Assessing the Forecast Performance of Models of Choice

by

Dale O. Stahl

Malcolm Forsman Centennial Professor

Department of Economics

University of Texas at Austin

September 4, 2017

ABSTRACT

We often want to predict human behavior. It is well-known that the model that fits in-sample data best is not necessarily the model that *forecasts* (i.e. predicts out-of-sample) best, but we lack guidance on how to select a model for the purpose of forecasting. We illustrate the general issues and methods with the case of Rank-Dependent Expected Utility versus Expected Utility, using laboratory data and simulations. We find that poor forecasting performance is a likely outcome for typical laboratory sample sizes due to over-fitting. Finally we derive a decision-theory-based rule for selecting the best model for forecasting depending on the sample size.

Key Words: forecast performance, over-fitting, cross-validation, lottery choice

J.E.L. Classifications: C52, C53, C91, D81

1. Introduction.

The desire to understand and predict human behavior motivates theorizing and building models of choice behavior, especially for situations in which the consequences of the choices are uncertain. Von Neumann and Morgenstern (1953) introduced Expected Utility (EU) theory, which has become the mainstream model in economics. From its inception, the EU model has faced heavy criticism outside and inside economics and has been subjected to laboratory testing. However, with rare exception (e.g. Wilcox, 2008), the testing has focused on showing how well EU and alternative models fit the data rather than assessing the model's ability to forecast (i.e. predict out-of-sample). In contrast, the focus of this paper is on assessing the forecast performance of alternative models of choice under uncertainty.

We will illustrate the general issues and methods with the case of Rank-Dependent Expected Utility (RDEU) versus EU. Since the RDEU model nests EU, RDEU can fit any data it is confronted with at least as well as EU. However, a better fit does not imply a better forecast, especially given the small sample sizes provided by laboratory experiments. On a small sample, RDEU may fit significantly better than EU because the extra parameter gives it the ability to fit the noise in the sample (called "over-fitting"), which leads to biased parameter estimates; hence, the RDEU forecast could be worse than the EU forecast.

To convince the reader that over-fitting is a real danger, we will demonstrate the problem using the data from Hey and Orme (1994; hereafter HO), which is one of the first papers to confront a variety of decision theories with experimental data from a large number of well-designed choice tasks. The standard statistical method to assess over-fitting is the split-sample method of cross-validation[1]: the data is divided into an "estimation" subset and a "holdout" subset. One estimates the model parameters on the estimation subset, and then tests whether this fitted model is the data generating process (DGP) for the holdout subset. We find that the answer is negative, and we further show that RDEU forecasts worse than EU.

There are two possible explanations for this finding: (i) RDEU over-fit the data, and/or (ii) the behavior of the humans was not governed by a single DGP throughout the experiment (i.e. the behavioral process was not stationary). The second question that arises is whether the

[1] There is a voluminous literature on cross-validation: e.g. Stone (1974), and Cawley and Talbot (2010).

poor forecast performance result for the HO data is statistically significant or an artifact of this particular data.

Both of these questions can best be addressed using simulation methods. First, in a simulation, the data generation process can be held fixed for both the estimation and the holdout pseudo-data, so non-stationary behavior is ruled out as a possible explanation of poor forecast performance. Second, a simulation can generate a good approximation of the properties of any statistic for a given sample size, so the question of statistical significance can be answered without relying inappropriately on asymptotic theory. Third, a simulation can determine how large a sample should be for the over-fitting danger to be negligible. Last, but not least, the simulations can be used to find an optimal decision rule for which model to use when forecasting based on the size of the in-sample data.

Our simulation exercise demonstrates that the poor forecast performance found using the HO data does not vanish when the DGP is fixed, and that such poor forecast performance should be expected given the typical laboratory sample sizes. Our simulations also indicate that for accurate estimation and forecasting of the RDEU model, we should have 200 or more binary lottery tasks - otherwise it would be better to use the EU model. Finally, we show that a decision-theory based conditional rule about which model to use for forecasting can improve forecast performance, but still one should have at least 100 binary lottery tasks in the estimation data.

The paper is organized as follows. Section 2 specifies the RDEU models. Section 3 describes the HO experiment and measures the forecast performance on that data. Section 4 describes the simulation exercise and presents the findings. Section 5 addresses the question of how to choose a model for forecasting. Section 5 concludes with a discussion.

2. The Rank-Dependent Expected Utility Model.

A convenient encompassing model is Rank-Dependent Expected Utility[2] (RDEU) [Quiggin (1982, 1993)], which nests EU. RDEU allows subjects to modify the rank-ordered

[2] This model is the same as the Cumulative Prospect (Tversky and Kahneman, 1992) model restricted to non-negative monetary outcomes.

cumulative distribution function of lotteries as follows. Let $Y \equiv \{y_0, y_1, \ldots, y_n\}$ denote the set of potential outcomes of a lottery, where the outcomes are listed in rank order from worst to best. Given rank-ordered cumulative distribution for a lottery on Y, and let F_j denote the cumulative probability up to and including y_j. It is assumed that the subject transforms F_j by applying an increasing function $H(F_j)$ with $H(0) = 0$ and $H(1) = 1$. From this transformation, the individual derives modified probabilities of each outcome:

$$h_0 = H(F_0), \; h_1 = H(F_1) - H(F_0), \ldots, \text{and} \; h_n = 1 - H(F_{n-1}). \tag{1}$$

