Statistical Power and the Individual Level Estimation of Risk Preferences

Brian Albert Monroe*

March 26, 2019

Abstract

Accurately estimating risk preferences is of critical importance when evaluating data from many economic experiments or behavioral interactions. I conduct power analyses over two lottery batteries designed to classify individual subjects as one of a number of alternative specifications of risk preference models. I propose a conservative case in which there are only two possible alternatives for classification and find that the statistical methods employed to conduct this classification result in type I and type II errors at rates far beyond traditionally acceptable levels. Following a Bayesian approach, I additionally find that the proportion of agents in a population that employ each model critically informs the probability that subjects are correctly classified.

*School of Economics, University of Cape Town, South Africa.
Email: brianalbertmonroe@gmail.com

Special thanks to Glenn Harrison, Don Ross, and Andre Hofmeyr for providing comments and feedback for this paper.

1 Introduction

In response to growing evidence that some subjects in economic experiments violate one or more axioms of Expected Utility Theory (EUT), several alternative models were proposed which allow for the apparent violations. Prospect Theory (Kahneman and Tversky 1979), Rank Dependent Utility (RDU) (Quiggin 1982), and Regret Theory (Bell 1982; Loomes and Sugden 1982) are among the best known of these alternative models. Many of the newly proposed theoretical explanations of the apparent violations of EUT have been tested experimentally. A well known example is the experiment of Hey and Orme (1994) (HO) to test if any of a variety of generalizations (and one restriction) of EUT can explain experimentally collected data significantly better than EUT. HO picked "winning" model specifications for each of their subjects on the basis of the estimates of each model and whether each model can be statistically distinguished from EUT using information criteria that punished the use of additional parameters. They conclude, "our study indicates that behavior can be reasonably well modeled (to what might be termed a 'reasonable approximation') as 'EU plus noise.'"

However, HO raise concerns that as the number of alternative specifications being tested increases, the probability that EUT will be selected as the "winning" model will decline, *even if EUT is the correct specification.* These concerns relate to statistical power, and to the weight economists should place on type I versus type II errors. The degree of confidence in the process employed by HO to pick winning models, and indeed most statistical tests in the economics literature, can be assessed through power analyses.

Power analyses are rarely conducted in parallel with econometric estimation.

McCloskey and Ziliak (1996, p. 105) find that only 4.4% of the 182 papers published in *The American Economic Review* in the 1980s reported the power of the test they were performing. Zhang and Ortmann (2013, p. 6) review all papers published in *Experimental Economics* for the years 2010-2012, and find that no paper stated the optimal sample size for their analyses, and only one paper mentions power as an issue.

Retroactive power analyses of published research and attempted replication of experiments has led to a recent reconsideration of claims of statistical significance in published research across many fields. De Long and Lang (1992) propose a measure of the fraction of unrejected null hypotheses that are, in fact, false, in economic journal articles, and infer that less than one third of unrejected null hypotheses are true. Ioannidis (2005) bluntly notes that in the medical sciences "It can be proven that most claimed research findings are false." Gelman and Loken (2014, p. 460) write, "There is a growing realization that reported 'statistically significant' claims in scientific publications are routinely mistaken."

Continuing the scrutiny around claims of statistical significance, I conduct power analyses of the ability of two risky lottery batteries to correctly distinguish between two possible data generating processes (DGPs), an EUT model and an RDU model. I analyze the original lottery battery proposed by HO and the battery proposed by Harrison and Ng (2016) (HN). The subjects in both studies made choices across many lottery pairs, a "winning" model was selected for each subject on the basis of that subject's choices, and the selected model was critical to the inferential objective of the study.

HO was directly concerned with whether their subjects systematically deviated from EUT, while HN was directly concerned with inferences about the consumer surplus for each of their subjects. The effect of deviations from EUT are of critical importance to the evaluation of subjective beliefs (Andersen, Fountain, Harrison and Rutström 2014), the calculation of consumer surplus (Harrison and Ng 2016), the calculation of discount factors (Andersen, Harrison, Lau and Rutström 2008), the applicability of the reduction of compound lotteries axiom (Harrison, Martínez-Correa and Swarthout 2015), and behavior in strategic interactions.

I begin by briefly describing the EUT and RDU DGPs, followed by a description of the power analysis process. I then discuss the experimental designs and inferential objectives of HO and HN. Finally, I present the results of a power analysis of the ability of the two experimental batteries used in HO and HN to classify subjects as either EUT or RDU, and apply these results to a hypothetical population using Bayes' Theorem.

2 The Data Generating Processes

HO and HN select a "winning" model specification for each of their subjects from 11 and 4 different candidate specifications, respectively. As noted by HO, as the number of alternatives increase, the frequency at which EUT is rejected as the true DGP will also increase, even if EUT is the true DGP. Accordingly, a conservative case for model classification is presented in which choice data produced by subjects can only be generated by two possible DGPs and experimenters seek only to discern which of these two DGPs an individual subject employs. The first of these two DGPs is an EUT model and the other is an RDU model with a probability weighting function (PWF) due to Prelec (1998). Since RDU nests

EUT as a special case, both DGPs can be defined using the RDU framework:

$$RDU = \sum_{c=1}^{C} [w_c(p) \times u(x_c)] \tag{1}$$

where c indexes the outcomes, x_c, of a lottery from $\{1, \ldots, C\}$ with $c = 1$ being the smallest outcome in the lottery and $c = C$ being the greatest outcome in the lottery, $u(\cdot)$ is a standard utility function, $w_c(\cdot)$ is a decision weight function associated the probability of outcome c given the distribution of probabilities in the lottery ranked by outcome, p. The decision weight function, $w_c(\cdot)$, takes the form:

$$w_c(p) = \begin{cases} \omega\left(\sum_{k=c}^{C} p_k\right) - \omega\left(\sum_{k=c+1}^{C} p_k\right) & \text{for } c < C \\ \omega(p_c) & \text{for } c = C \end{cases} \tag{2}$$

where the PWF, $\omega(\cdot)$, can take a variety of parametric or non-parametric forms. The special case of EUT, where the PWF gives the objective probabilities, is used as the first DGP:

$$\omega(p_c) = p_c \tag{3}$$

and an RDU model with the two parameter PWF proposed by Prelec (1998) as the second DGP:

$$\omega(p_c) = \exp(-\eta(-\ln(p_c))^{\phi}) \tag{4}$$

where $\phi, \eta > 0$.

To complete the model in (1), the utility function is defined as the constant relative risk aversion (CRRA) function:

$$u(x) = \frac{x^{1-r}}{1-r} \tag{5}$$

To account for randomness in the choices of real subjects, the RDU model in

(1) is combined with a stochastic specification in which the preference of A over B is related to the probability of A being chosen over B:

$$A \succeq B \Rightarrow Pr(A) \geq Pr(B) \tag{6}$$

The Contextual Utility (CU) stochastic model of Wilcox (2011) is used to relate the RDU of an option to its choice probability. Thus the probability that option A is chosen is given by:

$$\begin{aligned} Pr(A) &= Pr\left(\epsilon \geq \frac{1}{\lambda}\left[RDU(A) - RDU(B)\right]\right) \\ &= F\left(\frac{RDU(A) - RDU(B)}{D(A,B)\lambda}\right) \end{aligned} \tag{7}$$

where ϵ is a mean 0 error term, F is a symmetric cumulative distribution function (cdf) and λ is a precision parameter. The function $D(\cdot)$ provides the "contextualization" that gives CU its namesake and is defined as the difference between the utility of the maximum and minimum possible outcomes across lotteries A and B:

$$D(A,B) = max[u(x)] - min[u(x)], \quad st.\ w(p) \neq 0 \tag{8}$$

The logistic cdf is used for F for all calculations. Given that each choice considered here only involves two options, the probability of choosing option A can be defined as a multinomial logit function:

$$Pr(A) = \frac{\exp\left(\dfrac{RDU(A)}{D(A,B)\lambda}\right)}{\exp\left(\dfrac{RDU(A)}{D(A,B)\lambda}\right) + \exp\left(\dfrac{RDU(B)}{D(A,B)\lambda}\right)} \tag{9}$$

The two data generating processes therefore consist of an EUT model and an RDU model which have the utility function and stochastic specification in common, and differ only by the treatment of decision weights in (1).

