I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13580

Behavioral Welfare Economics and Risk Preferences: A Bayesian Approach

Xiaoxue Sherry Gao
Glenn W. Harrison
Rusty Tchernis

AUGUST 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13580

Behavioral Welfare Economics and Risk Preferences: A Bayesian Approach

Xiaoxue Sherry Gao
University of Massachusetts Amherst

Glenn W. Harrison
Georgia State University and University of Cape Town

Rusty Tchernis
Georgia State University, IZA and NBER

AUGUST 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Behavioral Welfare Economics and Risk Preferences: A Bayesian Approach*

We propose the use of Bayesian estimation of risk preferences of individuals for applications of behavioral welfare economics to evaluate observed choices that involve risk. Bayesian estimation provides more systematic control of the use of informative priors over inferences about risk preferences for each individual in a sample. We demonstrate that these methods make a difference to the rigorous normative evaluation of decisions in a case study of insurance purchases. We also show that hierarchical Bayesian methods can be used to infer welfare reliably and efficiently even with significantly reduced demands on the number of choices that each subject has to make. Finally, we illustrate the natural use of Bayesian methods in the adaptive evaluation of welfare.

JEL Classification: D6, C11, D81, G40

Keywords: behavioral welfare economics, Bayesian Analysis, risk preferences, insurance

Corresponding author:
Rusty Tchernis
Department of Economics
Andrew Young School of Policy Studies
Georgia State University
P.O. Box 3992
Atlanta, GA 30302-3992
USA

E-mail: rtchernis@gsu.edu

* We are grateful to Andre Hofmeyr for comments.

Table of Contents

Welfare evaluations of observed choices over risky lotteries depend on the assumed risk preferences that are used to make the evaluation. As a consequence, there are several burdens placed on the estimation of those risk preferences before one can reliably undertake normative evaluations of those choices. We propose a Bayesian approach to ease those burdens, and provide a rich case study of the evaluation of insurance purchase decisions.

The first burden arises from the recognition that risk preferences differ from individual to individual, so we ideally need to make inferences that entail collecting data at the individual level. In turn, that level of information on an individual can be time-consuming and expensive to collect, so we would like to have rigorous ways of pooling what individual responses we can collect in a cost-effective manner to generate informed priors about individual risk preferences. The second burden arises from the empirical observation that some, perhaps even many, individuals, are not well characterized statistically by available models of risk preferences using classical statistical methods. This can mean that we have estimates of their risk preferences but they are imprecise, that are *a priori* unlikely, or that estimation routines fail to produce estimates under the assumed model. This means we would like to have some disciplined way of "borrowing" information from other data points to better reflect the model when applied to each individual.

These considerations motivate a derived demand for conditioning inferences about individual risk preferences with priors from other sources, which is what Bayesian analysis allows one to do systematically and rigorously. We propose, and constructively illustrate, how to undertake a Bayesian analysis in this way for applications in behavioral welfare economics.[1] We focus initially

[1] In various forms Bayesian analysis has long been applied to condition inferences from experimental data. For example, see Harrison [1990] and the effect of priors over risk preferences on inferences about bidding behavior in first-price sealed bid auctions. Closer to our own implementation, Nilsson, Rieskamp and Wagenmakers [2011] and Murphy and ten Brincke [2018] employ hierarchical Bayesian methods to make inferences about risk preferences under Cumulative Prospect Theory, which is a structurally rich model and relatively hard to reliably estimate at the individual level.

on the canonical case in economics, evaluating the welfare consequences for an individual of some observed choices.[2] To illustrate the relevance for normative applications in a concrete manner, we re-examine the evaluations of decisions to purchase insurance from Harrison and Ng [2016].

A natural source of priors comes from estimates of models of risk preferences that pool data from a sample of subjects, using uninformative, diffuse priors over parameter values.[3] One can then estimate posterior distributions of these parameters, and use these predictions as informative, non-diffuse priors for Bayesian inferences for each individual. The posterior distributions that result for each individual are then a reflection of the overall prior and the sample generated by the individual subject. Bayesians call this "overall prior," that spans uninformative priors over the parameters characterizing the "representative agent" with informative priors over the parameters characterizing each individual agent, a *hierarchical* prior. A hierarchical prior describes a distribution for each individual, as well as the distribution of individuals in the population. When the data are relatively uninformative for a given individual, for one reason or another, the hierarchical prior will play a greater role in conditioning the posterior for that individual. The advantage of this approach is

---

[2] One might also be interested in measures of *social* welfare, derived from these individual welfare evaluations. Kitagawa and Tetenov [2018] consider a related issue, using a social welfare function defined directly over observable outcomes of individuals. They examine the determination of the sample of a population that should be treated by some intervention, when it is impossible to treat the full population with the available budget, and when one has baseline data with which to condition who to treat with what intervention. They explicitly recognize (p. 592) that when "multiple outcome variables enter into the individual utility (e.g., consumption and leisure), [the individual outcome measure] can be set to a known function of these outcomes." For us the challenge is to estimate this "known function" and account for the statistical properties of those estimates. The experimental task we use to estimate risk preferences is our counterpart of their baseline survey, albeit fully incentivized of course.

[3] An extension of this approach conditions inferences about each parameter on a list of observable demographic characteristics of the pooled sample. One can then generate predictions about the distributions of these parameters that condition on the specific value of the characteristics of each individual being normatively evaluated, and use these predictions as priors for Bayesian inferences that pool the sample data for that individual. We carefully evaluate this extension in Gao, Harrison and Tchernis [2020] and find that it adds no substantive insight for the sample from our population, although it does add considerable computational burden. This conclusion may be specific to our, relatively homogenous, population; we encourage examination of this extension for applications to field populations that are likely more heterogeneous.

that it will "always" generate informative priors for each individual. We focus on the role of this class of priors, since they are generally available.[4]

In Section 1 we review the data underlying these calculations, and the Bayesian framework for evaluating it. In Section 2 we discuss the normative evaluations of individual welfare based on that Bayesian framework, and contrast it with the Maximum Likelihood (ML) approach. Section 3 provides some extensions, showing how the Bayesian hierarchical approach allows dramatic savings in the experimental demands of subjects that is likely to be particularly attractive for field applications. We also show that a Bayesian approach lends itself naturally to "adaptive welfare evaluations" for individuals. Section 4 offers general conclusions.

1. Bayesian Estimation of Individual Risk Preferences

A. Data

We consider the data from Harrison and Ng [2016], where 111 subjects made 80 binary choices over risky lotteries with objective probabilities. For each individual we replicate the ML approach that they used, by estimating Rank Dependent Utility (RDU) models of risk preferences from the 80 choices that each individual made.[5]

---

[4] The use of Bayesian hierarchical models to infer individual preferences has a long tradition in marketing: see Rossi and Allenby [1993], McCulloch, Rossi and Allenby [1995], Allenby and Gintner [1995], Allenby and Rossi [1999] and Rossi, Allenby and McCulloch [2005]. Random coefficient (or mixed logit) models have been developed for similar applications: see Huber and Train [2001], Train [2009; chapter 11] and Reiger, Ryan, Phimister and Marra [2009] for expositions and comparisons with Bayesian hierarchical methods.

[5] However, we do not follow their approach of classifying certain individuals as having risk preferences consistent with Expected Utility Theory (EUT). The statistical reason, stressed by Monroe [2021], is that those subjects that are characterized as EUT by the test for "no probability weighting" still have standard errors around the probability weighting parameters, and potentially large ones. And, perhaps surprisingly, these standard errors can make a substantive difference in precisely the normative evaluations undertaken here. Hence there is no formal need to differentiate EUT and RDU decision makers for these calculations, because EUT is nested within RDU, even if there is an important normative insight in knowing that there are these different *types* of risk preferences in the sample.

In addition, and central to our normative application, Harrison and Ng [2016] also asked each individual to make 24 binary choices over an insurance product. The background risk that this product was defined over is, formally, a simple lottery. In the absence of having purchased insurance, the individual faced some known probability of a loss from some known endowment. The insurance product was a full indemnity, zero-deductible product with no co-pay and no coinsurance. Across the 24 choices there were two loss amounts, and various premia, presented in random order; the endowment and loss probability were held constant. Of course, to economists this is just a choice between the "safe lottery" of buying insurance and the "risky lottery" of not buying insurance. Hence the domain of the task is identical to the prior choices over 80 risky lotteries, apart from the framing of the task as the purchase of insurance. We return to this point in the conclusions: Bayesian analysis lends itself naturally to considering the use of risk preferences elicited in one domain to evaluating "target choices" from another domain, which will be needed for broader applications of this normative approach to welfare evaluation.