Common parametric specifications of the transformation functions are

$$H(F_j) \equiv (F_j)^\beta/[(F_j)^\beta + (1-F_j)^\beta] , \tag{2a}$$

$$H(F_j) \equiv (F_j)^\beta/[(F_j)^\beta + (1-F_j)^\beta]^{1/\beta} , \tag{2b}$$

$$H(F_j) \equiv (\alpha F_j)^\beta/[\alpha(F_j)^\beta + (1-F_j)^\beta] , \; \alpha > 0, \tag{2c}$$

where $\beta > 0$. Arguing from symmetry that H(0.5) should equal 0.5, Quiggin (1982) recommended eq(2a). Tversky and Kahneman (1992) suggested eq(2b) because it allows the interior fixed point to differ from 0.5. Lattimore, et al. (1992) suggested eq(2c) which allows a greater range on the shape and fixed point.[3] For ease of reference, RDEU0 will refer to the EU model (i.e. β=1 and α=1); RDEU1 will refer to the model with eq(2a), RDEU2 to the model with eq(2b), and RDEU3 to the model with eq(2c).

Given value function $v(y_j)$ for potential outcome y_j, the *rank-dependent expected utility* is

$$U(F) \equiv \sum_j v(y_j)h_j(F). \tag{3}$$

To confront the RDEU model with binary choice data (F^A *vs*. F^B), we assume a logistic choice function:

$$\text{Prob}(F^A) = exp\{\gamma U(F^A)\}/[\, exp\{\gamma U(F^A)\} + exp\{\gamma U(F^B)\} , \tag{4}$$

[3] Prelec (1998) provides an axiomatic foundation for an alternative two-parameter transformation with a fixed point of H() near 1/3; however, because we found that the Prelec specification fit the data much worse than any of the other specifications, we do not pursue this specification in this paper.

where $\gamma \geq 0$ is the precision parameter. Without loss of generality, we can assign a value of 0 to the worst outcome and a value of 1 to the best outcome.[4] Accordingly, we specify $v_0 \equiv v(y_0) = 0$ and $v_n \equiv v(y_n) = 1$. This leaves n-1 free utility parameters: $v_j \equiv v(y_j)$ for j=1,..., n-1, with the monotonicity constraint that $v_j \geq . v_{j-1}$ for j=1,..., n. Hence, the empirical RDEU0 model entails n parameters: (γ, $\underline{v}$), the RDEU1 and RDEU2 models entail n+1 parameters: (γ, $\underline{v}$, β), and the RDEU3 model entail n+2 parameters (γ, $\underline{v}$, β, α). It is obvious that RDEU3 nests RDEU1 (when $\alpha = 1$), and RDEU1 and RDEU2 nest RDEU0 (when $\alpha = 1$ and $\beta = 1$).

Next, to specify the likelihood function for our data, let $\underline{x}_i \equiv \{x_{i1}, ..., x_{iT}\}$ denote the choices of subject i for T lottery pairs indexed by $t \in \{1, ... T\}$, where $x_{it} = 1$ if lottery A was chosen, and 0 otherwise. Then the probability of the T observed choices of subject i is the product of the probability of each choice given by eq(4).[5] For notational convenience, let $\theta_i \equiv (\gamma_i, \underline{v}_i, \beta_i, \alpha_i)$. Then, in log-likelihood terms:

$$LL(\underline{x}_i, \theta_i) \equiv \sum_{t=1}^{T} \left[x_{it} \, ln\left(Prob[F^A(\theta_i)] \right) + (1 - x_{it}) \, ln\left([1 - Prob[F^A(\theta_i)]] \right) \right]. \qquad (5)$$

Then, we define the total log-likelihood of the data as

$$LL(\mathbf{x}, \boldsymbol{\theta}) \equiv \sum_i LL(\underline{x}_i, \theta_i) \, , \qquad (6)$$

where $\boldsymbol{\theta} \equiv \{\theta_i \, , i = 1, ..., N\}$, and $\mathbf{x} \equiv \{\underline{x}_i, i = 1, ..., N\}$.

[4] Since we estimate one precision parameter for all choice tasks, this scale specification is not simply the assumption of affine invariance; it is also an assumption about the magnitude of "noise" implicit in the logistic function relative to the payoffs. Wilcox (2008) argues for a re-scaling for each choice task. While we agree that re-scaling may be needed for diverse choice tasks, we feel that in the context of the HO tasks, since all four payoffs were encountered many times in succession, a re-scaling for the entire set is more appropriate. To test our intuition, we estimated the Wilcox-type EU model for the HO data (which he used), and we found it fit slightly worse than a EU model without rescaling for each task. This different finding may be due to our using only the first 100 tasks of HO and estimating individual parameters rather than a random coefficient specification.

[5] As pointed out by Harrison and Swarthout (2014), this specification implicitly assumes the "compound independence axiom". Since we view EU and RDEU as behavioral models, we are comfortable with this implicit assumption.