3 The Studies Under Consideration

3.1 Hey and Orme (1994)

HO conducted an experiment over four days in which 80 subjects completed a single task on each day. In two of these tasks, *Circles 1* and *Circles 2*, subjects were presented with 100 lottery pairs on a computer screen and asked whether they would prefer to play out the lottery on the left, the lottery on the right, or if they didn't care which lottery would be played out. Subjects were told that once they had answered all 100 questions, one would be chosen at random and their choice played out for money. If they selected "don't care" the experimenter selected which lottery was played out. The 100 lottery pairs comprised 25 lottery pairs repeated 4 times with the order of the pairs presented to the subjects at random. The same lotteries were used in *Circles 1* and *Circles 2*, with the order of the pairs re-randomized and the position of the lottery on the screen randomly reversed.

HO estimate 11 different model specifications of choice under risk using maximum likelihood (ML) on the data from *Circles 1*, *Circles 2*, and *Circles 1* and *Circles 2* combined (*Circles 3*) for each subject. Of these specifications, one was a risk neutral (expected value) model, one was an EUT model, and two were RDU specifications. Of the remaining seven specifications, all but one nested EUT and expected value as special cases.

The process of choosing a "winner" across all 11 model specifications involved testing for a statistical difference from nested models using a likelihood ratio test, and then ranking the specifications on the basis of the Akaike information criteria (AIC).[1] First, all of the models were tested to see if they were statistically different from expected value at the 1% level; if none were, expected value won. If EUT and at least one of the 8 specifications that nested EUT significantly deviated from expected value, then the non-EUT models were tested to see if they deviated from EUT; if none did, EUT won. If only one deviated from EUT, it won, but if two or more specifications were different from EUT, the winner was chosen on the basis of the AIC.

HO report that the EUT model generally wins across more subjects in the *Circles 1* and *Circles 2* datasets than any other model, though it does not win for a majority of subjects across any single dataset. HO also report that for any given binary test of EUT and a specification that nests EUT, EUT cannot be rejected at the 1% level as the DGP for more than half of their subjects using any data set.

3.2 Harrison and Ng (2016)

HN conduct an experiment in which 111 subjects responded to two tasks. In the first task subjects were asked to make binary choices over 80 lottery pairs. In the second task subjects were asked to make 24 binary choices in an insurance task. A "winning" model specification estimated from the first task was then used to calculate the consumer surplus of the choices made in the insurance task.

The battery of lotteries was specifically designed to establish whether experimental subjects' behavior was more consistent with EUT or some RDU specification. HN follow the design of Loomes and Sugden (1998) in this regard, with 40 lottery

[1] The Akaike information criteria is given by $AIC = -2logL(\hat{\alpha})/T + 2k/T$, where $L(\hat{\alpha})$ is the log-likelihood of the model at its estimated maximum, k is the number of parameters for that model, and T is the number of observations.

pairs on the border of a Marschak-Machina (MM) triangle, and 40 lottery pairs in the interior. HN estimate three RDU and one EUT specification for every subject. All four specifications employed the CRRA function defined in (5) as the utility function, and the CU stochastic function defined in (7) and (8). For the three RDU specifications, HN employed as the PWFs the "power" function, the "Inverse-S" function, and the two parameter function described in (4).

To select a winner, HN first used a non-linear Wald test to determine if the PWF of each RDU specification was significantly different from a linear function, the special case of EUT. If not, the RDU model was dropped from consideration. Of those RDU models remaining, the model with the greatest log-likelihood was selected as the "winner." Using this process, HN found that EUT won for nearly half of their subjects, with the RDU model employing the Prelec (1998) PWF a close second, and the RDU models employing the "Inverse-S" and "Power" PWFs distant runners up.

4 Power Analysis Procedure

HO and HN both classified their subjects by first testing if the RDU specification was statistically different from EUT, and then selected a "winner" on the basis of either the log-likelihood for HN or the AIC criterion for HO. The null hypothesis for both studies was that the subject did not employ probability weighting. The following power analyses estimate the probability that a subject with an EUT DGP is falsely classified as employing an RDU DGP, a type I error, and the probability that a subject with an RDU DGP is falsely classified as employing an EUT DGP, a type II error.[2] With two possible DGPs, and two models to estimate, there are four possible results of a classification when both models have converged, shown in Table 1.

Table 1: Possible Results of Classification

	EUT DGP	RDU DGP
Classified EUT	Null correctly unrejected	Type II error
Classified RDU	Type I error	Null correctly rejected

Simulation methods similar to those described by Feiveson (2002) are used to analyze the power of the batteries used by HO and HN. Feiveson (2002, p. 108) briefly describes a simulation method for determining the power of an experiment by repeatedly generating hypothetical data, and then calculating the proportion of rejections of the null hypothesis as an estimate of power.

A simulated subject is represented by an assigned DGP and an associated set of parameters. Each DGP uses the CRRA utility function defined in (5) and the CU stochastic model in equations (7) and (8), with the RDU model additionally employing the PWF defined in (4). For EUT subjects, the parameter set consists of $\{r, \lambda\}$, and for RDU subjects $\{r, \phi, \eta, \lambda\}$, where r gives the CRRA parameter, λ the CU precision parameter, and ϕ and η the probability weighting parameters.

Each DGP's parameter sets are drawn from a joint uniform distribution with uncorrelated marginal distributions over the parameters needed. For the EUT

[2] Typically, when a test indicates the probability of a type I error to be less than 5%, social scientists consider this result "statistically significant," and when researchers engage in *ex ante* power analysis, they typically aim for a probability of a type II error less than 20% (Cohen 1988; Gelman and Loken 2014). These values are based on convention, and are somewhat arbitrary. Ronald Fisher disagreed with picking the same level of statistical significance for every analysis: "[...] no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (Fisher 1956).

DGP, the marginal distribution for r is $r \in [0, 1]$ and for λ is $\lambda \in [0.05, 0.30]$. For the RDU DGP the marginal distributions are $r \in [0.4, 0.6]$, $\lambda \in [0.1, 0.15]$, $\phi \in [0.5, 2.5]$ and $\eta \in [0.5, 2.5]$. These values roughly conform to the ranges of parameter estimates on data generated by real, human subjects.[3] The marginal distributions of the r and λ distributions are narrower for the RDU DGP in order to focus on how the probability weighting parameters affect the classification of RDU subjects.

The simulation process is as follows. First, a simulated subject is assigned a DGP, either EUT or RDU, and a set of parameters is drawn from the associated joint distribution defined above. Second, for each battery, the choice probability of lottery A and lottery B is calculated for every lottery pair using the assigned DGP and the associated parameter set drawn for the subject. Finally, a random number is then drawn from an univariate uniform distribution. If the choice probability calculated for the A option exceeds the random number the subject "chooses" A, otherwise they choose B. This ensures that choices are made probabilistically with respect to the subject's assigned DGP and drawn parameter set.[4]

The EUT and RDU models are then estimated over the simulated subject's choices. Any model that does not converge with a gradient close to 0 and a positive definite Hessian matrix is dropped. Subjects are then classified as one of the two possible models following the HN classification process. To classify subjects, the probability weighting parameters are jointly tested for equality to 1, and then the log-likelihoods of the EUT and RDU model are compared. If the p-value of the

[3] See the Appendix of HN for estimates of typical university students in the United States, and Harrison and Rutström (2008) for additional reviews of studies with real human subjects.

[4] Consider a choice probability calculated to be 0.90 for option A, and therefore 0.10 for option B. A random number drawn from an univariate uniform distribution has a 90% chance of being less than or equal to 0.90, so option A would be chosen 90% of the time by the simulated subject.

test of the probability weighting parameters is not less than 5%, the subject is classified as EUT. If it is less than 5%, the subject is classified as the model with the greatest log-likelihood. If only one of the two models converge, the subject is classified as the converged model, if neither model converge, the subject is not classified and is dropped from the dataset.

5 Results

For each DGP, the probabilities of being correctly classified are presented for the HO and HN batteries For the EUT DGP the simulated $\{r, \lambda\}$ parameter space is partitioned into a 16 × 16 equally spaced grid. For the RDU DGP the simulated $\{\phi, \eta\}$ parameter space is partitioned into a 16 × 16 equally spaced grid. For the EUT DGP this grid represents the entire parameter space needed to define the model, but for the RDU DGP this grid only shows variation across the probability weighting parameters.

5.1 EUT DGP

Figures 1 and 2 show how the probability of being correctly classified varies across r and λ for the HO and HN batteries, respectively. In both figures, darker colors represent lower probabilities of correct classification and lighter colors represent higher probabilities of correct classification. The probability that a subject is correctly classified is displayed numerically every third cell.