Harrison and Ng [2016] take the estimates from the risk preferences of each individual subject from the initial 80 choices, and use them to infer the Certainty Equivalent (CE) of each of the 24 binary choice options. The difference in the CE of buying or not buying insurance defines the expected Consumer Surplus (CS) of purchasing insurance, and hence provides a rigorous measure of individual welfare of the observed choice. From a policy perspective, the insight from behavioral welfare economics is that an individual may be observed to make an insurance choice that involves a negative CS.[6] In addition, this approach provides a quantification of the CS, whether gained or lost, from the observed choices.

[6] The methodological basis of this insight is discussed by Harrison and Ng [2016; p.111-116], Harrison and Ross [2018; p. 59-63] and Harrison [2019].

B. Models of Risk Preferences

In the evaluation of lottery prizes, assume individuals perfectly integrate the prizes with their endowments and behave as if they evaluate Constant Relative Risk Aversion (CRRA) utility functionals $u(e, x_k) = (e + x_k)^{(1-r)}/(1-r)$ for any k = 1, ... , K, and where x_k refers to prize k, e is some endowment, and r is the utility curvature parameter. To ease notation, and unless the context needs it, we dispense with subscripts for core risk preference parameters.

Under **Expected Utility Theory** (EUT) a lottery is evaluated by the weighted sum of utilities of prizes, with the weights being the objective probabilities associated with the prizes. Then, we have

$$EU = \sum_{k=1,K} [\, p_k \times (e + x_k)^{(1-r)}/(1-r) \,]. \tag{1}$$

In our battery K=4. Define the latent index for choice t by subject i as the difference between the EU of the left and right lottery subject to a Fechner noise parameter μ_i and a random noise term ε_{it}:

$$y_{it}^{*} = \nabla EU_{it}(r_i, \mu_i) + \varepsilon_{it} = \{ [(EU_{it}^{L}(r_i) - EU_{it}^{R}(r_i)) / \nu_{it}] / \mu_i \} + \varepsilon_{it} , \tag{2}$$

where ν_{it} is the "contextual utility" term specific to choice t to normalize utilities of prizes between 0 and 1, and r_i and μ_i are the parameters for subject i we want to estimate. Assume that subject i selects the left lottery in lottery pair t whenever the latent index y_{it}^{*} is greater or equal to 0:

$$\text{Prob}\ (y_{it} = 1) = \Lambda(y_{it}^{*}) , \tag{3}$$

where $\Lambda(.)$ is the logistic function.

Under **Rank-Dependent Utility** (RDU) theory, due to Quiggin [1982], a lottery is evaluated by the weighted sum of utilities of prizes, where the weights are the associated *decision weights*. RDU departs from EUT in the manner in which decision weights depend on objective probabilities; under EUT the decision weight for each prize *is* the corresponding objective probability, as in (1). Under RDU we first rank the prizes from best to worst, such that $x_1 \geq x_2 \ldots \geq x_K$. The decision weight associated with each prize is calculated as follows:

$$\pi(x_1) = \omega(p_1) \quad (4a)$$
$$\pi(x_2) = \omega(p_{1+}p_2) - \omega(p_1) \quad (4b)$$
$$\dots$$
$$\pi(x_K) = \omega(1) - \omega(p_1 + \dots + p_{K-1}) \quad (4c)$$

where $\omega(.)$ is the probability weighting function (PWF): a strictly increasing and continuous function with $\omega(0) = 0$ and $\omega(1) = 1$. The flexible PWF that we use is due to Prelec [1998]:

$$\omega(p) = \exp(-\eta (-\ln p)^{\varphi}) \quad (5)$$

with $\eta > 0$ and $\varphi > 0$. EUT is nested within RDU when $\eta = \varphi = 1$. The RDU of a lottery is then calculated as

$$\text{RDU} = \sum_{k=1,K} [\pi_k \times (e + x_k)^{(1-r)}/(1-r)], \quad (6)$$

which is the same as the definition of the EU of a lottery in (1) apart from p_k being replaced by π_k. Define the latent index as the difference between the RDU of the left and right lottery subject to a Fechner noise parameter μ_i and a random noise term ε_{it}. We therefore have

$$y_{it}^{*} = \nabla \text{RDU}_{it}(r, \eta, \varphi) + \varepsilon_{it} = \{ [(\text{RDU}_{it}^{L}(r_i, \eta_i, \varphi_i) - \text{RDU}_{it}^{R}(r_i, \eta_i, \varphi_i)) / \nu_{it}] / \mu_i \} + \varepsilon_{it} , \quad (7)$$

where ν_{it} is again the term to normalize utilities of prizes between 0 and 1 in choice t by subject i, and r_i, η_i, φ_i and μ_i are the parameters we want to estimate. The subject is again assumed to select the left lottery in a pair whenever the latent index y_{it}^{*} is greater or equal to 0, as specified in (3).

C. Bayesian Analysis

We specify a Hierarchical Bayesian model in formal terms, and then explain how it is interpreted in terms of historically popular terminology about "shrinkage priors."[7]

The data-generating process revolves around core parameters r_i, η_i, φ_i and μ_i. We posit *two* hyper-parameters that describe the distribution that characterizes *each* of

[7] Gao, Harrison and Tchernis [2020] provide full details of implementation for simpler and more complex models, extensive simulation evidence of the reliability of estimators to reliably recover risk preferences, and software in *Stata*.

- r_i, the curvature of the utility function of individual i;
- η_i, one of the parameters of the probability weighting function of individual i;
- φ_i, the other parameter of the probability weighting function of individual i; and
- μ_i, the Fechner noise parameter of individual i.

Hence we estimate 8 hyper-parameters in all, based on the pooled data across all N subjects. In addition, we estimate r_i, η_i, φ_i and μ_i for each individual i = 1, ..., N. In all, therefore, we jointly estimate 8 + (4 × N) parameters for the full hierarchical model. Since N = 111 in our data, we jointly estimate 452 parameters.

Although we specify the prior distribution separately for each parameter, the posterior distribution of each parameter is correlated with other parameters, both within a subject and across subjects. In essence, the RDU model decomposes the risk premium presumed to drive the observed choices by subject i into two components: utility curvature governed by parameter r_i, and probability weighting governed by parameters η_i and φ_i.[8] There is a well-understood tradeoff between the two components explaining the risk premium, which introduces the correlation between the three parameters in the sampling of their joint posterior distribution.

Turning to the specific prior distributions assumed, it is important with hierarchical Bayesian models to be explicit and verbose so that the full specification is clear. Specifically, we assume that r_i is characterized by a Normal *prior*:

$$r_i \sim N(m_r, \sigma^2_r), \tag{8}$$

where there is a diffuse Normal *hyper-prior* for m_r given by

$$m_r \sim N(0, 100), \tag{9}$$

and there is a diffuse Inverse Gamma *hyper-prior* for σ^2_r given by

$$\sigma^2_r \sim IG(\sigma_r, 0.001, 0.001). \tag{10}$$

[8] In the extreme case of EUT the risk premium is solely determined by utility curvature. In the extreme case of "dual theory" the risk premium is solely determined by the probability weighting function (Yaari [1987]).

The essential idea is that there is an informative, non-diffuse prior specified in (8), where the values for m_r and σ^2_r come from the posterior distributions generated by the data for all subjects *and* the diffuse priors in (9) and (10). We can restate (8) in conditional form as

$$r_i \mid m_r, \sigma^2_r \sim N(m_r, \sigma^2_r), \tag{8′}$$

to remind us that if we knew the mean and the variance of the prior we would have much more information about the individual r_i values.

Although it is important that these estimations are undertaken jointly, (8′) reminds us that it is as if one Bayesian model was estimated for the pooled data just assuming the diffuse priors (9) and (10), and *then* the "point estimates" (averages) from the resulting posterior distributions for m_r and σ^2_r were used as the informative priors for each r_i, which are *then* estimated one individual at a time. The joint distribution is the product of conditional distributions and marginal distributions. In this manner a hierarchical prior achieves two goals. First it restricts parameters of individual distributions to a specific family. Second, it communicates that *a priori* those distributions are diffuse. As we will see, the resulting posterior distributions will be combining information from the prior and the likelihood. Thus, we will be informing the posterior for a specific individual using information from other individuals.