3. The Hey-Orme Experiment and Performance Tests.

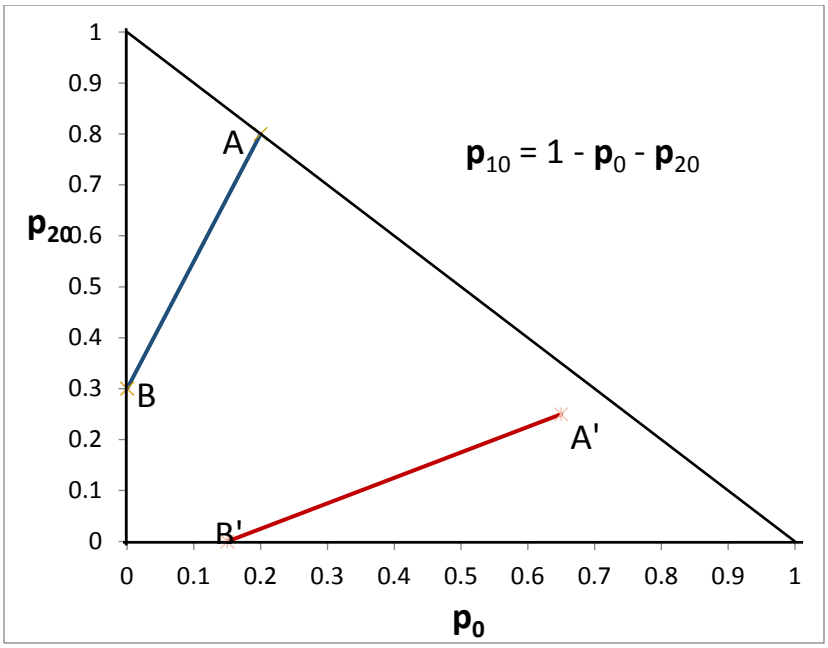
a. The Experiment.

Hey and Orme (1994; hereafter HO) is one of the first papers to confront a variety of decision theories with experimental data from a large number (100) of choice tasks.[6] Each task was a choice between two lotteries with three prizes drawn from the set {£0, £10, £20, £30}[7]. A crucial design factor was the ratio of (i) the difference between the probability of the high outcome for lottery A and the probability of the high outcome for lottery B to (ii) the difference between the probability of the low outcome for lottery A and the probability of the low outcome for lottery B. It is insightful to represent this choice paradigm in a Machina (1982) triangle, as shown in Figure 1.

Figure 1. Example of Lottery Choice Pairs

[6] These 100 tasks were presented to the same subjects again one week later. We do not consider that data here because the test that the same model parameters that best fit the first 100 choices are the same as those that best fit the second 100 choices fails. Possible explanations for this finding are (i) that learning took place between the sessions, (ii) preferences changed due to a change in external (and unobserved) circumstances, and (iii) the subjects did not have stable preferences. Therefore, we focus our attention on the first 100 choice tasks.

[7] £ is the British pound.



The ratio for the A-B pair is the slope of the dotted line connecting A and B, which is greater than 1. The ratio for the A'-B' pair (dashed line) is clearly less than 1. According to EU indifference curves are parallel straight lines with positive slope in this triangle, and the indifference curves of a risk neutral subject would have slope equal to 1. A wide range of ratios was used in order to identify indifference curves and to test the implications of EU (as well as alternative theories).

After all choices were completed, one task was randomly selected and the lottery the subject chose was carried out to determine monetary payoffs. On average, the difference in the expected monetary value of the two lotteries was about 5% of £30 = £1.5, so the expected monetary incentive for each choice task was £1.5/100 = £0.015 ≈ $0.02. To each decision theory, the authors appended a probit-like stochastic error specification, and computed maximum likelihood estimates of the model and error parameters for each of 80 subjects. They conclude: "Our study indicates that behavior can be reasonably well modelled (to what might be termed a 'reasonable approximation') as EU plus noise. Perhaps we should now spend some time on

thinking about the noise, rather than about even more alternatives to EU."[8,9] Our results reinforce their conclusion.

The low expected payoffs in this experiment raises the suspicion that the economic incentives may not have been sufficient to elicit careful effort.[10] Over-fitting is an increasing danger when there is substantial noise in the data.

b. Statistical Tests.

The estimation and holdout subsets of the data were selected as follows. Forecasting in practice typically consists of using past data to forecast future choices. Since all 100 tasks were unique, it is therefore natural to spit the data into the first 50 and second 50 tasks.[11] To find the maximum likelihood parameter estimates, we used simulated annealing (Goffe, et al., 1994) followed by the Nelder-Mead (1965) algorithm.[12] Note than for the HO experiment, since there were four distinct outcome prizes (n = 3), so there are two free utility parameters: $v_1 \leq v_2$.

i) Parameter Stability

Assuming all the data comes from a single DGP, say RDEUn, then if the data sample is sufficiently large and split into equal subsets, the maximum likelihood estimates of the model parameters on each subset should be the same. Thus, the first test of over-fitting is to estimate the parameters on the estimation subset of the data, and then test whether this fitted model is the DGP for the holdout subset. Using that fitted model, we predict the log-likelihood of the holdout subset of data: call it LL^p. Next, we estimate the parameters on the holdout subset, and compute the maximized log-likelihood of that subset: call it LL^h. To compare, we compute $LL^h - LL^p$ which is presented in Table 1. Twice this difference is asymptotically distributed Chi-square

[8] Loomes and Sugden (1998) is a similar study as Hey and Orme (1994), except that their analysis of the data is based on non-parametric tests involving the number of "reversals" and violations of dominance

[9] Wilcox (2008, 2011) uses the entire HO data to carefully study alternative stochastic specifications and his "contextual utility" model. Briefly, contextual utility essentially rescales the payoffs for each of the choice tasks to a [0, 1] scale based on the minimum and maximum payoff in that choice task. He estimates a random parameter econometric model and finds that contextual utility fits and forecasts best.

[10] The problem of flat payoff functions was forcefully raised by Harrison (1989).