Figures 1 and 2 show that the HO and HN batteries have similar patterns of correct classification across the r and λ parameters. As expected, as λ increases, the probability that a subject with an EUT DGP is correctly classified monotonically decreases. The “noisiness” of the data is directly influenced by λ; as λ rises, so does the noise. As the noise in any data increases, we expect the probability of a type I error to increase.

For both batteries, the probability correct classification peaks for values of r between 0.4 and 0.6, the middle of the range of r values considered. In other words, the power of these batteries to correctly classify a subject as EUT is lower for subjects that are either very risk averse or not risk averse at all compared to subjects that are moderately risk averse, though not by much in the considered range of parameters.

In general, the affect of the r and λ parameters on the probability of correctly classifying a subject are as expected. There is every reason to believe that as data gets noisier, the probability of type I errors increases (the effect of the λ parameter), and that the power of an battery to identify risk aversion is greatest in the ranges indicating moderate risk aversion (the effect of the r parameter). However, of greater interest are the absolute probabilities of correct classification.

Consider the range of parameters for the EUT DGP where $r \in (0.31, 0.37)$ and $\lambda \in (0.10, 0.11)$.[5] Subjects in this range would fall in the cell four columns from the left and six rows from the bottom in Figures 1 and 2, outlined by a red square in each figure. Subjects in this range have a 92% chance of being correctly classified as EUT with the HO battery and an 85.22% chance with the HN battery. This implies a type I error rate of 8% for the HO battery and 14.78% for the HN battery.

Both batteries have wide ranges of parameter values where the probability of a type I error exceeds the typically required 5%. The probability of a type I error can be as high as 23.51% for the HO battery and 35.43% for the HN battery at the

[5] The r parameter of subject 8 given as an example in HN (pg. 104) would fall in this range.

extreme edges of the parameter ranges considered here. But within the range of r values that give both batteries their greatest power, $r \in [0.31, 0.62]$, and the range of λ values that are generally estimated from human subject data, $\lambda \in [0.10, 0.16]$, the probability of a type I error is 7.74–14.82% for HO and 13.73–25.6% for HN. Both batteries have rates above the 5% typically cited as the threshold for statistical significance in the social sciences.

5.2 RDU DGP

Now presenting the results for the RDU DGP, the relevant parameter space is again partitioned into an equally sized 16×16 grid. This time, however, Figures 3 and 4 show how the probability of an RDU subject being correctly classified changes with the probability weighting parameters, ϕ and η.

In contrast to the r and λ parameters in Figures 1 and 2, the ϕ and η parameters in Figures 3 and 4 show more interaction in determining the probability of RDU subjects being correctly classified. When $\phi = \eta = 1$, the PWF becomes linear and the RDU model reduces to EUT. As expected, the probability of correctly classifying RDU subjects is extremely low near these values for both batteries, and at it greatest for values of ϕ or η much larger, or much smaller than 1.

Even when excluding parameter values $0.75 < \phi, \eta < 1.38$, i.e. excluding values near the special case where RDU reduces to EUT, the probability of a type II error[6] can be up to 93.44% and is never lower than 24.11% for the HO battery and up to 88.58% and never lower than 37.51% for the HN battery.

For a subject with $\phi = .74$ and $\eta = .70$,[7], outlined in red in Figures 3 and 4,

[6] Recall that a type II error in this analysis is 1 minus the probability of correctly classifying an RDU subject.

[7] These are the values estimated for subject 8 in HN (p. 104).

Figure 1: Hey and Orme (1994)

Probability of Correct Classification, EUT DGP

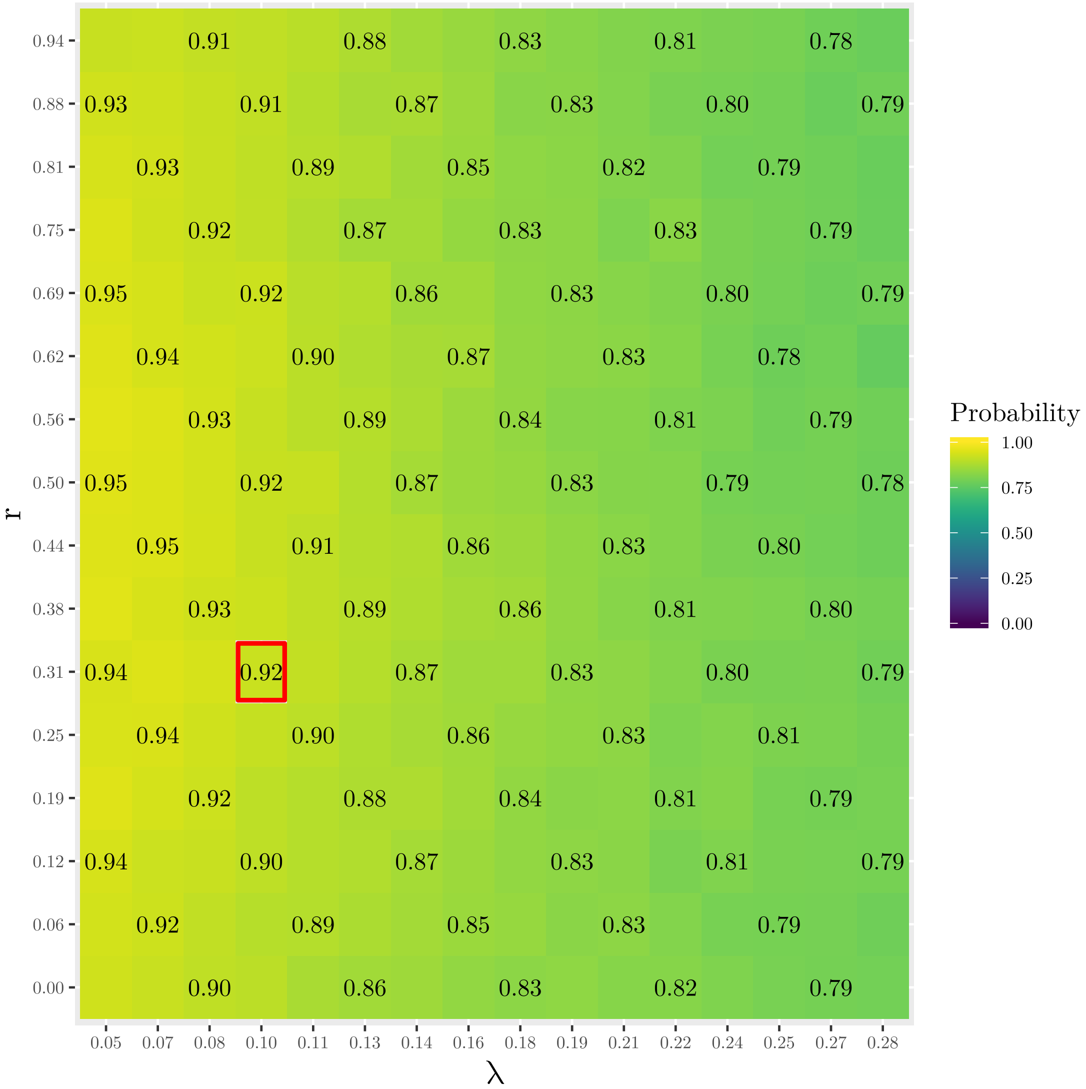


Figure 2: Harrison and Ng (2016)