The remaining prior distributions are similar, and can be interpreted similarly. The only difference is that we want to ensure that the core parameters η_i, φ_i and μ_i are each non-negative, for obvious theoretical reasons. Therefore we use log-normal priors for each, and conventional hyper-priors. Assume that η_i is characterized by a log-normal *prior*

$$\ln(\eta_i) \sim N(m_{\ln\eta}, \sigma^2_{\ln\eta}) \tag{11}$$

where there is a diffuse Normal *hyper-prior* for $m_{\ln\eta}$ given by

$$m_{\ln\eta} \sim N(0, 100), \tag{12}$$

and there is a diffuse Inverse Gamma *hyper-prior* for $\sigma^2_{\ln\eta}$ given by

$$\sigma^2_{\ln\eta} \sim \text{IG}(\sigma_{\ln\eta}, 0.001, 0.001). \tag{13}$$

Assume that φ_i is characterized by a log-normal *prior*

$$\ln(\varphi_i) \sim \text{N}(m_{\ln\varphi}, \sigma^2_{\ln\varphi}) \tag{14}$$

where there is a diffuse Normal *hyper-prior* for $m_{\ln\varphi}$ given by

$$m_{\ln\varphi} \sim \text{N}(0, 100), \tag{15}$$

and there is a diffuse Inverse Gamma *hyper-prior* for $\sigma^2_{\ln\varphi}$ given by

$$\sigma^2_{\ln\varphi\mu} \sim \text{IG}(\sigma_{\ln\varphi}, 0.001, 0.001). \tag{16}$$

Finally, assume that μ_i is characterized by a log-normal *prior*

$$\ln(\mu_i) \sim \text{N}(m_{\ln\mu}, \sigma^2_{\ln\mu}) \tag{17}$$

where there is a diffuse Normal *hyper-prior* for $m_{\ln\mu}$ given by

$$m_{\ln\mu} \sim \text{N}(0, 100), \tag{18}$$

and there is a diffuse Inverse Gamma *hyper-prior* for $\sigma^2_{\ln\mu}$ given by

$$\sigma^2_{\ln\mu} \sim \text{IG}(\sigma_{\ln\mu}, 0.001, 0.001). \tag{19}$$

In effect, all that these priors are saying is that we let the pooled sample data determine the posterior distribution for the representative agent,[9] and then use that distribution as the prior for the sample data for each and every individual subject. The key implication of these priors being presented jointly, and then the joint estimation of the posterior over the risk preferences of the representative agent *and* N individual agents, is that the estimation of the posterior for the representative agent respects the fact that each individual agent can have different risk preferences.

[9] The "representative agent" just refers to a model of the complete sample of individuals that assumes that each individual has the same risk preferences. A variant allows for conditioning on observable demographics of the sample, such as gender. Our model of the representative agent has no such conditioning.

D. Historical Connections

The prior we employ to infer individual risk preferences is known historically as a "shrinkage" prior, since it uses pooled data for the sample of N individuals that includes the individual to generate a prior for the individual. The term "shrinkage" refers to the idea that the posterior distribution for each individual is pulled towards the posterior distribution for the pooled sample of N, hence the effect is to reduce (i.e., shrink) the cross-individual variability in posterior distributions. This is also sometimes referred to as an "empirical Bayes" approach, to reflect the fact that the data for a sample of N individuals is being used to form a prior for the individual in question.[10]

Modern Bayesians refer to these instead as hierarchical Bayesian models, where the information provided by the rest of the sample is used to condition the prior for the individual in question. Detailed reviews can be found in Gelman et al. [2013; ch. 5], Kruschke [2013; ch. 9], Kruschke and Liddell [2018; p. 197ff.], Kruschke and Vanpaemel [2015], Leamer [1978; ch. 5], Rossi, Allenby and McCulloch [2005] and Train [2009; chapter 12].

2. Normative Application

A. Estimates of Risk Preferences

We replicate the ML estimates obtained by Harrison and Ng [2016]. The first observation is that of the 111 subjects we want to make welfare evaluations for, 9 simply drop out because it was not possible to generate ML estimates for their risk preferences. This is true for all of the models they considered, and not just the most demanding in terms of numbers of free parameters to be estimated. As happens when estimating risk preferences at the individual level, even with 80 binary

[10] There are "jacknife" variants that use the N-1 individuals in the sample other than the individual in question, but for large enough samples this is not likely to make an appreciable difference quantitatively.

choices chosen carefully to allow estimates of models of risk preferences such as these, standard numerical methods can simply fail to converge.[11] An immediate corollary is that one is left without any normative judgement for these 9 individuals. Our Bayesian approach generates posterior estimates for those 9 individuals.

For simplicity we focus attention solely on the most general model of risk preferences considered by Harrison and Ng [2016], the RDU model with Prelec probability weighting. The second observation to make is that there are no ML estimates for *this* model for 22 of the 102 individuals for whom *one* of the models of risk preferences did converge. Given the generality of the RDU model with Prelec probability weighting, this is a caution that one or more of the parametric restrictions for less general RDU models[12] was needed to even obtain ML estimations. Relying on parametric restrictions that have no *a priori* support to even obtain estimates is problematic, from a Bayesian and classical perspective. Again, for all of these 22 subjects we were also able to obtain Bayesian posterior estimates using the most general RDU model.

For those less familiar with Bayesian methods, it is useful to explain how we do this, as if by magic, for the 31 = 9 + 22 subjects abandoned by ML. The reason is simple: the ML approach rests on numerical methods finding a set of estimates that characterizes a maximum log-likelihood for the observed binary choices. If the likelihood function has some "flatness" around the maxima, standard methods, particularly derivative-based methods, can fail to converge. Critically, there is no difficulty evaluating the log-likelihood for a wide range of possible estimates, just a difficulty finding the one best set of estimates. A Bayesian is not bothered by this latter difficulty at all, and just needs the

[11] In comparable calculations Harrison and Ross [2018; p. 54] report having to drop 19 of 193 subjects for effectively the same reason.

[12] Specifically, using Power or Inverse-S probability weighting functions. Each is effectively nested in the Prelec probability weighting function. When $\varphi = 1$ the Prelec function collapses to the Power function, and when $\eta = \varphi = 0$ or $\eta = 1$ it collapses to the Inverse-S function.

likelihood function evaluations in order to derive the posterior distribution. Of course, if the likelihood function is globally flat, the posterior will just be a replica of the prior, and the data from the subject non-informative, but that is a separate matter: there will still be a posterior, albeit derived solely from the prior. In general we "never" observe such globally uninformative data, but we do observe data that are locally uninformative, as evidenced by the 9 individuals callously tossed overboard by Harrison and Ng [2016] for the purposes of welfare evaluation. Moreover, the posterior estimates for these 9 subjects are not just replicas of the posterior distribution of the representative agent: their likelihoods can be evaluated and averaged, even if they are hard to numerically optimize with derivative-based algorithms.

The Bayesian hierarchical model generates estimates of the pooled behavior over all 111 subjects, which we might think of as the risk preferences of a representative agent. Of course this is just a stepping stone to the estimates from the same model for each of the 111 individuals, but it is a valuable one to help understand where the informative prior comes from for the individual posterior distributions.

Figure 1 compares "point estimates" for the risk preference estimates of the representative agent using ML methods (the top two panels) and then using Bayesian methods (the bottom two panels). For the Bayesian model these point estimates refer to means of the posterior distributions for the representative agent, since "point estimate" makes no formal sense to a Bayesian. Consistent with the use of a diffuse prior for the representative agent, we observe virtually no difference between the ML estimates and the Bayesian posterior estimates in Figure 1.

But the modest step summarized in Figure 1 is just the beginning for the Bayesian hierarchical model, whose primary inferential objective is to estimate individual risk preferences in the form of posterior distributions that are reduced to "point estimates" in Figure 1. These distributions across individuals are illustrated in Figure 2. Here we again reduce a posterior

distribution to a "point estimate," but in this instance it is a full posterior distribution for each and every individual. This posterior distribution for each individual is estimated by the informative prior obtained from the posterior distribution for the sample as a whole as well as the observed data for each individual. The posterior distribution for each individual combines the information from that individual and the information from other individuals, which is communicated through the prior. This prior is referred to by Gelman et al. [2013; p.559] as "a common backbone from which a hierarchical model *for borrowing information* can be built" (our emphasis).

The dashed lines in Figure 2 are the average Bayes estimates displayed in Figure 1. Now, in Figure 2, we start to see the distribution of individual risk preferences that we need for behavioral welfare evaluation.

