first train extensively on them, we show that it becomes difficult to trust any theoretical conclusions drawn from the model having been fit very well to one data set in particular for modelling risky choices, because the dataset contains a single sample of an inherently stochastic process.

For risky 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 amount of variance in responses captured by a primed sampler model presented in Erev et al. (2010) with our benchmark values, suggesting that the primed sampler model is already close to an optimal predictor for this task, as was also discovered empirically in the tournaments with winning models unable to improve significantly upon the baseline (Erev et al., 2010). 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.

The disjoint pattern of results seen across both decisions from description and from experience in our experiments is consistent with a fairly commonsensical explanation. We saw that choices across sessions were maximally consistent within individuals ($\kappa > 0.7$) when the same person saw a sequence of samples that yield approximately the same expected value difference between lotteries on both occasions. When sequences with different expected value differences are included, the choice-level consistency for decisions from experience reduces considerably ($\kappa = 0.33$), becoming nearly equivalent to that seen in decisions from description ($\kappa = 0.32$). Thus, the sampling history for a DFE trial contains much information about the individual's upcoming choice.

Therefore, a crucial difference between DFE and DFD modelling is that DFE models have access to the sequence of samples drawn before each choice is made, data that are clearly relevant to an observer's choice. The presence of such rich data, prior to each binding choice being made, permits DFE models to be simpler and more mechanistic in design, since much of the explanation of the observer's choice lies within the data itself. On the other hand, DFD models have basically no more information than what generalized expected utility models can use, and therefore it becomes unrealistic to expect better fits to data without complexifying models. Thus, our conclusion that expected utility models are already explaining most of the explainable variance in decisions from description should not be unexpected. If the only prediction target for

data from such experiments remains aggregate choice proportions, overly complex models cannot genuinely improve upon simple models that are already using prospect information appropriately.

For modelling studies focusing exclusively on choice proportions in the decisions from description paradigm, our results suggest a shift towards evaluating choice models based on their ability to predict individual choices, or identify individual-level differences, as attempted in, e.g. Glöckner and Pachur (2012) (see also Chuang and Schechter (2015) for a broader review of such efforts). More generally, our results favor modelling (Gonzalez & Dutt, 2011) and theoretical (Stewart, Chater, & Brown, 2006) proposals that seek to bridge the gap between decisions from experience and decisions from description instead of trying to optimize for accuracy in any one paradigm. Such proposals commonly impute some degree of sampling of simulated lottery outcomes as the mechanistic link between the two types of decisions. Thus, response variables sensitive to such mechanistic claims should be treated as prediction targets in conjunction with raw choice data.

Recent studies measuring gaze data in conjunction with choices in decisions from description have uncovered interesting inconsistencies between classic models' predictions and the actual link between gaze and choice in decisions from description, showing for instance that gaze duration predicts choices, but valuations don't predict gaze durations (Glöckner, Fiedler, Hochman, Ayal, & Hilbig, 2012; Stewart, Hermens, & Matthews, 2016), stimulating several new theoretical proposals (Gluth, Kern, Kortmann, & Vitali, 2020; Sepulveda et al., 2020; Smith & Krajbich, 2019). In line with these exciting developments, by showing that goodness-of-fit to choice proportions can no longer differentiate models beyond a certain threshold, this paper highlights the necessity of shifting towards studying decisions from description using a wider repertoire of response variables. Paralleling the imperative to unite explanations from both brain and behavior in the context of decision-making, our results suggest that good models of choice behavior must be differentiated based on consilience with mechanistically meaningful and observable decision correlates (Glimcher & Rustichini, 2004).

More broadly, increases in computing power and data collection capabilities are placing larger datasets and more complex models within the reach of behavior scientists (Peterson et al., 2021). It has even been suggested that, since psychological constructs are frequently difficult to test in designs that control for all possible confounds, the field may benefit from ignoring explanation in favor of seeking to maximize prediction using large amounts of data to implicitly control for the large number of confounds that bedevil small sample studies (Yarkoni & Westfall, 2017). While this is an enticing prospect, for good science to happen prediction and explanation must advance in tandem (Kuhn, 1970).