Probability of Correct Classification, EUT DGP

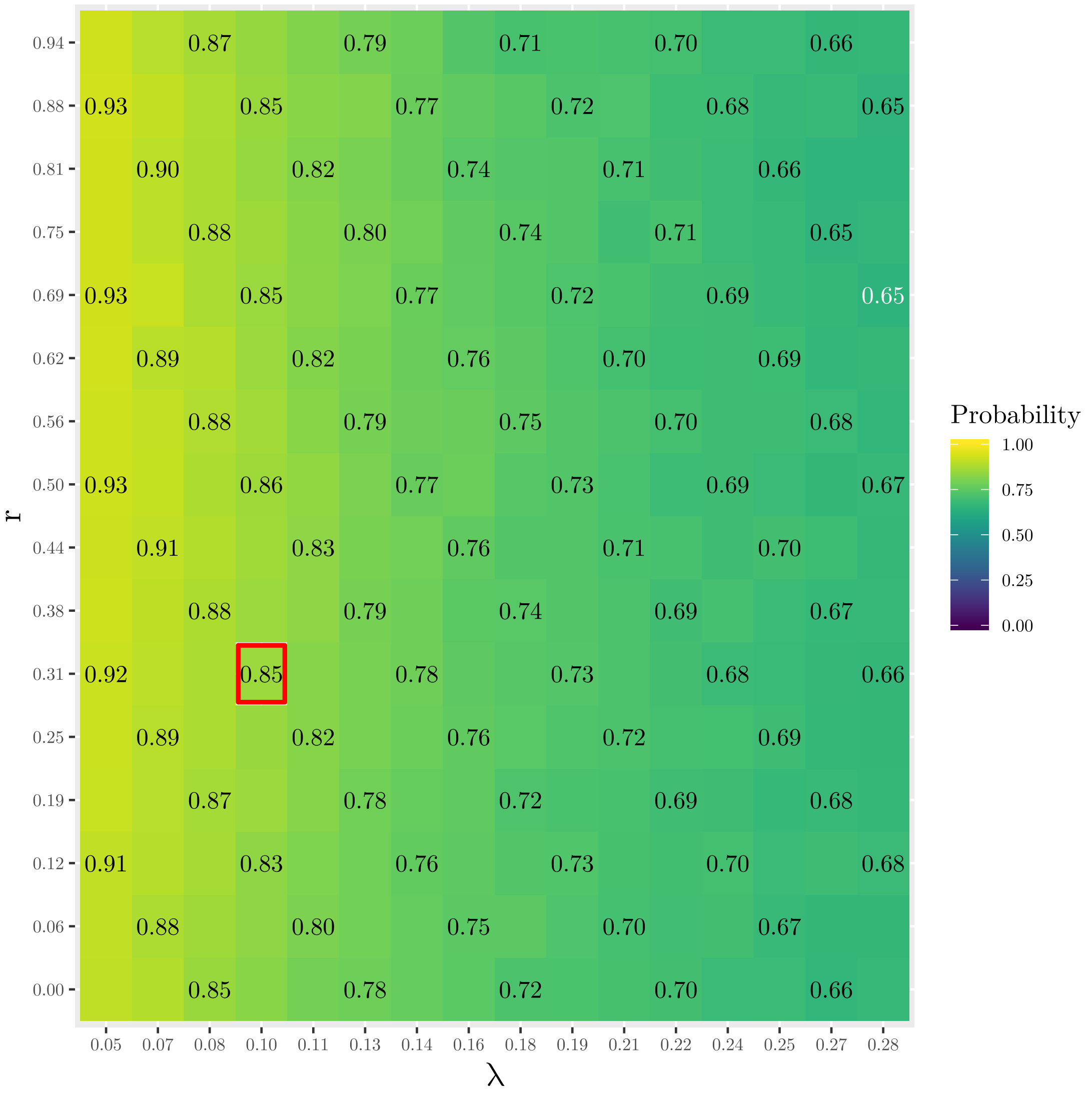


Figure 3: Hey and Orme (1994)

Percentage of RDU Winners, RDU DGP

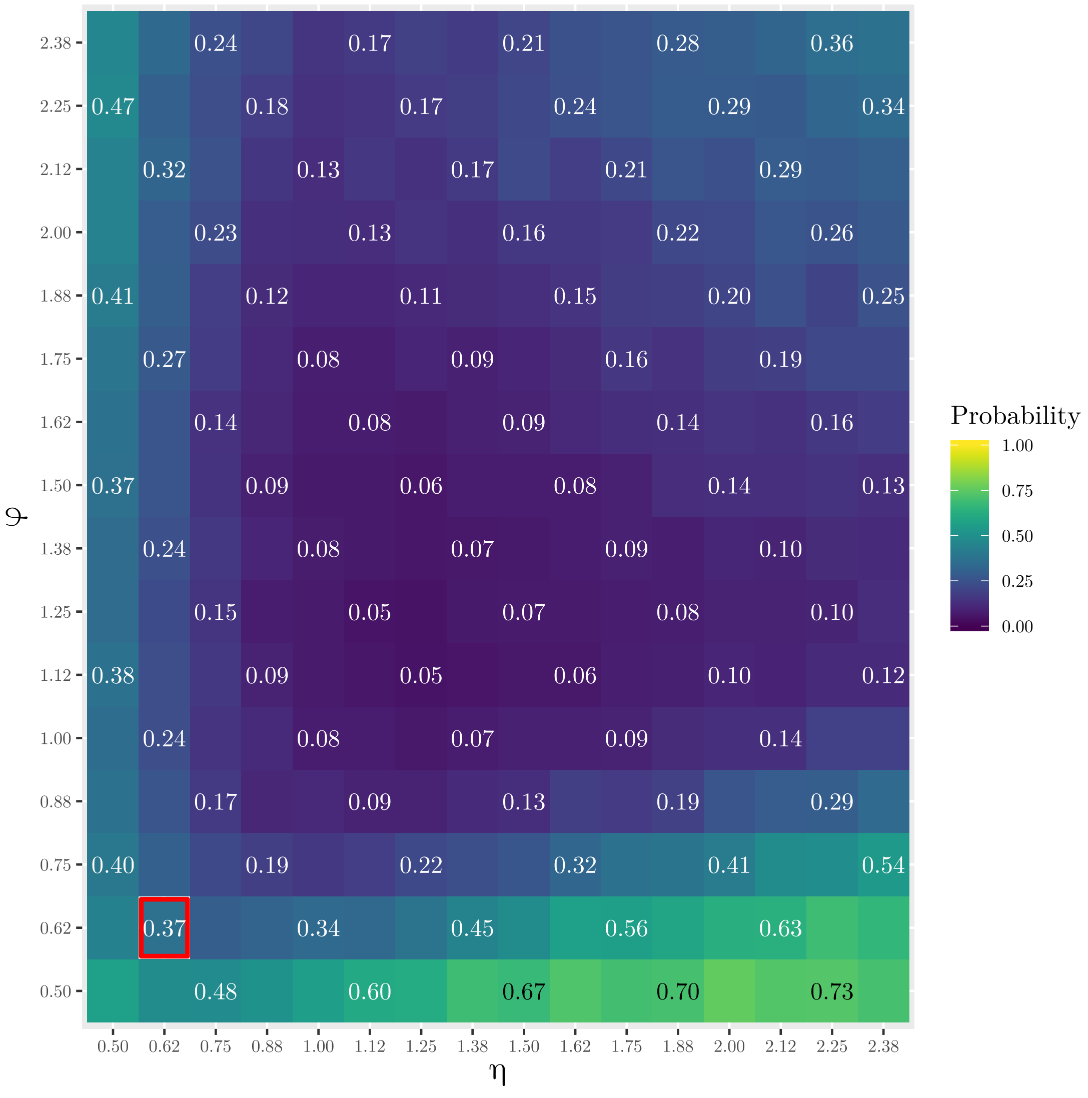
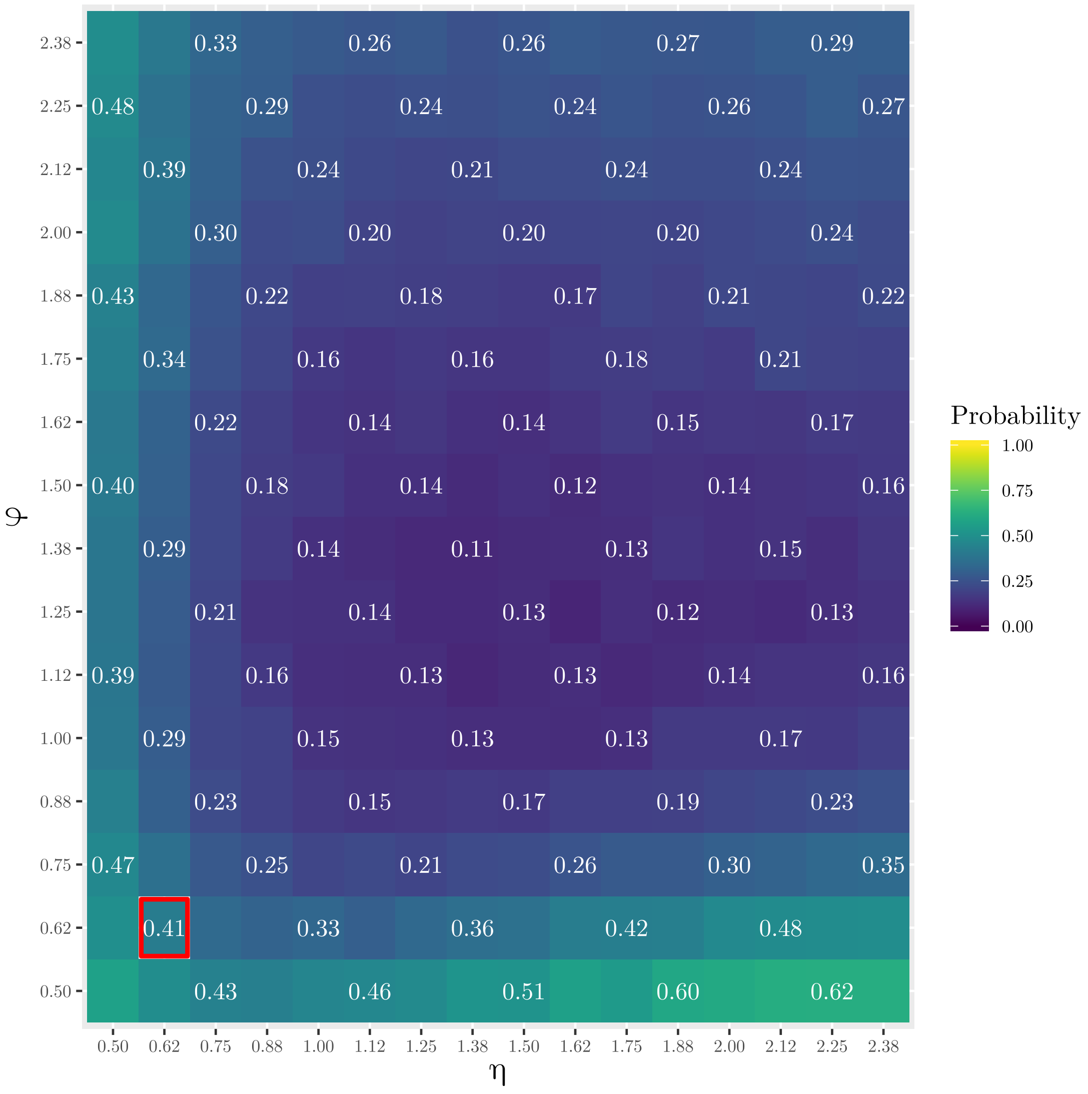