We can directly compare the ML estimates for the remaining 80 subjects with our Bayesian estimates for all 111 subjects. For the moment just focus on the estimate of the CRRA parameter for the utility function, since that is the critical parameter for the evaluation of CE and CS for the insurance choice options. We find 6 subjects for whom the ML estimate implies convex utility, but the Bayesian estimate implies concave utility. And we find 3 subjects for whom the ML estimate implies concave utility, but the Bayesian estimate implies convex utility. Set aside whether these are statistically significant or credible differences, to use the classical or Bayesian counterparts for such inferences. This qualitative difference in the point estimates has dramatic implications for the individual welfare evaluation for these subjects. As a sample, it may end up being a wash, but that is not generally, or reliably, the point.

Three examples demonstrate the contrasts between ML and Bayesian estimates. Figure 3 illustrates an individual whose ML estimates show sharply convex utility with extreme probability pessimism, and whose Bayesian estimates show mildly concave utility with modest probability pessimism. For given RDU evaluations of the safe "buy insurance" lottery and the risky "do not buy

insurance" lottery these utility functions generate very different CE. These estimates also show the difference between selecting the single *maximum* LL estimates and *averaging* a weighted array of LL estimates. In the ML case the utility function, *ceteris paribus*, generates risk loving behavior; and the probability weighting function, *ceteris paribus*, generates risk averse behavior. These two, strong, opposing gross effects lead to a modest risk premium. In the Bayesian case, the estimates exhibit virtually minimal concavity in the utility function, and modest probability pessimism, jointly resulting in the same, modest risk premium. Figure 3 is an example of a wider class of subjects, where the Bayesian estimates lead to less extreme specifications of utility curvature and probability weighting.

Figure 4 shows a case in which the ML and Bayesian estimates more or less agree on the concavity of the utility function, but show different degrees of probability weighting patterns. The qualitative nature of probability weighting is the same with ML and Bayesian estimates, but clearly the ML estimates are more extreme. In both cases there is local pessimism for low (decumulative) probabilities and local optimism for high (decumulative) probabilities. Since the insurance contract is full indemnity there is no risk if the contract is purchased: the individual received the endowment less the premium no matter what the state of nature. And if the insurance contract is not purchased, the endowment outcome receives the higher decumulative weight (hence optimism) and the outcome in which the endowment is reduced by the loss amount receives the lower decumulative weight (hence pessimism). Thus the probability weighting alone means that the no-insurance RDU being evaluated in the ML case is not perceived as risky at all, compared to when it is evaluated in the Bayesian case (and is then only modestly risky). And since the utility functions are virtually the same, we would see a lower CE to purchasing insurance with the Bayesian estimates, and hence a lower CS.

Finally, Figure 5 displays a case that is modal and typical. The ML point estimates change slightly in quantitative terms, and do not change in qualitative terms. Modestly concave utility with

the ML estimates become more concave with the Bayesian estimations. And the roughly "power" probability weighting with ML, that indicates significant probability pessimism, become modestly pessimistic with Bayesian estimation methods. Although specific to this instance, Figure 5 also illustrates the nature of the RDU trade-off between utility curvature and probability weighting nicely. With the ML estimates much more of the risk premium is due to probability weighting than we find with the Bayesian estimates, but both types of estimates end up at the same risk premium due to offsetting adjustments to utility curvature.

As a general matter, we find that most of the Bayesian posterior estimates for individuals are close to their ML counterpart. Figure 6 displays this, by showing scatter plots of the ML and Bayesian estimates, along with 45° lines. A large number of observations are clustered around modest deviations of the 45° line. The serious deviations are all from the perspective of extreme ML estimates: very low estimates of r, and very high estimates of η or φ.

B. Welfare Effects

The top panel of Figure 7 displays the implied calculations of CS gains or losses from each of the 24 decisions that each individual subject make, evaluated with the ML or Bayesian estimates for that subject.[13] As explained above, we have 80 subjects with ML estimates, and 111 subjects with Bayesian estimates.[14] The distribution indicates a difference between the two sets of estimates: less extremes with the Bayesian estimates, a clear tendency for more CS gains up to +$4, and a clear

[13] Harrison and Ng [2016; p. 110/111] show how one can bootstrap the CS calculations to reflect the covariance matrix of ML estimates for each individual. And similar exercises can, and should, be undertaken with the Bayesian posterior distributions for each individual. In the interests of exposition we focus here solely on the effects of using different point estimates. We consider the calculation of posterior predictive *distributions* of welfare in §3.B.

[14] Virtually identical distributions are generated if we restrict to the 80 individuals with both ML and Bayesian estimates, but one point of the exercise is not to do that.

tendency for more small CS losses up to -$1.

Because the some of the 24 product offering are better than others, we often consider the percentage of the total CS that the individual *realizes* over all observed decisions compared to the total CS that the same individual *would have realized* over all decisions if all decisions were correct. This is called Efficiency by experimental economists, and effectively normalizes across subjects for the different product offerings, since each individual faces the same set of 24 product offerings by design.

The bottom panel of Figure 7 displays the implied calculations of Efficiency for each individual, across all 24 decisions, evaluated with the ML or Bayesian estimates for that subject. The distribution of Efficiency with the Bayesian estimates of risk preferences is clearly higher than with the ML estimates of risk preferences. The Efficiency results complement the CS results, by informing us of the agent-specific welfare effects. Thus the clear tendency for more small CS losses up to -$1, with Bayesian estimates, is swamped by the virtual elimination of extreme losses greater than -$5. Similarly, the fortunate tail of extreme CS gains greater than +$5 with ML estimates does not offset their absence with Bayesian estimates.

Figure 8 shows a scatter plot of Efficiency outcomes to allow a literal "head to head" comparison of the effects of using Bayesian estimates rather than ML estimates.[15] Many are indeed virtually identical, as shown on the 45° line. But we see a large number of individuals for whom the estimates are strikingly different. And the majority of deviations *below* the 45° line correspond to the improvements in Efficiency that flow from using the Bayesian estimates (per the bottom panel of Figure 7).

We make no formal inferences about the effects of using Bayesian estimates instead of ML

[15] In this case it is appropriate to limit the sample to those that have both ML *and* Bayesian estimates.

estimates on *average* CS or average Efficiency. We could, from inspection of Figures 7 and 8, but we stress that welfare evaluation in the context of preference heterogeneity must not be about central tendencies. It should always be about *distributions* of welfare effects.

3. Extensions

The Bayesian approach illustrated here was designed to solve a specific problem that arises in behavioral welfare economics: ascertaining reliable and *a priori* sensible estimates of risk preferences for individuals, which are in turn used to condition normative inferences about some other choices. The approach is quite general. There are some exciting extensions that can be considered.

A. Reducing the Number of Choices Each Subject Has to Make

One extension is to evaluate settings in which each individual was only presented with a random sub-set of the full range of risky lottery choices. In our experiment every subject was asked the same 80 questions, albeit in random order that varied from subject to subject. What if we had selected 60 for each subject, at random and without repetition? Or 50, or 40? Would we have obtained comparable estimates? By selecting a smaller set of choices at random for each subject, we ensure "coverage" over the full range of questions for the pooled sample of individuals, which can be important for addressing different aspects of the structure of risk preferences relevent to the target choice for normative evaluation.[16] Having full coverage of the complete battery allows the hierarchical model to generate good estimates of the posterior for the pooled sample that is used as

[16] For example, Harrison and Ng [2016; p. 99][2018; p. 49-51] discuss in detail why different types of lottery questions are included in their full battery for different type of normative inferences. In the latter case, focused on compound risks from non-performance of insurance contracts (e.g., due to fraud or bankruptcy), it was critical to estimate risk preferences that included compound lotteries.

an informative prior for the inferences about individual subjects.

This is not just an idle technical question. Reducing the number of questions any one individual has to make can be particularly valuable in field settings. Invariably in those settings one is under time pressure in terms of how long the subject can be expected to focus on artefactual tasks of this kind, even with compensation. This is particularly true when estimating risk preferences is not the primary focus of the field experiment: in some cases it is just a "nuisance parameter" that would be valuable to have, but not something that can take up the entire session. Even in the field settings of policy interest to us, evaluating various insurance options where knowing risk preferences is foundational to the behavioral welfare evaluation, we must have multiple tasks as well as the risk preference elicitation.[17] Hence time is a crtical factor in experimental design, and it would be valuable to know the trade-off with accuracy that comes with reducing the number of choices each subject has to make.

We can explore this trade-off with our data, to illustrate. Consider the restriction to ask subjects only 20 questions, rather than 80. As suggested, allow those 20 questions to be drawn at random for each subject, without replacement, from the full battery. Then re-estimate the Bayesian hierarchical model with just these 20 questions over the 111 subjects, and compare results with the estimates using the full battery.