This article presents a case example emphasizing the need for caution before permitting prediction to overtake explanation in psychological science,

for instance, by treating deep neural networks as scientific models of behavior (Cichy & Kaiser, 2019). We show that predictions generated by complex models trained on large datasets may end up being too precise to be true if the stochasticity underlying prediction targets is not properly recognized. The classic practice of obtaining test-retest consistency measures for prediction targets, as demonstrated in this article, can help prevent such mishaps across a variety of behavioral tasks, reining in models from making over-precise predictions based on single samples of highly stochastic events.

Finally, we note that an appreciation of fundamental limits on the predictability of many forms of behavior is a necessary, but not sufficient condition, for engaging with the generalizability crisis confronting behavioral research, wherein researchers use quantitative measurements that are weakly related to the underlying qualitative phenomenon they are actually interested in understanding (Yarkoni, 2022). As part of the commodification of science, researchers are frequently tempted to substitute precision of prediction or measurement in place of precision of understanding as markers of scholarly achievement (Stark & Saltelli, 2018). Since behavioral constructs are intrinsically noisy, behavioral disciplines have had to bear the brunt of this generalizability crisis disguised as a replication crisis (Loken & Gelman, 2017; Maxwell, Lau, & Howard, 2015). Constructs that vary within individuals across multiple elicitations, therefore, must be subjected to special cautions in both measurement and interpretation. As we show in this paper, risky choice proportion in certainty equivalence experiments is one such construct, and therefore, should be treated accordingly.

Declarations

- Funding, DST India DST/CSRI/2017/334
- Conflicts of interest: none
- Consent to publish: Not applicable
- Ethics approval: all data collection approved by IITK's IRB
- Consent to participate: informed consent was recorded as a form item online at the time of participation in our experiment.
- All code and data are available at https://osf.io/zjd38/

References

Bell, D.E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*(5), 961–981.

Bhatia, S. (2014). Sequential sampling and paradoxes of risky choice. *Psychonomic Bulletin & Review*, *21*(5), 1095–1111.

Bhatia, S., & Loomes, G. (2017). Noisy preferences in risky choice: A cautionary note. *Psychological Review*, *124*(5), 678-687.

Bishop, C.M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Bourgin, D.D., Peterson, J.C., Reichman, D., Russell, S.J., Griffiths, T.L. (2019). Cognitive model priors for predicting human decisions. *International Conference on Machine Learning* (pp. 5133–5141).

Brandstätter, E., Gigerenzer, G., Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*(2), 409-432.

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

Busemeyer, J.R., & Townsend, J.T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432-459.

Chuang, Y., & Schechter, L. (2015). Stability of experimental and survey measures of risk, time, and social preferences: A review and some new results. *Journal of Development Economics*, *117*, 151–170.

Cichy, R.M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317.

Diecidue, E., & Van De Ven, J. (2008). Aspiration level, probability of success and failure, and expected utility. *International Economic Review*, *49*(2), 683–700.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A. (2015). Generalization in adaptive data analysis and holdout reuse. *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), 369-409.

Erev, I., Ert, E., Roth, A.E., Haruvy, E., Herzog, S.M., Hau, R., ... Lebiere, C. (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*, *30*(11), 1283–1300.

Fishburn, P.C. (1977). Mean-risk analysis with risk associated with below-target returns. *The American Economic Review*, *67*(2), 116–126.

Fudenberg, D., Kleinberg, J., Liang, A., Mullainathan, S. (2019). Measuring the completeness of theories. *arXiv preprint arXiv:1910.07022*.

Glimcher, P.W., & Rustichini, A. (2004). Neuroeconomics: the consilience of brain and decision. *Science*, *306*(5695), 447–452.