[11] An alternative split into the first 75 and the last 25 yields qualitatively the same results.

[12] All the data and fortran programs for this paper can be found at https://laits.utexas.edu/~stahl/Data&Fortran.zip.

with degrees of freedom equal to the number of extra model parameters times the number of subjects (80).

Table 1: Parameter Stability Tests

	EU	RDEU1	RDEU2	RDEU3
$LL^h - LL^p$	457.49	617.84	734.86	885.72
% failure[a]	52.50	63.75	68.75	68.75

[a] Percentage of individual subjects who fail the Chi-square test

From the computed LL^h - LL^p shown in the first row of Table 1, it is clear that we can strongly reject the hypothesis that the fitted models on the estimation data subsets are the DGPs for the holdout subsets.[13] Rather than conducting the test on the sum of all the individual log-likelihood differences (as was reported in the first row of Table 1), an alternative is to compute the test for each individual subject. Doing so, and accounting for the degrees of freedom, the percentage of individual subjects who fail the Chi-square test is given in the second row of Table 1 labeled "% failure". These percentages are unacceptably high and are a strong indication of over-fitting and/or a non-stationary DGP.[14]

ii) Forecast Performance

The second test is to compare the forecast performance of the models. For each model we use the estimated the parameters for each individual subject on the estimation subset and use this fitted model to compute the log-likelihood of the holdout subset. The first row of Table 2 presents the maximized log-likelihoods (LL) summed over all subjects for the estimation subset.

[13] The p-values are essentially 0, thereby rejecting the joint hypothesis of (i) a single DGP for the first 50 and the second 50 tasks, and (ii) no over-fitting.

[14] One possible explanation for non-stationarity is that the subjects experimented with different heuristics for the first dozen or so tasks and then settled into a consistent strategy. Unfortunately, we do not have a widely accepted model of such "start-up" behavior, so any attempt to enhance the EU and RDEU models to allow for start-up behavior would be ad hoc. Investigating such enhancements can be pursued in future research.

Table 2: Estimated LL for Estimation Subset and Forecasted LL for Holdout Subset

	EU	RDEU1	RDEU2	RDEU3
LL 1st 50	-1575.84	-1458.60	-1438.50	-1395.87
LL 2nd 50 [a]	-1719.86	-1728.34	-1844.27	-1900.69
ΔF_n [b]		-8.48	-124.41	-180.83
ΔQS_n [c]		8.65	18.76	26.11

[a] LL summed over all subjects given MLE parameters from 1st 50.

[b] $\Delta F_n \equiv LL_{RDEUn} - LL_{EU}$ forecast of 2nd 50.

[c] ΔQS_n is the quadratic score, eq(7).

Clearly, on the estimation subset, the EU model fits worst and the RDEU3 model fits best.[15] The second row of Table 2 presents forecasted log-likelihoods (LL) summed over all subjects for the holdout subset using the estimated parameters for the estimation subset. The third row displays the forecast difference on the second 50: $\Delta F_n \equiv LL_{RDEUn} - LL_{EU}$. Remarkably, EU forecasts better than all the RDEUn models and RDEU3 forecasts worst. This is clear indication of over-fitting and/or a non-stationary DGP.

Since the *ln* function on [0, 1] is unbounded below, ΔF_n is unbounded. A common alternative measure of forecasting performance without this unboundedness feature is the quadratic score.[16] To construct this measure we create 26 bins {[0, 0.02), [0.02, 0.06), [0.94, 0.98) [0.98, 1]},[17] and for each bin (k) we count the number of times the forecasted Prob(F^A) of the model falls into bin k (denoted N_k) and the number of those times the subject chose lottery A (denoted NA_k). With 80 subjects and 50 tasks, we have N = 4000 forecast observations sorted

[15] On an individual basis, at the 5% significance level, we can reject EU in favor of RDEU3 for 38 out of 80 of the HO subjects.

[16] For a general introduction see Gneiting and Raftery (2007), or Clements and Harvey (2010).

[17] The two end bins are half the length of the other 24 bins to adjust for the fact that nearly half the observations land in the end bins.

into 26 bins. We then compute the weighted sum of squared deviations from the midpoint of each bin:

$$QS_n \equiv \sum_{k=1}^{25}\left(\frac{N_k}{N}\right)[(100\,NA_k/N_k) - q_k]^2, \text{ and} \qquad (7)$$

$$\Delta QS_n \equiv QS_n - QS_0, \text{ for n=1, 2 and 3,}$$

where q_k is the midpoint of bin k, expressed as a percent on [0, 100]. ΔQS_n is the difference in the quadratic score for RDEUn and the score for EU. The larger is ΔQS_n, the worse is the forecast of $RDEU_n$ relative to EU. ΔQS_n is shown in the last row of Table 2. Again, EU forecasts better than all the RDEUn models.

4. Simulations of Forecast Performance.

We turn now to simulations for three major reasons. First, since we do not know the small sample statistical properties of our forecast performance measures, simulations are required to know whether the above results are specific to the HO data or to be expected for this class of models. Second, simulations allow us to focus on the over-fitting hypothesis by fixing the DGP for both the estimation and holdout subsets of the simulated data. Third, we want to know how the sample size[18] affects forecast performance. In other words, how large should the sample size be for over-fitting to be negligible.