Figure 9 displays the results of this exercise in restricting the number of questions asked of each subject to 25% of the total. In each panel we display a scattergram of the estimate for an individual of some risk preference parameter (r, η or φ) or welfare measure (Efficiency). These are, again, based on the posterior average "point estimate" for each individual. Remarkably, the

[17] Apart from the obvious need to ask questions about insurance purchases, in field settings we are also interested in eliciting preferences about time preferences, subjective beliefs, intetemporal risk aversion, and possibly even social preferences.

correlation for Efficiency, the target or normative evaluation, is 0.79 in this instance. For the utility curvature parameter r the correlation is slightly higher, and for the probability weighting parameters η or φ it is considerably lower. Given the dramatic reduction in the number of questions required of each subject, we view this as likely to be an acceptable trade-off for many field researchers.

If we consider, instead, a reduction of the number of questions for each subject to 50% of the full battery, which is 40 questions for each subject, the results are dramatic. Figure 10 provides a comparable display. Now we achieve a correlation of 0.90 for Efficiency when we use the reduced task for each subject. Again, the probability weighting parameters have the lowest correlations, particularly η at 0.68, but if the focus of analysis is Efficiency, and r, η or φ are "nuisance parameters," then this relatively low correlation is of no concern. Just to round out the evaluation, if we reduce the number of questions to 75% of the full battery, which is 60 questions for each subject, we achieve a correlation with Efficiency of 0.97, and 0.97, 0.90 and 0.94 for the r, η or φ risk preference parameters, respectively.

Obviously these are valuable trade-offs when it comes to field, or even lab, experiments. Our methodological point is to stress how they flow naturally from thinking about pooled data being used to inform priors for inference about individuals. The reason we get such high correlations for Efficiency with just 20 or 40 questions per subject, rather than all 80, is that the pooled data spans all 80 questions.

In a similar vein, another type of extension would be to evaluate the use of disjoint samples from the same population. One might imagine one sub-sample being asked all 80 questions, to help condition the posterior distribution of the representative agent, and then the other sub-sample being asked far fewer questions.[18] Again, field settings are natural here: one might have a large-scale survey

[18] In principle one could also identify *which* of the full battery of questions are most informative to ask, which is just a "pre-posterior" analysis to a Bayesian. Lindley [1972; p, 20ff.] provided the first general,

of tens of thousands, and can afford the time and money to ask only a few risky lottery choices. One could then then have a much smaller sample, drawn appropriately from the same population, that is recruited for a longer, more demanding series of risky lottery choices.

B. Inferring the Distribution of Welfare

For comparability to the traditional ML analysis employed by Harrison and Ng [2016] and others, we focused on inferences about welfare that used a "point estimate" from the posterior distribution of risk preference parameters r, η or φ. The correct inferences should take into account the fact that these are *full* posterior distributions.[19] Due to the significant non-linearity of the prediction measure, the mean of the *distribution* of CS evaluated over the *distribution* of r, η and φ can be quite different from the CS evaluated at the *mean* of r, η and φ. In Bayesian jargon, we should calculate the *posterior predictive distribution* of welfare for each insurance choice of an individual. The predictive distribution is just a distribution of unobserved data (the expected insurance choice given the actuarial parameters offered) conditional on observed data (the actual choices in the risk lottery task).[20] All that is involved is marginalizing the likelihood function for the insurance choices with respect to the posterior distribution of model parameters from the risk lottery choices. The upshot is that we predict a *distribution* of welfare for a given choice by a given individual, rather than a *scalar*.

---

formal statement of Bayesian experimental design, and Chaloner and Verdinelli [1995] a valuable literature review. Gelman et al. [2013; ch. 8] review complementary literature on how various experimental designs impact Bayesian analyses.

[19] As noted earlier, Harrison and Ng [2016; p. 110/111] show how one can bootstrap the welfare calculations to reflect the covariance matrix of ML estimates for each individual. So the ML approach also allows one to calculate distributions of welfare, although with a very different interpretation.

[20] Perhaps a simpler and more familiar way to think of a posterior predictive distribution is to imagine that the subject was faced with a new battery of risk lotteries and we use the observed behavior from the old battery of risk lotteries to infer what choices would be made for the new battery. The posterior estimates of r, η or φ from the old choices are used to characterize the data-generating process, and then infer the distribution of expected choices for the new battery. In our case we substitute insurance choices for a new risk lottery battery, but the statistical principles are the same.

We can then report that distribution as a kernel density, or select some measure of central tendency such as the mean or median.

We consider the mean of the posterior predictive distribution of Efficiency for each individual. Figure 11 displays a scattergram of these means for the smaller sample sizes assumed for each subject (20, 40 or 60) against the means for the full sample size (80). Again, there is a quantified tradeoff in reliability that is apparent as the sample size is reduced, and these appear again to be relatively small tradeoffs for the savings in the number of tasks required of each subject. Of course these judgments must be made by the researcher, or those funding the research, but it is critical that they be quantified to inform that judgment.

C. Adaptive Welfare Evaluation

Some of our subjects gain from virtually every opportunity to purchase insurance, and sadly some lose with equal persistence over the 24 sequential choices. Armed with posterior predictive estimates of the welfare gain or loss distribution for each subject and each choice, can we adaptively identify *when* to withdraw the insurance product from these persistent losers, and thereby avoid them incurring such large welfare losses? Important recent research by Caria et al. [2020], Hadad et al. [2020] and Kasy and Sautmann [2019] considers this general issue. The challenges are significant, from the effects on inference about confidence intervals, to the implications for optimal sampling intensity, to the weight to be given to multiple treatment arms, and so on.

We consider a simple application of our Bayesian approach to behavioral welfare economics to illustrate some important issues. Assume that the experimenter could have decided to stop offering the insurance product to an individual at the mid-point of their series of 24 choices, so the

sole treatment arm was to discontinue the product offering or continue to offer it.[21] Recall that the order of insurance products, differentiated by their actuarial parameters, was randomly assigned to each subject.[22] Figure 12 displays the sequence of welfare evaluations possible for subject #1. The two solid lines show measures of the CS: in one case the average gain or loss from the observed decision in that period, and in the other case the cumulative gain or loss over time. Here the average refers to the posterior predictive distribution for this subject and each decision. Since this is a distribution, we can evaluate the Bayesian probability that *each* decision resulted in a gain or no loss, reflecting a qualitative Do No Harm (DNH) metric enshrined in the *Belmont Report* as applied to behavioral research.[23] This probability is presented in Figure 1, in cumulative form, by the dashed line and references the right-hand vertical axis.

Although there are some gains and losses in average CS along the way, and the posterior predictive probability declines more or less steadily towards 0.5 over time, the probability of DNH is always greater than 50:50 for this subject. And there is a steady, cumulative gain in expected CS over time. These outcomes reflect a common pattern in our data, with small CS losses often being more than offset by larger CS gains. Hence one can, and should, view these as a temporal series of "policy lotteries" which are being offered to the subject, if the policy of offering the insurance contract is in place (Harrison [2011]). In this spirit, we can think of the probabilities underlying the posterior predictive probability of DNH as the probabilities of positive or negative CS outcomes, given the

[21] Evaluation of multiple treatment arms for a comparable insurance product, and using similar evaluations of individual welfare, are provided by Harrison, Morsink and Schneider [2020].

[22] A more sophisticated "targeting" policy might use the information from the first 12 insurance choices to adaptively determine the actuarial parameters that might lead each subject to make better decisions in the remaining 12 decisions.

[23] See Teele [2014] and Glennerster [2017] for discussion of the *Belmont Report* and the ethics of conducting randomized behavioral interventions in economics. Even when randomized clinical trials were not adaptive, or even sequential in terms of stopping rules, it has long been common to employ termination rules based on extreme, cumulative results (e.g., the "3 standard deviations" rule noted by Peto [1985; p. 33]).

risk preferences of the subject. So the fact that the EV of this series of lotteries is positive, even as the probability approaches 0.5, reflects the asymmetry of CS gains and losses in quantitative terms and the policy importance of such quantification. For now, we can think of the *policy maker* as exhibiting risk neutral preferences over policy lotteries, but recognizing that the evaluation of the purchase lottery by the subject should properly reflect her risk preferences.

Consider comparable evaluations for four individuals from our sample in Figure 13. Subject #5 is a "clear loser," despite the occasional choice that generates an average welfare gain. It is exactly this type of subject one would expect to be better off if not offered the insurance product after period 12 (or, for that matter and with hindsight, at all). Subject #111 is a much more challenging case. By period 12 the qualitative DNH metric is around 0.5, and barely gets far above it for the remaining periods. And yet the EV of the policy lottery is positive, as shown by the steadily increasing cumulative CS. This example sharply demonstrates the "policy lottery" point referred to for subject #1 in Figure 12.