Glöckner, A., Fiedler, S., Hochman, G., Ayal, S., Hilbig, B. (2012). Processing differences between descriptions and experience: A comparative analysis using eye-tracking and physiological measures. *Frontiers in Psychology*, *3*, 173.

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*(1), 21–32.

Gluth, S., Kern, N., Kortmann, M., Vitali, C.L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, *4*(6), 634–645.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological Review*, *118*(4), 523-551.

Handa, J. (1977). Risk, probabilities, and a new theory of cardinal utility. *Journal of Political Economy*, *85*(1), 97–122.

He, L., Analytis, P.P., Bhatia, S. (2021). The wisdom of model crowds. *Management Science*, *68*(5), 3635-3659.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523.

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.

Jia, J., Dyer, J.S., Butler, J.C. (2001). Generalized disappointment models. *Journal of Risk and Uncertainty*, *22*(1), 59–78.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kuhn, T.S. (1970). *The structure of scientific revolutions* (Vol. 111). Chicago University of Chicago Press.

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

Lattimore, P.K., Baker, J.R., Witte, A.D. (1992). The influence of probability on risky choice: A parametric examination. *Journal of Economic Behavior & Organization*, *17*(3), 377–400.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585.

Luce, R.D., & Suppes, P. (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. *37th Annual Meeting of the Cognitive Science Society* (pp. 1512–1517).

Maxwell, S.E., Lau, M.Y., Howard, G.S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487-498.

Peterson, J.C., Bourgin, D.D., Agrawal, M., Reichman, D., Griffiths, T.L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D.D., Peterson, J.C., ... Erev, I. (2019). Predicting human decisions with behavioral theories and machine learning. *ArXiv*, *abs/1904.06866*.

Plonsky, O., Erev, I., Hazan, T., Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. *AAAI conference on Artificial Intelligence.*

Prelec, D. (1998). The probability weighting function. *Econometrica*, 497–527.

Recht, B., Roelofs, R., Schmidt, L., Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? *International conference on machine learning* (pp. 5389–5400).

Roelofs, R., Fridovich-Keil, S., Miller, J., Shankar, V., Hardt, M., Recht, B., Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 9179–9189).

Sepulveda, P., Usher, M., Davies, N., Benson, A.A., Ortoleva, P., De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *Elife*, *9*, e60705.

Smith, S.M., & Krajbich, I. (2019). Gaze amplifies value in decision making. *Psychological Science*, *30*(1), 116–128.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

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

Stark, P.B., & Saltelli, A. (2018). Cargo-cult statistics and scientific crisis. *Significance*, *15*(4), 40–43.

Stewart, N., Chater, N., Brown, G.D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Stewart, N., Hermens, F., Matthews, W.J. (2016). Eye movements in risky choice. *Journal of Behavioral Decision Making*, *29*(2-3), 116–136.

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

Wall, L., Gunawan, D., Brown, S.D., Tran, M.-N., Kohn, R., Hawkins, G.E. (2021). Identifying relationships between cognitive processes across tasks, contexts, and time. *Behavior Research Methods*, *53*(1), 78–95.

Yaari, M.E. (1987). The dual theory of choice under risk. *Econometrica*, *55*(1), 95–115.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115.

Appendix A List of evaluated models

Table A1 Models for historical analysis. Prominent risky choice model labels as used in the paper in chronological order, and the reference paper for implementation details.