Figure 4: Harrison and Ng (2016)

Percentage of RDU Winners, RDU DGP



the probability of a type II error is 63.25% for the HO battery, and 58.61% for the HN battery. These rates of error are also far beyond the 20% typically referenced as a target limit for type II errors.

6 Power in a Hypothetical Sample

The previous analyses show how the probability that a subject with a given DGP will be correctly classified depends on the parameters they employ. Bayes' Theorem can be used to additionally infer the probability that a subject is correctly classified given their observed classification and knowledge of the subject's population. Bayes' Theorem stipulates:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{10}$$

Applying Bayes' Theorem to this problem, A indicates a subject actually employs a particular DGP, and B indicates the subject is classified as a particular DGP. Consider the case where A indicates a subject employs EUT and B that the subject is classified as RDU. $P(A)$ is therefore the probability that a subject actually employs the EUT, and $P(B)$ is the probability that a subject is classified as RDU. $P(B|A)$ gives the probability of a subject being classified as RDU given that they actually employ EUT. Finally $P(A|B)$ is the probability that a subject actually employs EUT given that they are classified as RDU. The conditional probability $P(A|B)$ is important because the classification of the subject is observed, while the actually employed DGP is unobserved.

$P(A)$ and $P(B)$ can be calculated by assuming a hypothetical sample. Consider a population made of 70% EUT and 30% RDU agents. $P(A_{\text{EUT}})$ would therefore be 0.7 and $P(A_{\text{RDU}})$ 0.3. $P(B)$ can be calculated using the law of total probability as $P(B) = P(B|A_{EUT})P(A_{EUT}) + P(B|A_{RDU})P(A_{RDU})$, where $P(B|A)$ is given by the analyses in the previous sections.

In the following, parameters grounded in the real data produced by the subjects in the HN laboratory experiments are used to calculate $P(B|A)$. The EUT and RDU models are estimated for each real, human subject in the HN data and used to classify the subject as either EUT or RDU. Then, all subjects classified as EUT (RDU) are pooled together and an unconditional, representative agent EUT (RDU) model is estimated over these pooled data. The estimated parameters from each pooled model are used to define a hypothetical joint distribution of parameters for the EUT and RDU DGPs. For the both DGPs, the r parameter is distributed as $r \sim N(0.5, 0.1)$ and the λ parameter as $\lambda \sim \mathit{Lognormal}(0.1, 0.02)$. For the RDU DGP, the ϕ and η parameters are additionally defined as $\phi \sim \mathit{Lognormal}(1.5, 0.1)$ and $\eta \sim \mathit{Lognormal}(0.7, 0.1)$.

Two hypothetical populations of EUT and RDU subjects are proposed, one that is 70% EUT and 30% RDU subjects, and one that is 70% RDU subjects and 30% EUT subjects. These values give $P(A)$ for each model. With the joint distributions of parameter sets defined for each of the EUT and RDU DGPs, each distribution is sampled 10,000 times, and the average $P(B)$, $P(B|A)$, and $P(A|B)$ are calculated for each population and each battery. Tables 2 and 3 show the results for the HO and HN batteries for the 70% EUT population. Tables 4 and 5 show the results for the HO and HN batteries for the 70% RDU population.

This Bayesian exercise presents a more complex picture than the simple power calculations. Take for instance, Tables 2 and 4, showing the HO battery with a 70% EUT population and a 70% RDU population, respectively. The probability that a subject employing RDU is correctly classified is the same across both populations,

$P(B|A) = 0.282$. However, since the classification of the subject, B, is observed and not the DGP they employ, A, these tables make clear that the proportion of subjects that employ the DGP in the general population, $P(A)$, critically informs whether the *observed classification* matches the DGP the subject *actually employs*. There is a 29% difference in the probability of the subject employing RDU given they have been classified as RDU, $P(A|B)$, solely due to the difference in the proportion of subjects that actually employ the RDU DGP in the two populations.

Table 2: Hey and Orme (1994), 100 choices, 70% EUT Sample

Model	$P(A)$	$P(B)$	$P(B\|A)$	$P(A\|B)$
EUT	0.7	0.861	0.922	0.750
RDU	0.3	0.139	0.282	0.608

Table 3: Harrison and Ng (2016), 80 choices, 70% EUT Sample

Model	$P(A)$	$P(B)$	$P(B\|A)$	$P(A\|B)$
EUT	0.7	0.812	0.863	0.744
RDU	0.3	0.188	0.307	0.490

Table 4: Hey and Orme (1994), 100 choices, 70% RDU Sample

Model	$P(A)$	$P(B)$	$P(B\|A)$	$P(A\|B)$
EUT	0.3	0.779	0.922	0.355
RDU	0.7	0.221	0.282	0.894

Table 5: Harrison and Ng (2016), 80 choices, 70% RDU Sample

Model	$P(A)$	$P(B)$	$P(B\|A)$	$P(A\|B)$
EUT	0.3	0.744	0.863	0.348
RDU	0.7	0.256	0.307	0.840

7 Conclusions

Accurate estimates of risk preferences are of critical importance when seeking to explain choice behavior of agents in a wide variety of economic environments. I present a conservative case to test the statistical power of two lottery batteries to distinguish between two possible DGPs. I conduct power analyses to estimate the probability of type I and type II errors when classifying subjects as either EUT or RDU and come to two general conclusions.

First, the probability of type I and type II errors are much greater than the 5% and 20% significance levels often cited in the social sciences as indicating statistically significant results. My analyses shows that the probability of a type I error for the HO battery is often above 10%, and often above 20% for the HN battery. The probability of a type II error can be as high as 95.21% for the HO battery and 89.89% for the HN battery.

Secondly, the conditional probability that a subject classified as EUT or RDU actually employs the EUT or RDU model critically depends on the percentage of subjects who actually employ each model in the population. This should be of no surprise to Bayesians, but analyses of individual level risk preferences are almost never conditioned on priors about the population. These analyses show that even when the probability of a type I error is low ($1 - P(A_{\text{EUT}}|B_{\text{EUT}}) < 10\%$), the conditional probability that a subject classified as EUT employs the EUT DGP can be less than 40% depending on the percentage of EUT subjects in the population.[8]

It is hard not to conclude that the effort of HO to determine if there was substantial evidence that subjects in experiments employed DGPs other than EUT was fraught with the statistical power issues that they suspected lingered behind their analyses. In comparing both the HO and HN batteries it is apparent that the lack of statistical power was not unique to HO, and may be a general problem when estimating risk preference models at the individual level.