The remaining subjects in Figure 13 illustrate different points: that we should also consider the preferences of the agent when evaluating the policy lottery of not offering the insurance product after period 12. Assume that these periods reflect non-trivial time periods, such as a month, a harvesting season, or even a year. In that case the temporal pattern for subject #67 encourages us to worry about how patient subject #67 is: the cumulative CS is positive by the end of period 24, but if later periods are discounted sufficiently, the subjective present value of being offered the insurance product could be negative due to the early CS losses.[24] Similarly, consider the volatility *over time* of the CS gains and losses faced by subject #14, even if the cumulative CS is positive throughout. In this

[24] This point has nothing to do with whether the subject exhibits "present bias" in any form. All that is needed is simple impatience, even with Exponential discounting. Andersen, Harrison, Lau and Rutström [2008] consider the joint estimation of risk and time preferences. Berry and Fristedt [1985; chapter 3] stress the importance of time discounting in sequential "bandit" problems in medical settings.

case a complete evaluation of the policy lottery for this subject should take into account the *intertemporal* risk aversion of the subject, which arises if the subject behaves consistently with a non-additive intertemporal utility function over the 24 periods.[25]

Applying the policy of withdrawing the insurance product after period 12 for those individuals with a cumulative CS that is negative results in an aggregate welfare gain of 108%, implicitly assuming a classical utilitarian social welfare function over all 111 subjects.

4. Conclusions

There are immediate reasons why one would want to use Bayesian estimates of risk preferences for the type of normative exercise illustrated here: more systematic control of the use of priors over plausible risk preferences, and the ability to make inferences for every individual in a sample.

There are also more general reasons for wanting to adopt a Bayesian approach, to make explicit the role for priors when making normative evaluations.

One general reason for a Bayesian approach derives from the ethical need to pool data from randomized evaluations and non-randomized evaluations. The ethical need first arises when *defining* the prior beliefs that justify a randomized trial with equal probabilities of control and treatment in the first place.[26] In general we need to be able to pool disparate sources of data, even observational

[25] The intertemporal risk aversion of a subject, also referred to as "correlation aversion," bears no necessary relationship to atemporal risk aversion. Andersen, Harrison, Lau and Rutström [2018] consider the joint estimation of atemporal risk preferences, time preferences, and intertemporal risk preferences.

[26] Commenting on the famous Extracorporeal Membrane Oxygenation (ECMO) adaptive randomization study for babies documented by Ware [1989], Royall [1989] and Berry [1989; p. 306] reject the claim that prior, well-known evidence from a randomized evaluation documented by Bartlett et al. [1985] supported such a perfectly diffuse prior. Kass and Greenhouse [1989; p. 313] raise similar concerns, but in the end explicitly, and reluctantly, assume that the study was "appropriately designed" to start with a diffuse prior. Royall [1989; p. 318] calculates the posterior probability that the ECMO treatment was inferior to be either 0.01 or 0.00003 based on previous data. Berry [1989; p.310] sharply concludes that "clinical equipoise is an invention used to avoid difficult ethical questions." In the context of economics experiments, that equipoise

studies, to form priors for ethical grounds prior to randomization, and that type of pooling is exactly what Bayesian analysis facilitates. The ethical need also arises *during and after* the trial, when determining what to make of the results in the context of many other sources of information that are *not* directly comparable (i.e., exchangeable). This issue arises so often that it cannot be set aside from the instant trial.[27]

Another general reason for a Bayesian approach derives from the methodological need for normative analysis to have estimates of risk preferences from choice tasks *other than the choice task one is making welfare evaluations about*. In settings of this kind, it is natural to want to debate and discuss the appropriateness of the risk preferences being used. In fact, the need for debate and conversation becomes more urgent when, as here, we infer significant losses in expected CS, and significant foregone Efficiency. How do we know that the task we used to infer risk preferences, or even the models of risk preference we used, are the right ones? The obvious answer: we don't. We can only hold prior beliefs about those, and related questions. And when it comes to systematically examining the role of alternative priors on posterior-based inference, one wants to be using Bayesian formalisms.

An example to illustrate this general point. Imagine one was designing a field experiment, say in rural Ethiopia, in which various interventions for a health insurance product were to be used to improve welfare. Assume a health insurance product focused on acute conditions, with significant mortality risk. The only priors on risk preferences you have come from university students in the United States. Should you go ahead and design interventions that, conditional on those risk preferences, lead to welfare losses for the same students, of the kind we have demonstrated? We

---

corresponds to claims that "anything *could* happen," as distinct from "here is what I believe *would* happen." Freedman [1987] first proposed the notion of clinical equipoise, controversially defining it in terms of priors that are presumed to be held in the broader research field, not the priors of the immediate investigators.

[27] See Yusuf et al. [1985], Peto [1985; p. 33] and Armitage [1985; p.19/20] for discussion.

suggest that, ethically speaking, you should not.

Now imagine you have been able to conduct comparable artefactual field experiments over *money* in Ethiopia that allow you to infer risk preferences, and assume that these experiments match the standard criteria we have for taking any experimental data seriously (e.g., financial incentives and incentive compatibility). These are obviously better priors for the eventual inference, and should be used. You completely discard the priors from students in the United States, or give them relatively lower weight in your hierarchical priors.

Then imagine that you have been able to conduct artefactual field experiments over *certain* health outcomes in Ethiopia that allow you to infer risk preferences. Assume that these health outcomes refer to morbidity risks, not mortality risks, but to real outcomes nonetheless. As any experimental economist knows, it is not easy to come up with morbidity outcomes that can be credibly and ethically delivered within the budgets we normally find ourselves in. Clearly the domain of risk preferences here is *closer* than the risk preferences defined over money, but would you now attach zero or negligible weight to the risk preferences over money by similar Ethiopians? Probably not. So how do you pool these priors to arrive at inferences? The answer is to be Bayesian.

Figure 1: Risk Preferences for Representative Subject

Maximum Likelihood *versus* Bayesian Estimates for all 111 subjects

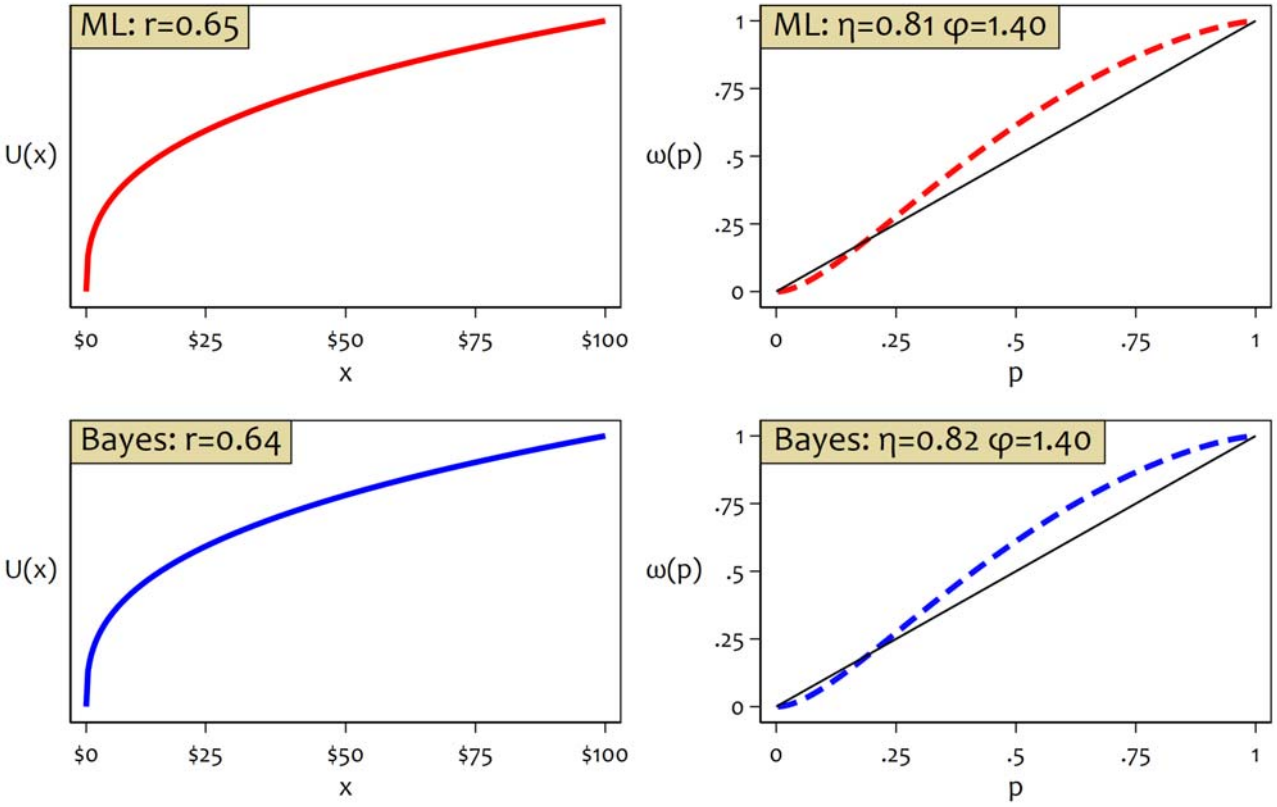


Figure 2: Distributions of Individual Risk Preference Parameters from Hierarchical Bayesian Model