Evaluation statistics estimated	
Model Label	Reference Paper
EV	1700
EUT	1738
Target	Fishburn (1977)
Cert Eq.	Handa (1977)
PT	Kahneman and Tversky (1979)
Regret	Bell (1982)
Disapp.	Jia, Dyer, and Butler (2001)
Dual(hyp)	Yaari (1987)
Dual(Quad)	Yaari (1987)
Cum PT	Tversky and Kahneman (1992)
CPT LBW	Lattimore, Baker, and Witte (1992)
DFT	Busemeyer and Townsend (1993)
CPT Prelec	Prelec (1998)
Aspiration	Diecidue and Van De Ven (2008)
TPTWinner	Erev et al. (2010)
DDFT	Bhatia (2014)
Evaluation statistics reported	
BEAST	Erev et al. (2017)
CPCWinner	Erev et al. (2017)
PsychForest	Plonsky, Erev, Hazan, and Tennenholtz (2017)
ContextNN	Peterson et al. (2021)

Appendix B Data: Problems and Responses

Here, we report the actual problems used in our experiment for completeness, and show scatter plots corresponding to the correlation values reported in the main text to verify that no statistical anomalies confound our interpretation of the correlation numbers.

Table B2 DFD : Observed response variables The 60 problems from Estimation set in Technion Choice Prediction tournament in column 2-5. Observed risky choice rate for four subsets of our dataset (RR) alongside the observed R-rate for TE dataset in the last column. Data subsets labels specified in text.