For a simulation study to be valid and relevant to RDEU models in general, such a study must be designed carefully. The first design question to address is what tasks should be used for the simulation study. Clearly, we want tasks for which the RDEU and EU models are likely to yield different choice probabilities. Fortunately, this design problem has been carefully considered by Hey and Orme (1994) and Harrison and Rutström (2009, hereafter HR), providing us with two sets of tasks chosen independently of our study. HO specified 100 unique lottery pairs over outcomes {0£, 10£, 20£, 30£}. HR specified 30 unique lottery pairs over outcomes {$0, $5, $10, $15}. By using the HO tasks for the in-sample estimation step, we can vary the sample size

[18] By "sample size" we mean the number of choice tasks, not the number of subjects.

from T = 1 to T = 100 without duplicating any tasks. By using the HR tasks for the forecasting step, we have a truly out-of-sample set for the forecast step.

The next design question is what parameter values should be used for the data generation process. Since we want our results to be robust to those parameters, we want a variety of parameter values. One approach would be to draw the parameter values from an uninformative prior distribution, but many of the randomly drawn parameters would be empirically irrelevant in the sense that they are unrepresentative of the distribution of parameter values in the subject population and they could confound the simulation results. To construct an empirically relevant set of parameter values, we use the maximum likelihood estimates of the parameters of the EU model (γ, v_1, v_2) on the first 100 tasks of the HO data. Since there were 80 subjects in the HO experiment, we generate 80 values of (γ, v_1, v_2). It is convenient to let $g(\gamma, v_1, v_2)$ denote this (discrete) distribution of parameter values for a typical subject pool.

a. <u>The Simulation Algorithm</u>

A simulation consists of M rounds. Round m has two steps.

1) For each value of (γ, v_1, v_2), we generate a pseudo-data sample $x_m(\gamma, v_1, v_2) \equiv \{x_{mt}, t = 1,...,T\}$ for the first T HO tasks, and we generate a pseudo-data sample $y_m(\gamma, v_1, v_2) \equiv \{y_{mt}, t = 1,...,30\}$ for the 30 HR tasks.

2) For each pseudo-data sample from step (1) we compute

   (a) the MLE of the EU model (call this $\boldsymbol{\theta}_{0m}$)[19] and the maximized $LL_0(\mathbf{x}_m)$;

   (b) the MLE of the RDEUn models (call this $\boldsymbol{\theta}_{nm}$) and the maximized $LL_n(\mathbf{x}_m)$;

   (c) $\Delta LL_{nm} \equiv LL_n(\mathbf{x}_m) - LL_0(\mathbf{x}_m) \geq 0$, n = 1, 2, and 3;

   (d) $LL_n(\mathbf{y}_m \mid \boldsymbol{\theta}_{nm})$ and

   $$\Delta F_{nm} \equiv \int [LL_n(y_m \mid \theta_{nm}) - LL_0(y_m \mid \theta_{0m})]dg(\gamma, v_1, v_2), \text{ n = 1, 2, and 3;}$$

[19] The bold typeface signifies the *set* of pseudo-data samples: one x_m and one θ_{0m} for each of the 80 values of (γ, v_1, v_2).

(e) $QS_n(\boldsymbol{\theta}_{nm})$ and $\Delta QS_{nm} \equiv QS_n(\boldsymbol{\theta}_{nm}) - QS_0(\boldsymbol{\theta}_{0m})$, n = 1, 2, and 3.

In step 2d, ΔF_{nm} is the log-likelihood of the out-of-sample simulated data for the 30 HR tasks under the RDEUn model less the log-likelihood of this out-of-sample simulated data under the EU model for simulation m averaged over the 80 values of (γ, v_1, v_2) in simulation m. In step 2e, ΔQS_{nm} is the difference between the quadratic score under the RDEUn model and the score under the EU model in simulation m.[20] A positive value of ΔQS_{nm} is a sign of poor forecast performance by RDEUn relative to EU.

We used in-sample sizes T = 25, 50, 100 and 200. For T = 200, we used each of the HO tasks twice. We set M = 200, so with 80 parameter vectors, for each T we simulated 16,000 pseudo-data sets, each consisting of T in-sample tasks and 30 out-of-sample tasks.

b. Simulation Results.

Since over-fitting as revealed by poor forecast performance is our main interest, we present the simulation results for forecast performance as revealed by ΔF and ΔQS. Specifically, we compute

$$\Delta F_n \equiv \frac{1}{M}\sum_{m=1}^{M} \Delta F_{nm} \,, \tag{8}$$

which is the mean of ΔF_{nm} over all M simulations. To estimate the variance of ΔF_n, we take the M values of ΔF_{nm} from which we compute the mean squared deviation from ΔF_n divided by M-1. The square root of this variance (or standard error) is presented in parentheses in the Tables that follow. Similarly, we compute

$$\Delta QS_n \equiv \frac{1}{M}\sum_{m=1}^{M} \Delta QS_{nm} \,. \tag{9}$$

[20] Unlike ΔF_{nm}, ΔQS_{nm} cannot be decomposed into QS measures for each (γ, v_1, v_2) because all 80 values of $\theta_{nm}(\gamma, v_1, v_2)$ are needed to ensure there are sufficient observationss in each of the 26 bins of eq(7) for each simulation round.

Table 3 displays ΔF_n and ΔQS_n for each RDEUn model and sample size T.

Table 3. Forecast Performance when DGP = EU.