What cannot be concluded from this analysis is the extent to which the lack of power in classifying the DGP of subjects leads to other inferential problems. Drawing inferences about subjective beliefs, consumer surplus, discount factors, and the applicability of the reduction of compound lotteries axiom all depend on the accuracy of risk preference estimates, and critically on whether the independence axiom of EUT is systematically violated by subjects. Additional analyses are needed to determine the rates of type I and type II errors in these extended inferential objectives, and the cost of these errors, that are due to the propagation of type I and type II errors in the model classification stage.

The paths available to improve the statistical power of batteries in identifying underlying DGP are somewhat unclear. A standard frequentist prescription might be to increase the sample size until sufficient statistical power is reached. While this may be appropriate if experimenters were concerned with inferences over a pooled sample, in which case only the number of subjects in an experiment would need to be increased, increasing the number of choices required by each subject to the degree

[8] Shown in Table 4.

needed is likely not feasible.[9] Increasing the statistical power of individual level classification will require the investigation of qualitative aspects of the batteries used and potentially the application of different econometric techniques.

[9] Analyses in Appendix A show that even increasing the size of the batteries to several hundred lottery pairs per subject is of limited value.

References

Andersen, Steffen, John Fountain, Glenn W. Harrison and E. Elisabet Rutström (2014). “Estimating Subjective Probabilities.” *Journal of Risk and Uncertainty* 48.3, pp. 207–229.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau and E. Elisabet Rutström (May 2008). “Eliciting Risk and Time Preferences.” *Econometrica* 76.3, pp. 583–618.

Bell, David E. (1982). “Regret in Decision Making under Uncertainty.” *Operations Research* 30.5, pp. 961–981.

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences.* Vol. 2. New York: Academic Press.

De Long, J. Bradford and Kevin Lang (1992). “Are all Economic Hypotheses False?” *Journal of Political Economy* 100.6, pp. 1257–1272.

Feiveson, Alan H. (2002). “Power by simulation.” *Stata Journal* 2.2, pp. 107–124.

Fisher, Ronald (1956). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver & Boyd, p. 175.

Gelman, Andrew and Eric Loken (2014). “The Statistical Crisis in Science.” *American Scientist* 102, pp. 460–465.

Harrison, Glenn W., Jimmy Martínez-Correa and J. Todd Swarthout (2015). “Reduction of compound lotteries with objective probabilities: Theory and evidence.” *Journal of Economic Behavior and Organization* 119, pp. 32–55.

Harrison, Glenn W. and Jia Min Ng (2016). “Evaluating the Expected Welfare Gain From Insurance.” *Journal of Risk and Insurance* 83.1, pp. 91–120.

Harrison, Glenn W. and E. Elisabet Rutström (2008). “Risk Aversion in the Laboratory.” *Research in Experimental Economics.* Ed. by James C Cox and Glenn W Harrison. Vol. 12. Bingley: Emerald Group Publishing Limited, pp. 41–196.

Hey, John D. and C. Orme (1994). “Investigating generalizations of expected utility theory using experimental data.” *Econometrica* 62.6, pp. 1291–1326.

Ioannidis, John P. A. (2005). “Why Most Published Research Findings Are False.” *Chance* 18.4, pp. 40–47.

Kahneman, Daniel and Amos Tversky (1979). “Prospect theory: An analysis of decision under risk.” *Econometrica* 47.2, pp. 263–292.

Loomes, Graham and Robert Sugden (1982). “Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty.” *Economic Journal* 92.368, pp. 805–824.

— (1998). “Testing different stochastic specifications of risky choice.” *Economica* 65, pp. 581–598.

McCloskey, Deirdre N. and Stephen T. Ziliak (1996). “The Standard Error of Regressions.” *Journal of Economic Literature* 34, pp. 97–114.

Prelec, Drazen (1998). “The Probability Weighting Function.” *Econometrica* 66.3, pp. 497–527.

Quiggin, John (1982). “A Theory of Anticipated Utility.” *Journal of Economic Behavior & Organization* 3, pp. 323–343.

Wilcox, Nathaniel T. (2011). “’Stochastically more risk averse:’ A contextual theory of stochastic discrete choice under risk.” *Journal of Econometrics* 162.1, pp. 89–104.

Zhang, Le and Andreas Ortmann (2013). “Exploring the Meaning of Significance in Experimental Economics.” *Working Paper.* Australian School of Business, University of New South Wales.

Appendix A: Scaled Batteries

The analyses in Section 5 show the differences in the rates of type I and type II errors for the EUT and RDU DGP across the HO and HN batteries. There are some mild differences in the error rates across the batteries for the EUT DGP and some more pronounced differences for the RDU DGP. However, the absolute rates of error across the two batteries are similar. Both have rates of type I errors in an acceptable range *if the noise in the data is sufficiently low*, and both batteries have exceedingly high rates of type II errors for the given parameter ranges.

Some of these differences in rates of error might be explained by the batteries having different numbers of choices per subjects. The simulation analysis described previously is repeated, but with each simulated subject responding to the HO battery 4 times and the HN battery 5 times. This results in each simulated subject making 400 choices for each battery. I refer to the scaled HO battery as HO_{400} and the scaled HN battery as HN_{400}. The results for the EUT DGP are presented in Figure A.1 for the HO_{400} battery, and Figure A.2 for the HN_{400} battery. The results for the RDU DGP are presented in Figure A.3 for the HO_{400} battery, and Figure A.4 for the HN_{400} battery.

As might be expected, the same patterns relating the parameters of interest and the probability a subject is correctly classified that was seen when subjects responded to the original HO and HN batteries are observed for the HO_{400} and HN_{400} batteries. For both the HO_{400} and HN_{400} batteries, the probability of a type I error increases monotonically with the λ parameter for the EUT DGP, and the probability of a type I error is lower for the range of r values in middle of the considered range, roughly $r \in (.3, .6)$. The probability of a type II error is greatest for the parameter values near $\phi = \eta = 1$, where the RDU model reduces to EUT.

The probability of type I and type II errors are lower for the HO_{400} and HN_{400} batteries than for the HO and HN batteries for the entire parameter ranges considered. This suggests, unsurprisingly, that the probability of type I and II errors monotonically decrease as the number of choices per subject increases. For the HO_{400} battery, the probability of a type I error is less than or equal to 14.31% for the entire range of parameters considered and the probability of a type I error for the HN_{400} battery is below 24.96% for the entire range of parameters. Both batteries show considerable increases in power over their original implementations. However, the probability of a type I error is still greater than 5% for both batteries across most of the parameter ranges considered, particularly when values of λ are large.

For parameter values such that $\phi, \eta > 1.38$ and $\phi, \eta < 0.75$, values far from the EUT special case, the probability of a type II error is always less than or equal to 68.62% for the HO_{400} battery and 67.9% for the HN_{400} battery. The median rate of a type II error in this range of ϕ and η values is 7.36% and 18.14% for the HO_{400} and HN_{400} batteries, respectively. Generally, the probability of correctly classifying RDU subjects is much greater than the original implementation of the HO and HN batteries.

It is intuitively sensible that the relationships between the values of the parameters and the probability that subjects are correctly classified should be similar between the HO and HN batteries and the HO_{400} and HN_{400} batteries. In scaling the batteries, the salient aspects of the lottery pairs remain unchanged for either battery, the only difference is the number of likelihood scores in the likelihood function. Thus, it is also sensible that there are large increases in the probability of correct classification (and therefore large decreases in the rates of type I and type II errors) for the HO_{400} and HN_{400} batteries given the additional likelihood scores.

However, these results present additional questions for experimenters concerned with statistical power. Firstly, while it is clear that one can improve the statistical power of these batteries by requiring subjects to respond to more questions, these results show that increasing the sample size is not a panacea. For both batteries, the RDU DGP with 400 choices per subject still produces type II errors with greater than 20% probability for large portions of the parameter ranges considered, even when the parameters are relatively far from $\phi = \eta = 1$. Subjects with an RDU DGP and parameters in the range of $1.37 < \phi < 1.5$ and $.6 < \eta < .75$, outlined in red in Figures A.3 and A.4, still have a 50.18% chance of producing a type II error with the HO_{400} battery and a 55.07% chance with the HN_{400} batteries.[10]

Secondly, 400 choices is generally beyond what many experimenters consider reasonable to ask subjects in a single experimental session, especially when accurate estimates of risk preferences are only a part of the inferential objectives of a study. Hey (2001) conducted an experiment in which subjects make choices over 500 lottery pairs, but do so over the course of 5 days. Background risk factors salient to subjects' choices over lottery pairs may change from day to day and experimenters may want subjects to make all of their choices in one session to help mitigate the effect of background risks.