	Risky Prospect			Safe	Observed Risky choice rate				
Problems	High	Low	P(High)	Medium	RR1	RR2	RR	RR-F	TE
1	-0.30	-2.10	0.96	-0.30	0.19	0.19	0.21	0.22	0.20
2	-0.90	-4.20	0.95	-1.00	0.20	0.17	0.22	0.24	0.20
3	-6.30	-15.20	0.30	-12.20	0.59	0.59	0.58	0.57	0.60
4	-10.00	-29.20	0.20	-25.60	0.83	0.63	0.72	0.76	0.85
5	-1.70	-3.90	0.90	-1.90	0.53	0.40	0.47	0.50	0.30
6	-6.30	-15.70	0.99	-6.40	0.32	0.35	0.40	0.43	0.35
7	-5.60	-20.20	0.70	-11.70	0.76	0.80	0.77	0.76	0.50
8	-0.70	-6.50	0.10	-6.00	0.81	0.70	0.72	0.72	0.75
9	-5.70	-16.30	0.95	-6.10	0.51	0.35	0.42	0.47	0.30
10	-1.50	-6.40	0.92	-1.80	0.29	0.36	0.38	0.40	0.15
11	-1.20	-12.30	0.02	-12.10	0.81	0.75	0.79	0.81	0.90
12	-5.40	-16.80	0.94	-6.40	0.56	0.59	0.59	0.59	0.10
13	-2.00	-10.40	0.05	-9.40	0.43	0.51	0.48	0.47	0.50
14	-8.80	-19.50	0.60	-15.50	0.72	0.79	0.75	0.72	0.70
15	-8.90	-26.30	0.08	-25.40	0.76	0.62	0.71	0.76	0.60
16	-7.10	-19.60	0.07	-18.70	0.74	0.59	0.62	0.64	0.55
17	-9.70	-24.70	0.10	-23.80	0.77	0.67	0.69	0.71	0.90
18	-4.00	-9.30	0.20	-8.10	0.64	0.64	0.58	0.55	0.65
19	-6.50	-17.50	0.90	-8.40	0.67	0.50	0.68	0.74	0.55
20	-4.30	-16.10	0.60	-4.50	0.10	0.24	0.16	0.12	0.05
21	2.00	-5.70	0.10	-4.60	0.40	0.44	0.41	0.40	0.65
22	9.60	-6.40	0.91	8.70	0.27	0.14	0.26	0.31	0.05
23	7.30	-3.60	0.80	5.60	0.44	0.53	0.44	0.40	0.15
24	9.20	-9.50	0.05	-7.50	0.47	0.33	0.39	0.41	0.50
25	7.40	-6.60	0.02	-6.40	0.74	0.74	0.72	0.71	0.90
26	6.40	-5.30	0.05	-4.90	0.85	0.64	0.67	0.69	0.65
27	1.60	-8.30	0.93	1.20	0.19	0.29	0.28	0.28	0.15
28	5.90	-0.80	0.80	4.60	0.38	0.45	0.41	0.40	0.35
29	7.90	-2.30	0.92	7.00	0.47	0.38	0.42	0.45	0.40
30	3.00	-7.70	0.91	1.40	0.52	0.65	0.58	0.55	0.40
31	6.70	-1.80	0.95	6.40	0.33	0.33	0.32	0.31	0.10
32	6.70	-5.00	0.93	5.60	0.56	0.48	0.47	0.47	0.25
33	7.30	-8.50	0.96	6.80	0.38	0.41	0.37	0.34	0.15
34	1.30	-4.30	0.05	-4.10	0.86	0.69	0.72	0.74	0.75
35	3.00	-7.20	0.93	2.20	0.40	0.40	0.40	0.40	0.25
36	5.00	-9.10	0.08	-7.90	0.59	0.56	0.56	0.57	0.40
37	2.10	-8.40	0.80	1.30	0.29	0.29	0.26	0.24	0.10
38	6.70	-6.20	0.07	-5.10	0.62	0.55	0.57	0.59	0.65
39	7.40	-8.20	0.30	-6.90	0.88	0.75	0.76	0.76	0.85
40	6.00	-1.30	0.98	5.90	0.34	0.28	0.31	0.33	0.10
41	18.80	7.60	0.80	15.50	0.54	0.50	0.58	0.62	0.35
42	17.90	7.20	0.92	17.10	0.31	0.46	0.38	0.34	0.15
43	22.90	9.60	0.06	9.20	0.84	0.89	0.86	0.84	0.75
44	10.00	1.70	0.96	9.90	0.20	0.28	0.25	0.24	0.20
45	2.80	1.00	0.80	2.20	0.57	0.32	0.42	0.47	0.55
46	17.10	6.90	0.10	8.00	0.38	0.63	0.50	0.43	0.45
47	24.30	9.70	0.04	10.60	0.43	0.50	0.44	0.41	0.65
48	18.20	6.90	0.98	18.10	0.30	0.35	0.35	0.34	0.10
49	13.40	3.80	0.50	9.90	0.21	0.25	0.24	0.24	0.05
50	5.80	2.70	0.04	2.80	0.70	0.80	0.69	0.66	0.70
51	13.10	3.80	0.94	12.80	0.31	0.38	0.35	0.33	0.15
52	3.50	0.10	0.09	0.50	0.45	0.62	0.49	0.43	0.35
53	25.70	8.10	0.10	11.50	0.21	0.24	0.26	0.28	0.40
54	16.50	6.90	0.01	7.00	0.57	0.77	0.66	0.60	0.85
55	11.40	1.90	0.97	11.00	0.50	0.46	0.48	0.48	0.15
56	26.50	8.30	0.94	25.20	0.50	0.57	0.48	0.43	0.20
57	11.50	3.70	0.60	7.90	0.35	0.42	0.43	0.43	0.35
58	20.80	8.90	0.99	20.70	0.34	0.31	0.37	0.40	0.25
59	10.10	4.20	0.30	6.00	0.40	0.32	0.42	0.47	0.45
60	8.00	0.80	0.92	7.70	0.38	0.44	0.42	0.41	0.20

Table B3 DFE : Observed response variables The 60 problems from Estimation set in Technion Choice Prediction tournament in column 2-5. Observed Mean sampling duration and risky choice rate are presented alongside similar statistics for TE dataset as the column in the respective sub-heading. Data subset labels specified in text.