	(a) ΔF_n			(b) ΔQS_n		
T	RDEU1	RDEU2	RDEU3	RDEU1	RDEU2	RDEU3
25	-10.79	-8.25	-18.76	72.23	75.70	132.8
	(0.23)	(0.13)	(0.31)	(1.27)	(1.38)	(1.7)
50	-1.29	-1.20	-3.56	10.99	11.23	29.89
	(0.09)	(0.04)	(0.11)	(0.64)	(0.67)	(0.77)
100	-0.266	-0.268	-0.663	2.85	2.58	6.04
	(0.017)	(0.010)	(0.03)	(0.44)	(0.44)	(0.52)
200	-0.084	-0.090	-0.211	0.401	0.636	1.56
	(0.004)	(0.004)	(0.007)	(0.412)	(0.424)	(0.46)

Table 3 reveals that when EU is the DGP, under both the ΔF and ΔQ criteria, all the RDEU models forecast worse than the EU model for all $T \leq 200$. Moreover, the five-parameter RDEU3 model forecasts the worst. On the other hand, as T increases the relative forecast performance of the RDEU models improves as it should.

To provide evidence that it is the extra parameters of the RDEUn models that are the major source of the over-fitting problem, Table 4 displays the root mean squared error (RMSE) of the MLE estimates of these extra parameters obtained in our simulations.

Table 4. RMSE of Parameter Estimates

	RMSE(β)			RMSE(α)
T	RDEU1	RDEU2	RDEU3	RDEU3
25	1.515	1.378	2.251	3.389
50	0.596	0.933	1.421	2.059
100	0.219	0.478	0.603	0.900
200	0.104	0.268	0.256	0.463

We see that even with a sample size of 200, the RMSEs for β and α are unusually high. Intuitively, with the increased ability to fit noise, a larger sample size is needed to reduce both the bias and the standard deviation of the estimates. In contrast, the RMSEs for the utility parameters are substantially smaller.

5. Deciding Which Model to Use for Forecasting.

While the results of Section 4b demonstrate the real and significant dangers of overfitting RDEU models it does not tell us which model to use for forecasting when the DGP is unknown. In that case we must take account of the possibility of type I errors (i.e. using the EU model when in fact the DGP is the RDEUn model). Which model to use for forecasting is a decision that should be made rationally and based on an objective function. Since the EU model can be taken as the default model, w.l.o.g. we measure forecast performance relative to the EU model. We have introduced two forecast performance measures: ΔF and ΔQS. For both measures, there are many objectives such as the expected gain of the RDEUn forecast (relative to the EU forecast). For illustration, we focusing on expected gain as the objective function, and we use our simulations to calculate the expected gain.

a. RDEUn- Always versus EU- Always.

We first compare the simple rules of *always using RDEUn*, and *always using EU*. When confronting actual data from humans, let $q_n \in (0, 1)$ denote our prior belief that RDEUn is the DGP. Since we are comparing RDEUn with EU, $1\text{-}q_n$ is our prior belief that EU is the DGP. Then, the unconditional expected forecast gain of RDEUn relative to the EU forecast is

$$E(\Delta F_n) \equiv q_n E(\Delta F_n \mid \text{RDEUn}) + (1\text{-}q_n) E(\Delta F_n \mid \text{EU}), \tag{10}$$

and similarly for $E(\Delta QS_n)$. The second term of eq(10) has been presented in Table 3. For the first term of eq(10), we use the same simulation algorithm but with RDEUn as the DGP. These latter results are given in Table 5.

Table 5. Forecast Performance when DGP = RDEUn.

	(a) $E(\Delta F_n \mid RDEUn)$			(b) $E(\Delta QS_n \mid RDEUn)$		
T	RDEU1	RDEU2	RDEU3	RDEU1	RDEU2	RDEU3
25	-6.41	-4.69	-13.18	51.39	47.66	101.0
	(0.20)	(0.11)	(0.27)	(1.11)	(1.06)	(1.7)
50	-0.128	0.037	-1.70	12.31	10.77	24.92
	(0.062)	(0.046)	(0.10)	(0.65)	(0.96)	(0.82)
100	0.528	0.628	0.621	2.82	1.90	5.13
	(0.024)	(0.019)	(0.03)	(0.49)	(0.53)	(0.69)
200	0.716	0.823	1.054	1.87	1.20	1.13
	(0.014)	(0.012)	(0.017)	(0.57)	(0.52)	(0.59)

The next step is to compute eq (10), but first we need to specify q_n. Classical hypothesis testing has an implicit prior of $q_n < 0.5$. Comparing Table 3 and 5, we see that the unconditional $E(\Delta F_n)$ is monotonically increasing in q_n and $E(\Delta QS_n)$ is monotonically decreasing in q_n. Therefore, to be as favorable to the RDEUn model as possible without violating the implicit constraint of classical hypothesis testing, we should have $q_n = 0.5$. Table 6 displays these results.

Table 6. Unconditional $E(\Delta F_n)$ and $E(\Delta QS_n)$ given $q_n = 0.5$.