The analysis of these batteries, both scaled so that they have the same number of lottery pairs, also raise additional questions about what *qualitative* aspects of the battery improve statistical power. Even when both batteries have the same

[10] The estimated parameters of subject 74 from HN falls in this range.

number of lottery pairs, it appears that the HO battery has increased statistical power in the parameter ranges considered over the HN battery. This suggests that qualitative differences in the construction of the lottery pairs used in the two batteries are what determine the differences in statistical power. As was stated previously, the HN battery was designed specifically with the intention of being able to distinguish between EUT and RDU subjects using the rationale of Loomes and Sugden (1998). It appears, however, that the HO battery performs better in this regard when comparing either the original battery, shown in Figure 3, or the scaled battery, shown in Figure A.3.

References

Harrison, Glenn W. and Jia Min Ng (2016). "Evaluating the Expected Welfare Gain From Insurance." *Journal of Risk and Insurance* 83.1, pp. 91–120.

Hey, John D. (2001). "Does Repetition Improve Consistency?" *Experimental Economics* 4, pp. 5–54.

Hey, John D. and C. Orme (1994). "Investigating generalizations of expected utility theory using experimental data." *Econometrica* 62.6, pp. 1291–1326.

Loomes, Graham and Robert Sugden (1998). "Testing different stochastic specifications of risky choice." *Economica* 65, pp. 581–598.