	Risky Prospect			Safe	Observed Mean Sampling Duration					Observed Risky Choice Rate				
Problems	High	Low	P(High)	Medium	RR1	RR2	RR	RR-F	TE	RR1	RR2	RR	RR-F	TE
1	-0.30	-2.10	0.96	-0.30	9.72	8.61	11.55	12.70	10.79	0.33	0.39	0.62	0.43	0.50
2	-0.90	-4.20	0.95	-1.00	14.19	11.31	11.74	11.98	10.11	0.77	0.85	0.64	0.78	0.64
3	-6.30	-15.20	0.30	-12.20	15.25	10.04	12.50	13.78	14.47	0.50	0.46	0.57	0.52	0.65
4	-10.00	-29.20	0.20	-25.60	12.22	11.41	12.51	13.15	11.16	0.37	0.37	0.59	0.39	0.58
5	-1.70	-3.90	0.90	-1.90	12.50	10.04	10.96	11.43	10.26	0.63	0.75	0.60	0.70	0.62
6	-6.30	-15.70	0.99	-6.40	10.68	8.27	11.65	13.26	10.00	0.82	0.82	0.66	0.83	0.59
7	-5.60	-20.20	0.70	-11.70	12.58	10.32	11.57	12.09	11.47	0.84	0.74	0.61	0.80	0.65
8	-0.70	-6.50	0.10	-6.00	21.92	13.35	16.36	18.07	14.11	0.42	0.42	0.60	0.33	0.54
9	-5.70	-16.30	0.95	-6.10	11.43	10.10	11.25	11.78	11.26	0.71	0.76	0.63	0.74	0.62
10	-1.50	-6.40	0.92	-1.80	12.05	16.10	13.68	12.63	12.26	0.85	0.80	0.60	0.83	0.69
11	-1.20	-12.30	0.02	-12.10	17.37	9.95	16.85	19.70	11.90	0.26	0.32	0.59	0.17	0.50
12	-5.40	-16.80	0.94	-6.40	11.89	10.11	11.91	12.65	11.15	0.79	0.89	0.63	0.80	0.57
13	-2.00	-10.40	0.05	-9.40	19.00	14.04	15.25	15.85	10.35	0.13	0.30	0.59	0.13	0.50
14	-8.80	-19.50	0.60	-15.50	12.09	9.00	11.55	12.83	12.10	0.61	0.74	0.63	0.65	0.59
15	-8.90	-26.30	0.08	-25.40	20.43	16.61	16.86	16.98	11.60	0.30	0.26	0.59	0.28	0.55
16	-7.10	-19.60	0.07	-18.70	26.96	11.87	17.78	20.74	11.00	0.13	0.13	0.62	0.20	0.52
17	-9.70	-24.70	0.10	-23.80	15.26	11.96	15.41	17.13	15.10	0.52	0.30	0.61	0.41	0.56
18	-4.00	-9.30	0.20	-8.10	18.38	16.63	16.44	16.35	11.15	0.29	0.38	0.61	0.35	0.55
19	-6.50	-17.50	0.90	-8.40	8.94	8.29	10.54	11.37	14.90	1.00	0.82	0.65	0.96	0.60
20	-4.30	-16.10	0.60	-4.50	14.62	14.38	12.10	10.80	10.85	0.27	0.23	0.53	0.26	0.48
21	2.00	-5.70	0.10	-4.60	15.67	15.14	13.15	12.24	9.05	0.33	0.33	0.58	0.26	0.58
22	9.60	-6.40	0.91	8.70	13.52	10.68	10.70	10.72	9.53	0.72	0.80	0.59	0.74	0.58
23	7.30	-3.60	0.80	5.60	13.15	10.96	11.00	11.02	11.16	0.73	0.73	0.60	0.76	0.63
24	9.20	-9.50	0.05	-7.50	9.41	7.82	11.32	12.61	15.26	0.24	0.29	0.59	0.15	0.58
25	7.40	-6.60	0.02	-6.40	16.10	10.90	13.48	14.65	8.89	0.29	0.14	0.57	0.24	0.50
26	6.40	-5.30	0.05	-4.90	16.14	12.57	14.54	15.43	13.89	0.14	0.14	0.60	0.22	0.65
27	1.60	-8.30	0.93	1.20	11.40	6.65	9.26	10.39	8.79	0.75	0.80	0.59	0.74	0.63
28	5.90	-0.80	0.80	4.60	13.82	9.68	11.49	12.59	11.05	0.57	0.75	0.65	0.65	0.64
29	7.90	-2.30	0.92	7.00	14.61	12.74	11.65	11.11	11.05	0.70	0.61	0.60	0.78	0.62