	(a) $E(\Delta F_n)$			(b) $E(\Delta QS_n)$		
T	RDEU1	RDEU2	RDEU3	RDEU1	RDEU2	RDEU3
25	-8.60	-6.48	-15.97	61.81	61.68	116.9
	(0.18)	(0.09)	(0.24)	(0.97)	(0.99)	(1.3)
50	-0.708	-0.584	-2.18	11.65	11.00	27.4
	(0.047)	(0.032)	(0.09)	(0.46)	(0.49)	(0.58)
100	0.131	0.180	-0.021	2.84	2.24	5.59
	(0.017)	(0.011)	(0.024)	(0.34)	(0.34)	(0.44)
200	0.316	0.367	0.421	1.14	0.92	1.35
	(0.007)	(0.006)	(0.009)	(0.35)	(0.33)	(0.38)

First observe from Table 6 that under both the ΔF and ΔQS criteria, EU-Always is better than RDEUn-Always for all $T \leq 200$. Second, from eq(10) and Tables 3 and 5, we deduce that $E(\Delta F_n)$ decreases and $E(\Delta QS_n)$ increases as q_n decreases below 0.5; i.e. EU-Always forecasts better than RDEUn-Always for all $q_n \leq 0.5$.

b. A Conditional Rule.

Next we consider a decision rule that selects the RDEUn model or the EU model for forecasting based on the log-likelihood difference ΔLL_{nm} on the estimation dataset. Specifically, consider the *conditional rule* (CR) to use RDEUn for out-of-sample forecasting only if ΔLL_{nm} is statistically significant. For example, suppose we use the commonly applied criteria that $2\Delta LL_{nm}$ exceed the inverse chi-squared value for the 5% level, given one degree of freedom for RDEU1 and RDEU2, and two degrees of freedom for RDEU3; call these thresholds $\Delta LL5_n$.[21]

We turn now to computing the expected gain for the CR rule. Since CR is the same as the EU-always rule whenever $\Delta LL_{nm} \leq \Delta LL5_n$ on the estimation dataset, the forecast performance measures ΔF_{nm} and ΔQS_n differ from zero only when $\Delta LL_{nm} > \Delta LL5_n$. Accordingly, we can focus on two subsets of the simulations: let M_0 denote the subset of simulations using EU as the DGP and $\Delta LL_{nm} > \Delta LL5_n$, and let M_n denote the set of simulations using RDEUn as the DGP and $\Delta LL_{nm} > \Delta LL5_n$. Thus, our estimate of the forecast performance of CR when DGP = EU is

$$E(\Delta F_n \mid EU) \equiv \frac{1}{M}\sum_{m\in M_0}\int \Delta F_{nm}\, dg(\gamma, v_1, v_2), \tag{11a}$$

and when DGP = RDEUn is

$$E(\Delta F_n \mid RDEUn) \equiv \frac{1}{M}\sum_{m\in M_n}\int \Delta F_{nm}\, dg(\gamma, v_1, v_2). \tag{11b}$$

Given a prior q_n that RDEUn is the DGP for the human data, the unconditional expectation of the forecast performance of CR is

$$E(\Delta F_n \mid CR) = q_n\, E(\Delta F_n \mid RDEUn) + (1-q_n)\, E(\Delta F_n \mid EU). \tag{12}$$

Obviously, analogous formulae apply when ΔQS_n is the forecast measure.[22] Table 7 shows these expected forecast differences for the CR rule given $q_n = 0.5$.

[21] I.e. $LL5_1 = LL5_2 = 1.92$, and $LL5_3 = 3.00$.

[22] When DGP is EU, we expect $\Delta LL_{nm} > \Delta LL5_n$ for 10% or fewer simulations (given the sample sizes), so 26 bins for QS is too many - many bins contain less than five observations which is too few for accurate frequency

Table 7 Expected Forecast Differences for CR Rule given $q_n = 0.5$.

	(a) $E(\Delta F_n \mid CR)$			(b) $E(\Delta QS_n \mid CR)$		
T	RDEU1	RDEU2	RDEU3	RDEU1	RDEU2	RDEU3
	-2.79	-2.12	-5.05	17.96	24.06	30.73
25	(0.11)	(0.11)	(0.14)	(0.29)	(0.30)	(0.10)
	-0.341	-0.239	-1.19	6.20	5.95	9.28
50	(0.034)	(0.026)	(0.06)	(0.25)	(0.33)	(0.28)
	0.229	0.243	0.201	1.03	-0.42	0.23
100	(0.011)	(0.009)	(0.094)	(0.44)	(0.27)	(0.34)
	0.327	0.347	0.478	-0.115	-1.54	-1.91
200	(0.001)	(0.007)	(0.008	(0.372)	(0.37)	(0.41)

First observe that under both the ΔF and ΔQS criteria, EU is the best model for forecasting when $T \leq 50$, and CR is the best model for forecasting when $T \geq 200$. For intermediate estimation sample sizes ($50 < T < 200$), the ΔF criteria favors EU while the ΔQS criteria favors CR for forecasting.

It is natural to be curious about how such a conditional rule would fare for the actual HO data. Accordingly, we revisit the exercise of forecasting the second 50 tasks of the HO experiment based on the maximum likelihood estimates for the first 50 tasks, as presented in Table 2. Instead of using the unconditional rule of using RDEUn for all subjects or EU for all subjects, we implement the conditional rule of using the RDEUn model for forecasting iff $dLL_n > dLL5_n$.

Table 11. Forecast Performance of the CR for the HO Experiment Data

	RDEU1	**RDEU2**	**RDEU3**
ΔF_n	25.66	34.85	-80.05
ΔQS_n	6.23	7.21	21.21

estimates. To compute a more reliable measure of QS, we used 14 bins: [0, 0.02), [0.02, 0.10), ..., [0.9, 0.98), [0.98, 1]. We keep the end bins unchanged because about half of all observations fall into those end bins. By doubling the size of the in-between bins, almost all bins contain at least five observations.