Figure A.1: Hey and Orme (1994), 4 Repetitions

Percentage of EUT Winners, EUT DGP

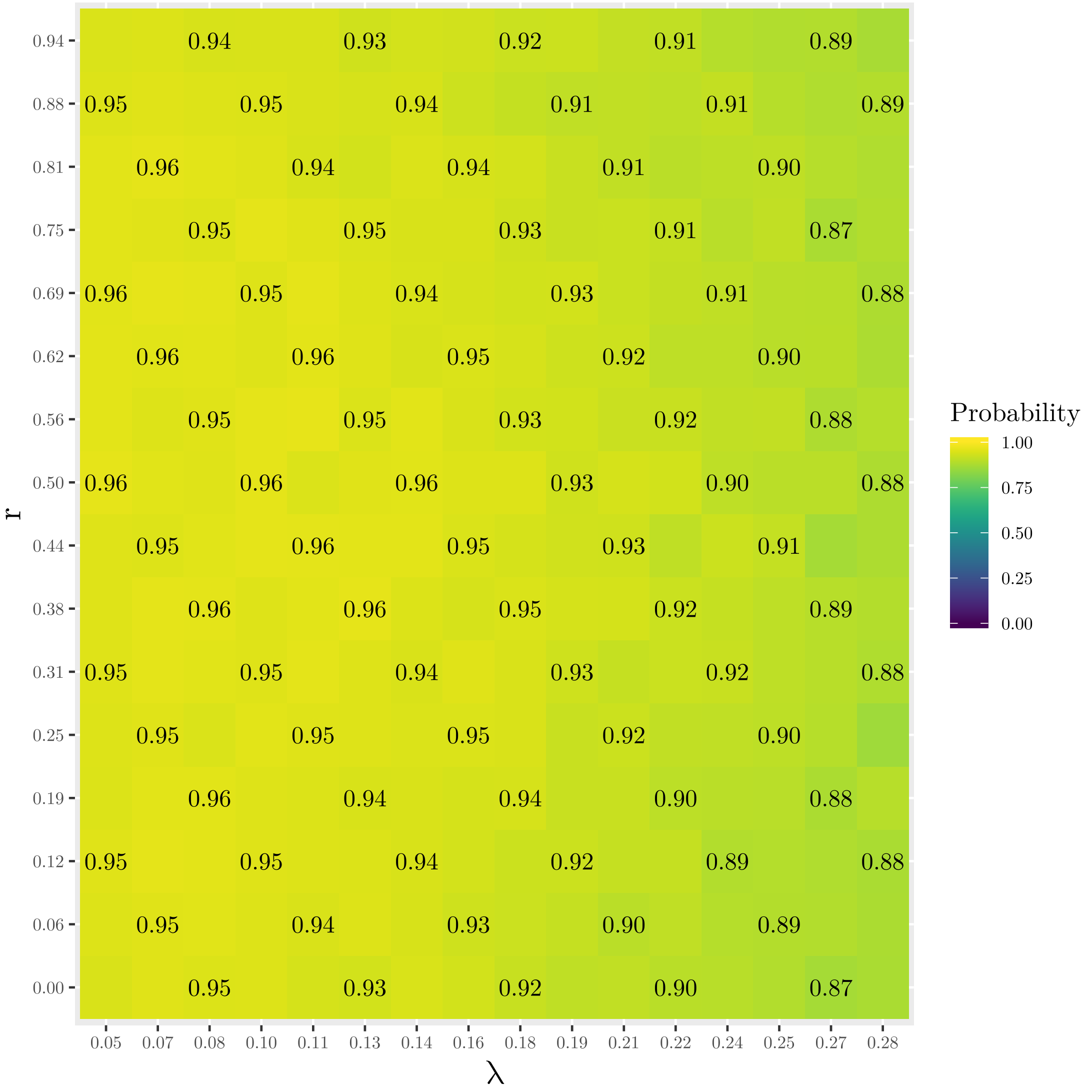


Figure A.2: Harrison and Ng (2016), 5 Repetitions

Percentage of EUT Winners, EUT DGP

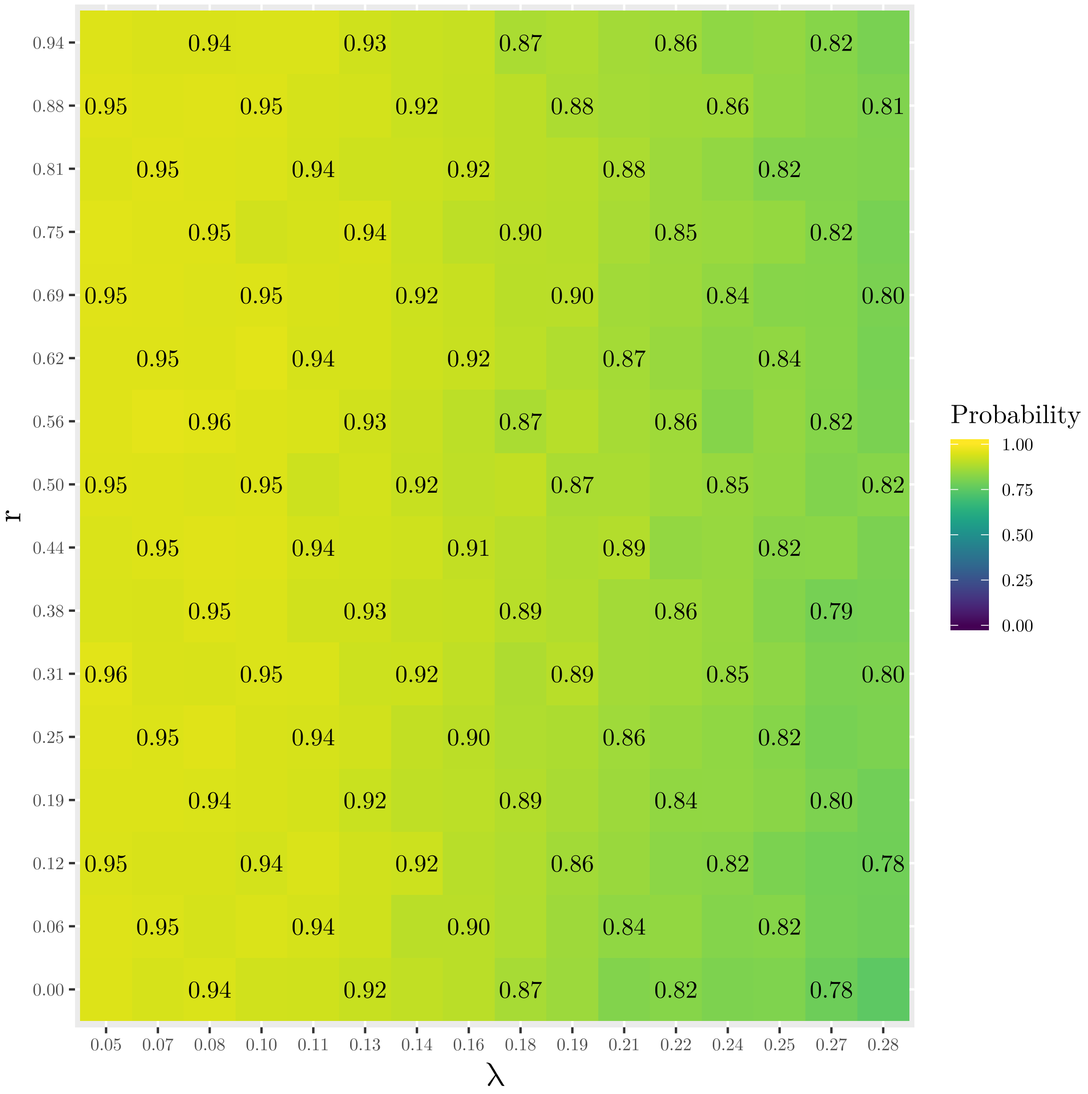


Figure A.3: Hey and Orme (1994), 4 Repetitions

Percentage of RDU Winners, RDU DGP

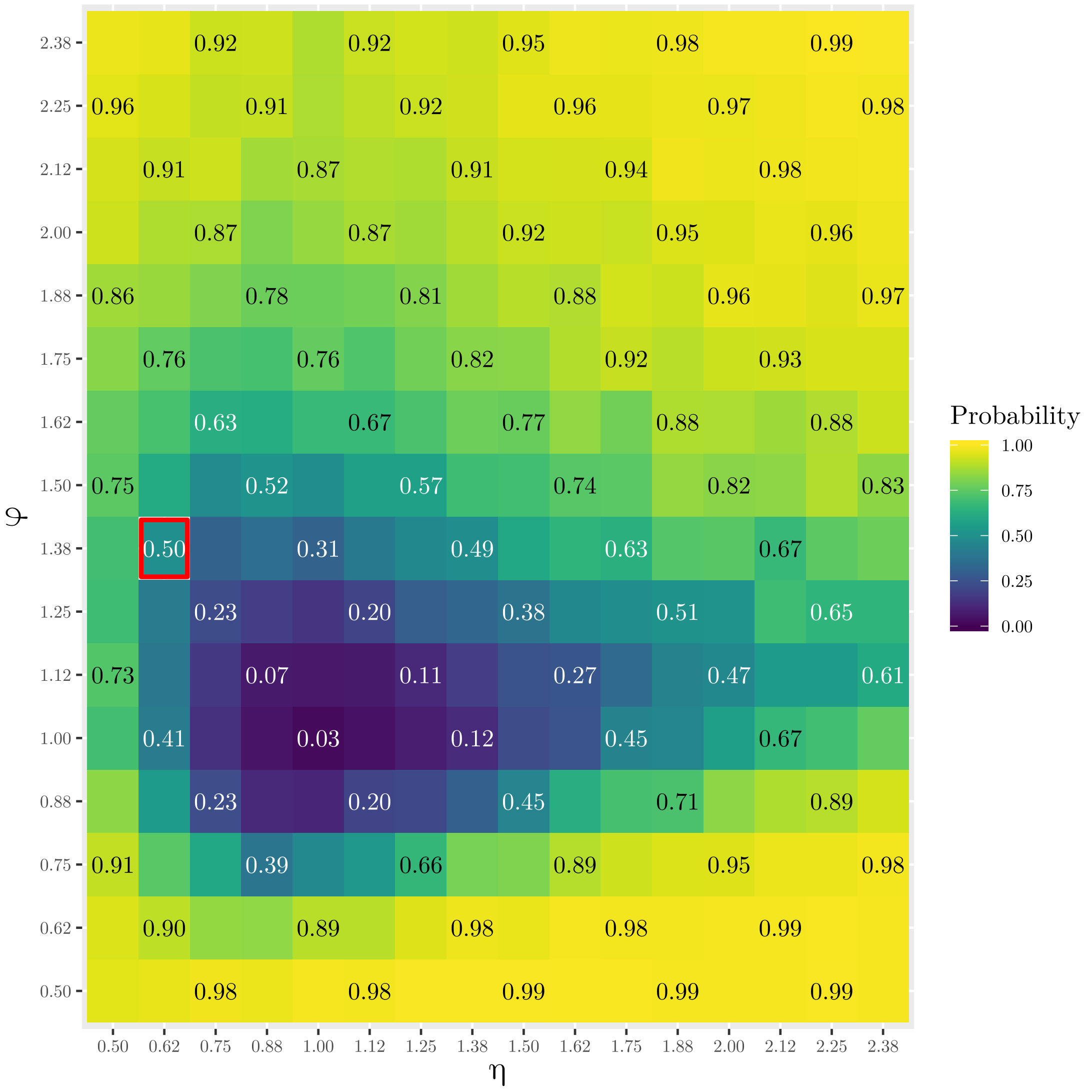
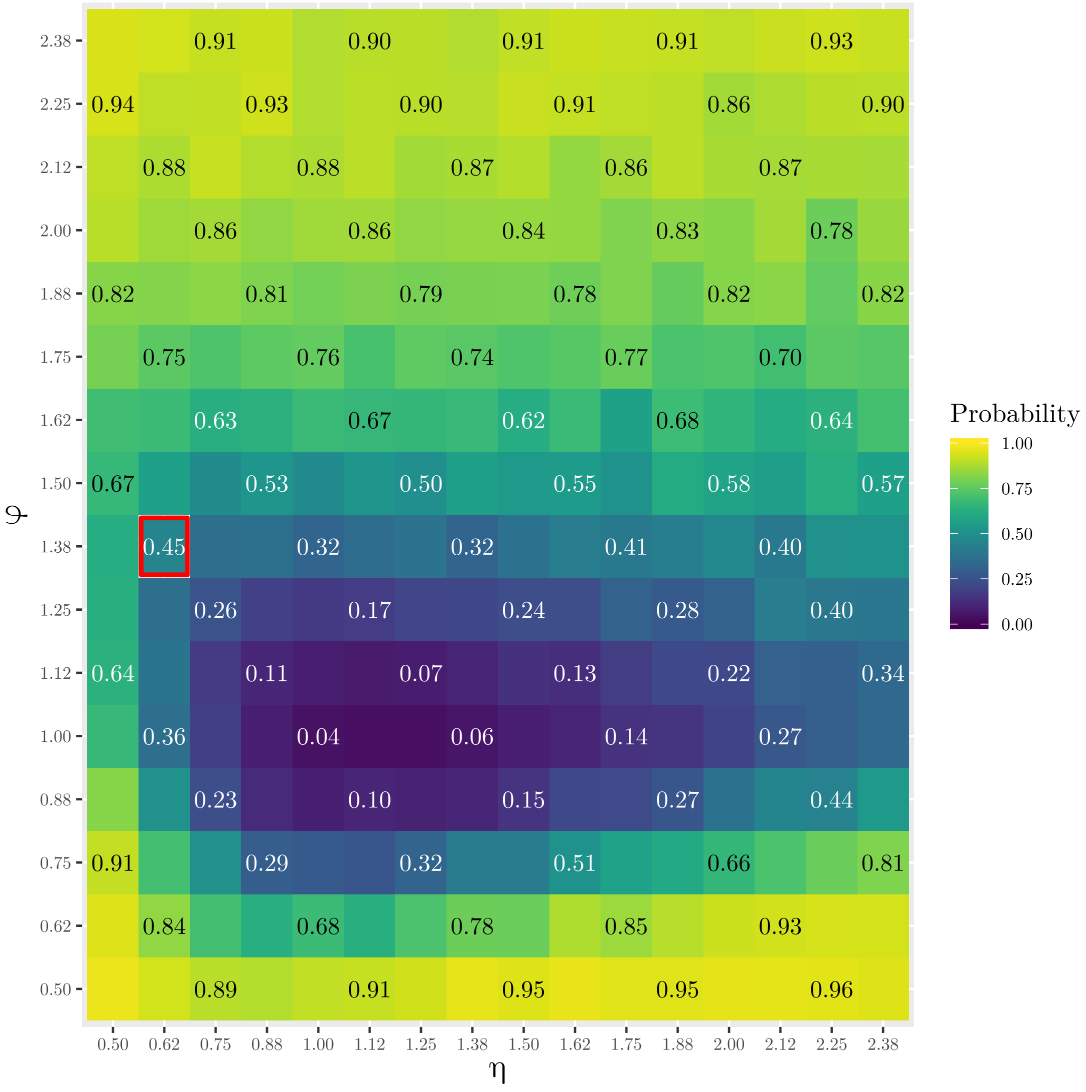


Figure A.4: Harrison and Ng (2016), 5 Repetitions

Percentage of RDU Winners, RDU DGP


φ
η
Probability