30	3.00	-7.70	0.91	1.40	15.44	13.63	13.19	12.93	10.16	0.70	0.93	0.63	0.74	0.64
31	6.70	-1.80	0.95	6.40	12.31	8.88	11.79	12.80	11.00	0.44	0.75	0.65	0.63	0.53
32	6.70	-5.00	0.93	5.60	10.00	6.64	10.08	12.17	10.95	0.71	0.82	0.64	0.70	0.54
33	7.30	-8.50	0.96	6.80	16.26	12.95	11.77	11.28	11.10	0.68	0.79	0.60	0.78	0.60
34	1.30	-4.30	0.05	-4.10	16.80	14.04	13.56	13.30	11.35	0.28	0.28	0.64	0.33	0.48
35	3.00	-7.20	0.93	2.20	9.86	10.05	10.13	10.17	12.80	0.82	0.73	0.61	0.80	0.55
36	5.00	-9.10	0.08	-7.90	19.84	14.84	14.70	14.63	14.60	0.40	0.20	0.53	0.35	0.49
37	2.10	-8.40	0.80	1.30	12.64	11.44	11.65	11.76	10.90	0.64	0.68	0.59	0.65	0.51
38	6.70	-6.20	0.07	-5.10	15.74	11.95	13.98	14.83	10.90	0.21	0.16	0.62	0.20	0.55
39	7.40	-8.20	0.30	-6.90	18.56	11.13	14.28	16.48	12.65	0.38	0.50	0.67	0.43	0.57
40	6.00	-1.30	0.98	5.90	12.31	8.50	12.05	13.28	13.50	0.81	0.81	0.67	0.89	0.61
41	18.80	7.60	0.80	15.50	11.33	9.67	10.58	10.93	9.37	0.72	0.78	0.61	0.76	0.61
42	17.90	7.20	0.92	17.10	14.52	8.10	10.33	11.35	11.11	0.57	0.62	0.66	0.74	0.67
43	22.90	9.60	0.06	9.20	10.00	6.94	10.11	11.35	10.32	0.83	0.83	0.65	0.87	0.62
44	10.00	1.70	0.96	9.90	14.30	14.04	12.67	11.98	10.47	0.83	0.74	0.66	0.85	0.57
45	2.80	1.00	0.80	2.20	13.92	10.88	10.97	11.02	20.32	0.62	0.85	0.62	0.63	0.71
46	17.10	6.90	0.10	8.00	16.13	11.04	11.96	12.41	9.37	0.17	0.35	0.55	0.22	0.60
47	24.30	9.70	0.04	10.60	13.52	9.71	12.81	14.89	12.32	0.26	0.16	0.55	0.28	0.55
48	18.20	6.90	0.98	18.10	13.46	9.00	11.53	13.07	9.37	0.71	0.64	0.66	0.74	0.64
49	13.40	3.80	0.50	9.90	15.69	9.77	11.89	13.09	9.21	0.35	0.65	0.60	0.35	0.59
50	5.80	2.70	0.04	2.80	16.17	9.22	12.06	13.48	10.32	0.35	0.35	0.58	0.33	0.52
51	13.10	3.80	0.94	12.80	10.61	7.43	9.65	10.76	8.95	0.87	0.78	0.67	0.83	0.60
52	3.50	0.10	0.09	0.50	11.52	10.14	11.82	12.59	11.85	0.29	0.19	0.63	0.37	0.44
53	25.70	8.10	0.10	11.50	14.23	11.92	12.89	13.43	9.00	0.35	0.15	0.57	0.33	0.49
54	16.50	6.90	0.01	7.00	24.60	16.35	16.26	16.22	13.40	0.30	0.10	0.59	0.24	0.56
55	11.40	1.90	0.97	11.00	15.21	12.86	12.23	11.85	9.55	0.75	0.82	0.66	0.78	0.62
56	26.50	8.30	0.94	25.20	13.70	9.39	10.81	11.52	14.25	0.74	0.83	0.65	0.76	0.61
57	11.50	3.70	0.60	7.90	14.52	10.28	11.49	12.15	10.00	0.64	0.68	0.60	0.63	0.54
58	20.80	8.90	0.99	20.70	13.31	8.96	11.35	12.70	12.90	0.69	0.69	0.63	0.72	0.65
59	10.10	4.20	0.30	6.00	14.63	12.79	12.29	12.02	10.10	0.42	0.42	0.66	0.46	0.55
60	8.00	0.80	0.92	7.70	16.83	13.54	12.07	11.30	10.20	0.83	0.75	0.65	0.78	0.54