From the first row of Table 11, we observe that under the ΔF criterion the conditional rule yields an improvement over using the EU model for the four parameter RDEU1 and RDEU2 models. That the five parameter RDEU3 model performs worse than the other models is consistent with our simulations. The last row of Table 11 shows that EU forecasted better than all RDEU models under the ΔQS criterion - again consistent with our simulations.

6. Conclusions and Discussion.

Our findings strongly suggest that with the sample sizes available from laboratory experiments, RDEU models are prone to over-fit the data and forecast poorly. If we value forecast performance and have a limited sample size for estimation ($T < 200$), our analysis suggests that one should use the simple EU model rather than an RDEU model.

The central lesson is that it is vitally important to guard against over-fitting data. Of course, if we had a very large number of observations for each individual, over-fitting would eventually cease to be a non-negligible problem. Our simulations indicate that for accurate estimation and forecasting of the RDEU model, it is advisable to have 200 or more binary lottery tasks - otherwise it would be better to use the EU model.

Unfortunately, there are natural limits to the sample size from laboratory experiments. The data set used in this paper is quite large in comparison to most experiment data. Obtaining hundreds of observations for each individual invites behavioral noise via fatigue and boredom. Collecting observations from the same individual over several sessions opens the door to learning through talking with friends and experts, and changes in preferences due to changing individual circumstances.

Furthermore, there is the problem that providing monetary incentives for the choice tasks and identifying indifference curves are conflicting objectives. [23] The latter requires lotteries for which the subject is nearly indifferent, but consequently the monetary incentives for such choice

[23] Harrison (1992) and Harrison and Rutström (2008) make a similar argument.

tasks are very small. This limitation is not restricted to the multiple choice design of the HO and HR experiments. [24]

Given these limitations for laboratory data, there is a limit to the information we can obtain about indifference curves in lottery space. Consequently, there is a corresponding limit to how confident we can be in our theories about lottery choices.

[24] For an extensive critique, see Harrison and Rutström (2008).

References

Cawley, G., and Talbot, N. (2010). "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *Journal of Machine Learning Research,* **11**, 2079-2107.

Clements, M., and Harvey, D. (2010). "Forecast Encompassing Tests and Probability Forecasts," *Journal of Applied Econometrics*, **25**, 1028-1062.

Gneiting, T., and Raftery, A. (2007). "Strictly Proper Scoring Rules, Prediction and Estimation," *Journal of the American Statistical Association*, **102**, 359-378.

Goffe, W., Ferrier, G., and Rogers, J. (1994). "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, **60**, 65-100.

Grether, D. and Plott, C. (1979). "Economic Theory of Choice and Preference Reversal Phenomena," *Am. Econ. Rev.*, **69**, 623-638.

Harrison, G. W. (1989). "Theory and Misbehavior of First-Price Auctions," *Am. Econ. Rev.*, **79**, 749-762.

Harrison, G. W. (1992). "Theory and Misbehavior of First-Price Auctions: Reply," *Am. Econ. Rev.*, **82**, 1426-1443.

Harrison, G. W. and Rutström, E. (2008). "Risk Aversion in the Laboratory," In J. C. Cox and G. W. Harrison, eds., *Research in Experimental Economics,* **12**, 41-196.

Harrison, G. W. and Rutström, E. (2009). "Expected Utility and Prospect Theory: One Wedding and Decent Funeral," *Experimental Economics*, 12, 133-158.

Harrison, G. W., Swarthout, J. T. (2014). "Experimental Payment Protocols and the Bipolar Behaviorist," *Theory and Decision,* 77, 423-438.

Hey, J. and Orme, C. (1994). "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica,* 62, 1291-1326.

Lattimore, P., Baker, J., and Witt, A. (1992). "The Influence of Probability on Risky Choice: A Parametric Examination," *J. of Econ. Behavior and Organization,* 17,377-400.

Loomes G., and Sugden, R. (1998). "Testing Alternative Stochastic Specifications for Risky Choice," *Economica*, **65**, 581-598.

Machina, M. (2008). "Non-expected Utility Theory," in *The New Palgrave Dictionary of Economics* (2nd edition). Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan.

Nelder, J., Mead, R. (1965). "A Simplex Method for Function Minimization," *Computer Journal* **7**, 308-313.

Prelec, D. (1998), "The Probability Weighting Function," *Econometrica*, 66, 497-527.

Quiggin, J. (1982). "A Theory of Anticipated Utility", *J. of Econ. Behavior and Organization*, **3**, 323-343.

Quiggin, J. (1993). *Generalized Expected Utility Theory: the Rank-Dependent Model*, Kluwer Academic Publishers.

Stahl, D. O. (2016). “Framing Lottery Choices,” working paper, University of Texas at Austin.

Stone, M. (1974). “Cross-Validatory Choice and Assessment of Statistical Predictors,” *J. of Royal Statistical Society, Series B,* **36**, 111-147.

Tversky, A., and Kahneman, D. (1992). "Cumulative Prospect Theory: An Analysis of Decision Under Uncertainty," *Journal of Risk and Uncertainty*, **5**, 297-323.

von Neumann, J. and Morgenstern, O. (1953), *Theory of Games and Economic Behavior*, Princeton University Press.

Wilcox, N. (2008). “Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison”. In J. C. Cox and G. W. Harrison, eds., *Research in Experimental Economics*, **12**, 197-292.

Wilcox, N. (2011). “Stochastically more risk averse:’ A contextual theory of stochastic discrete choice under risk,” *Journal of Econometrics*, **162**, 89–104.