Mean posterior estimate for each of N=111 subjects
Dashed lines indicate posterior averages

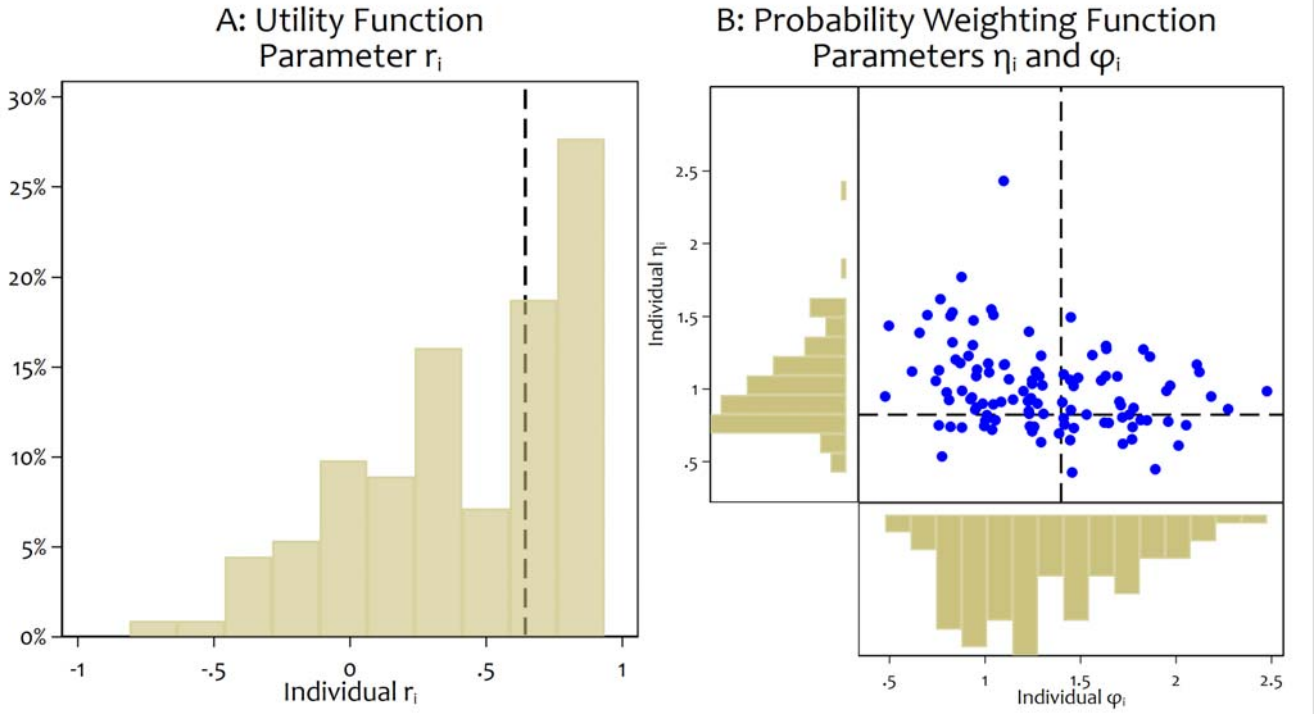


Figure 3: Risk Preferences for One Subject

Maximum Likelihood *versus* Bayesian Estimates for the same subject

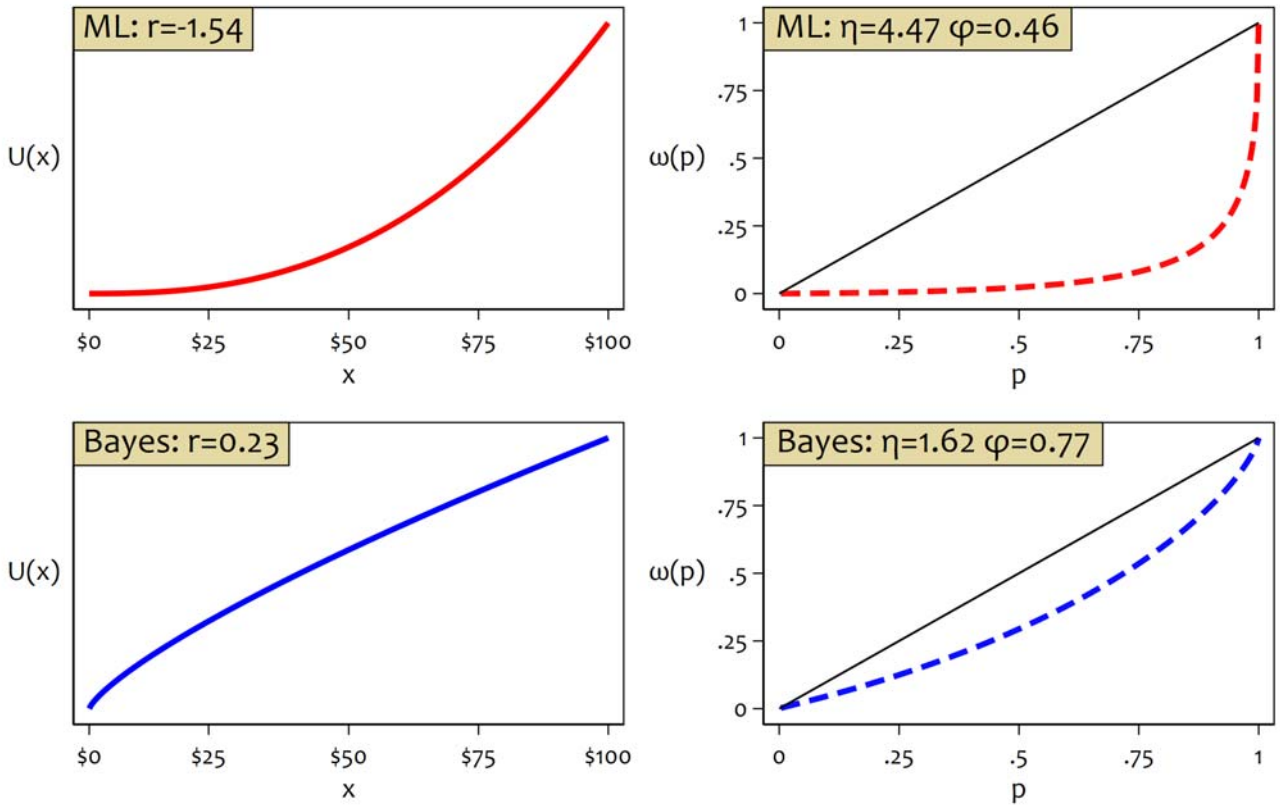


Figure 4: Risk Preferences for A Second Subject

Maximum Likelihood *versus* Bayesian Estimates for the same subject

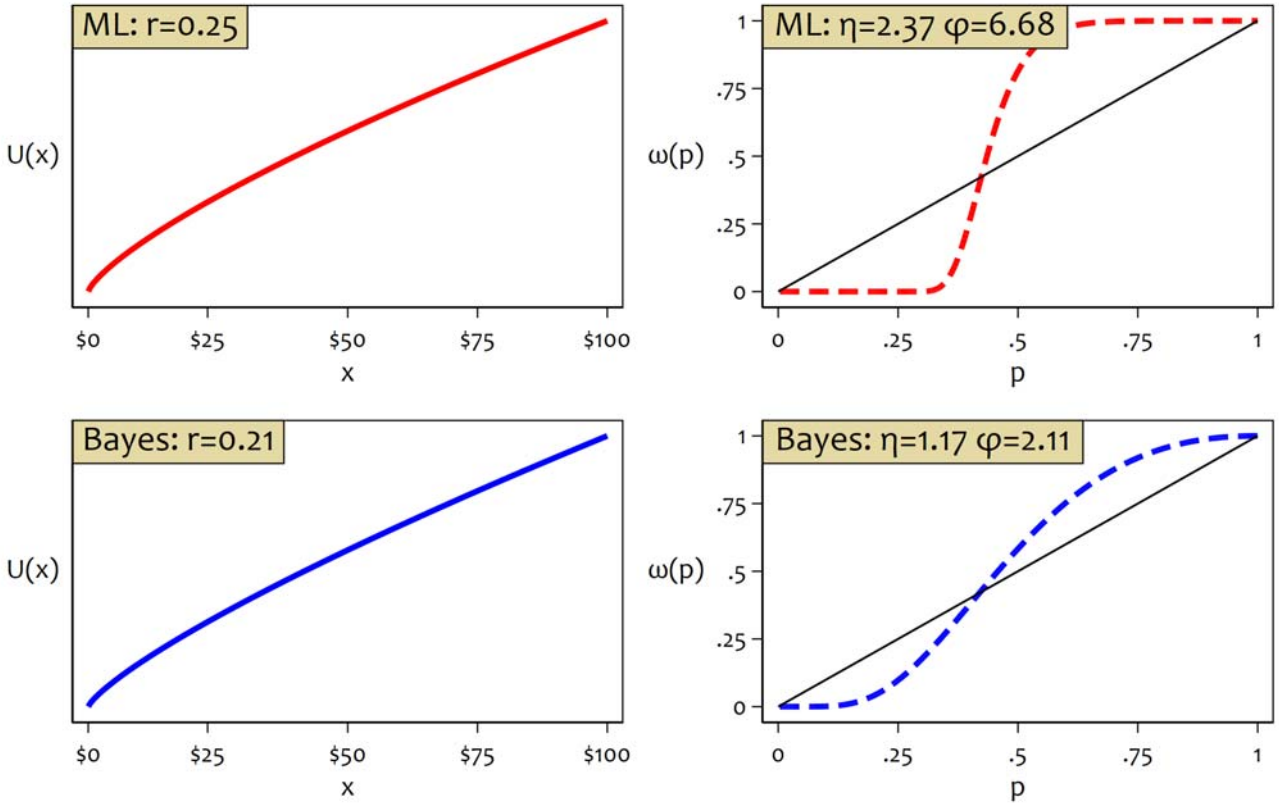


Figure 5: Risk Preferences for A Third Subject

Maximum Likelihood *versus* Bayesian Estimates for the same subject