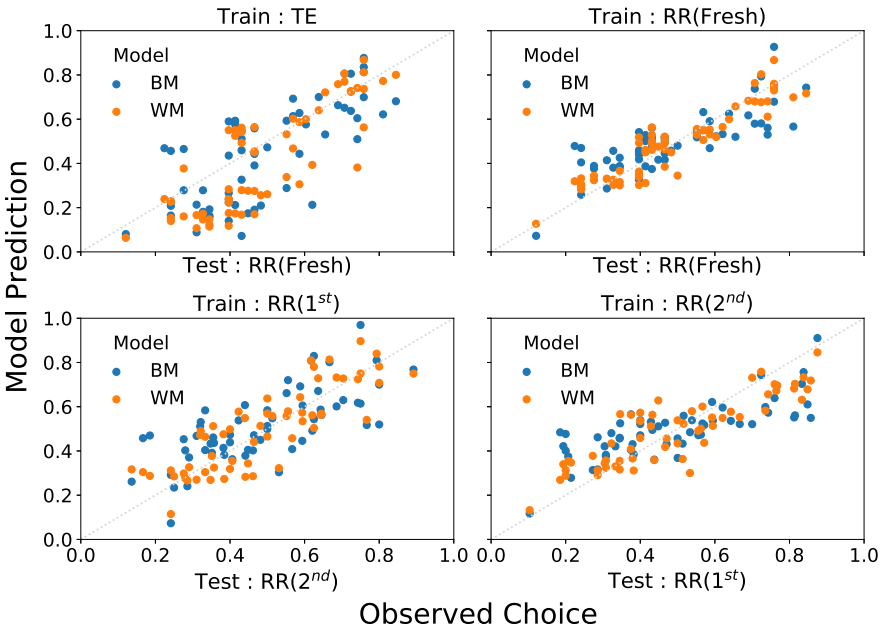

Fig. B1 Scatterplot between observed and predicted R-rate for Decisions from description data subsets Predicted R-rate by baseline (BM) and winning (WM) model from Technion Choice Prediction tournament as trained on indicated datasets, plotted against the observed choice in the test dataset (indicated in respective subplot).

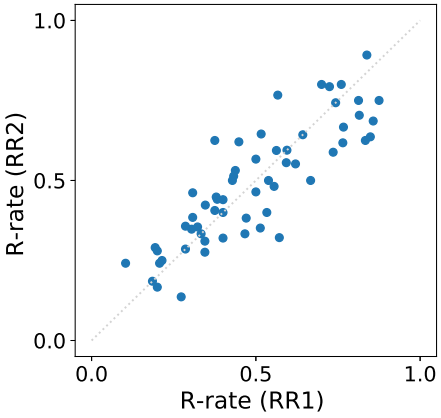

Fig. B2 Scatterplot for Retest R-rate in RR-DFD Observed risky choice rate for RR1 plotted against similar statistic for RR2 for decisions from description (DFD)