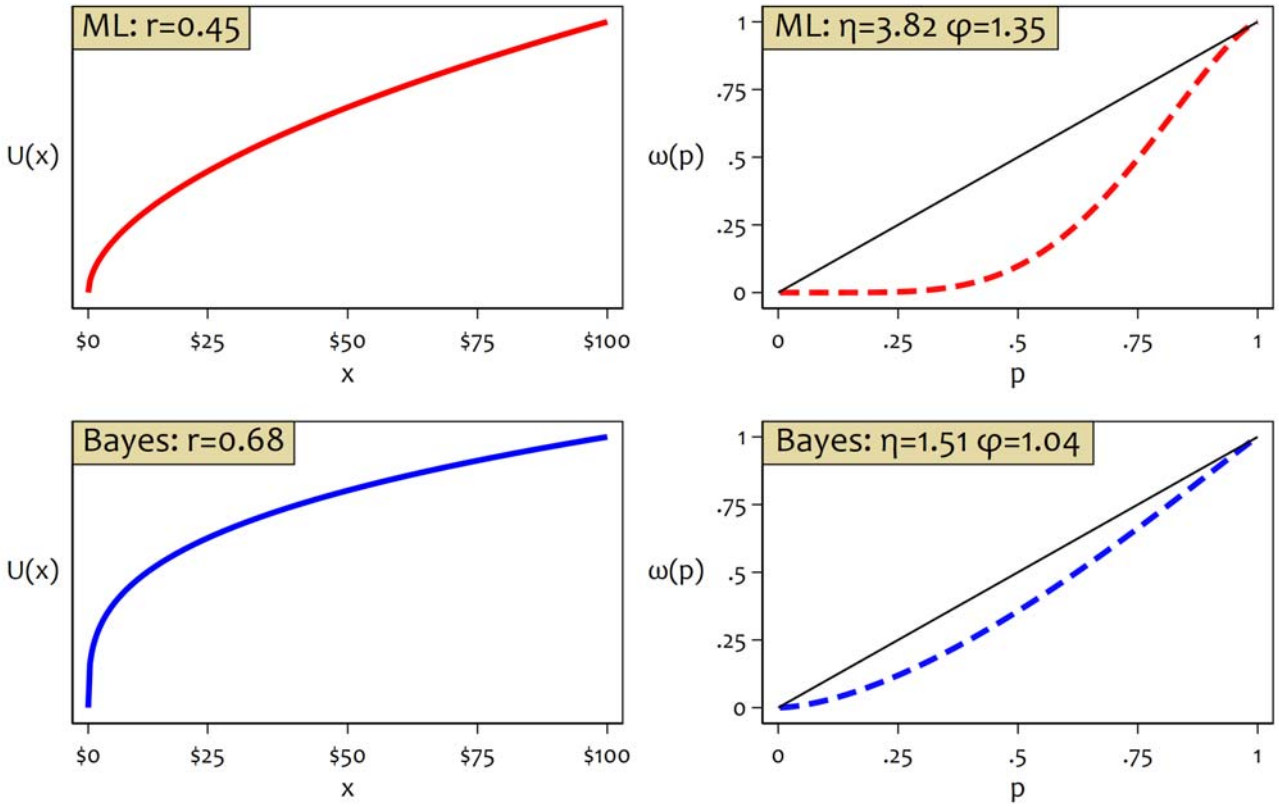
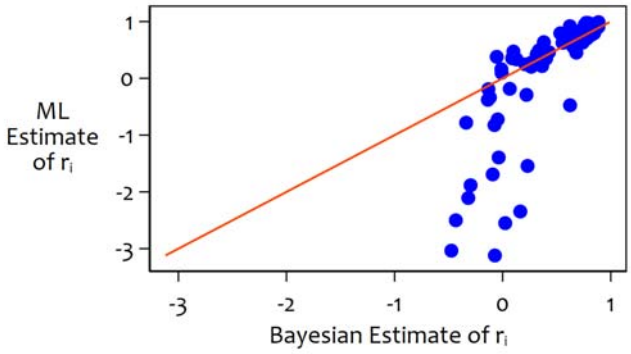
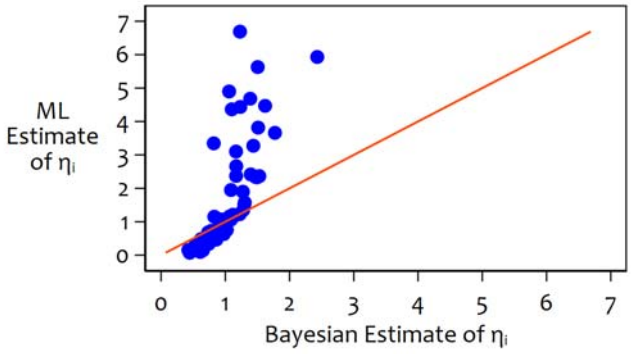


Figure 6: Comparison of ML and Bayesian Estimates of Individual Risk Preference Parameters

Mean posterior estimate for each of N=111 subjects





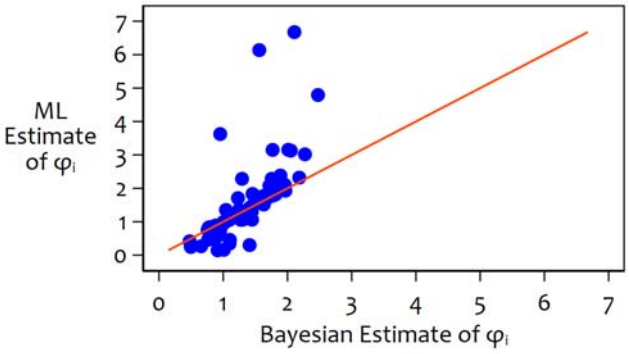


Figure 7: Effects of Inferences About Risk on Welfare

Consumer Surplus based on 24 insurance purchase decisions per individual
Efficiency defined over all 24 decision of each individual

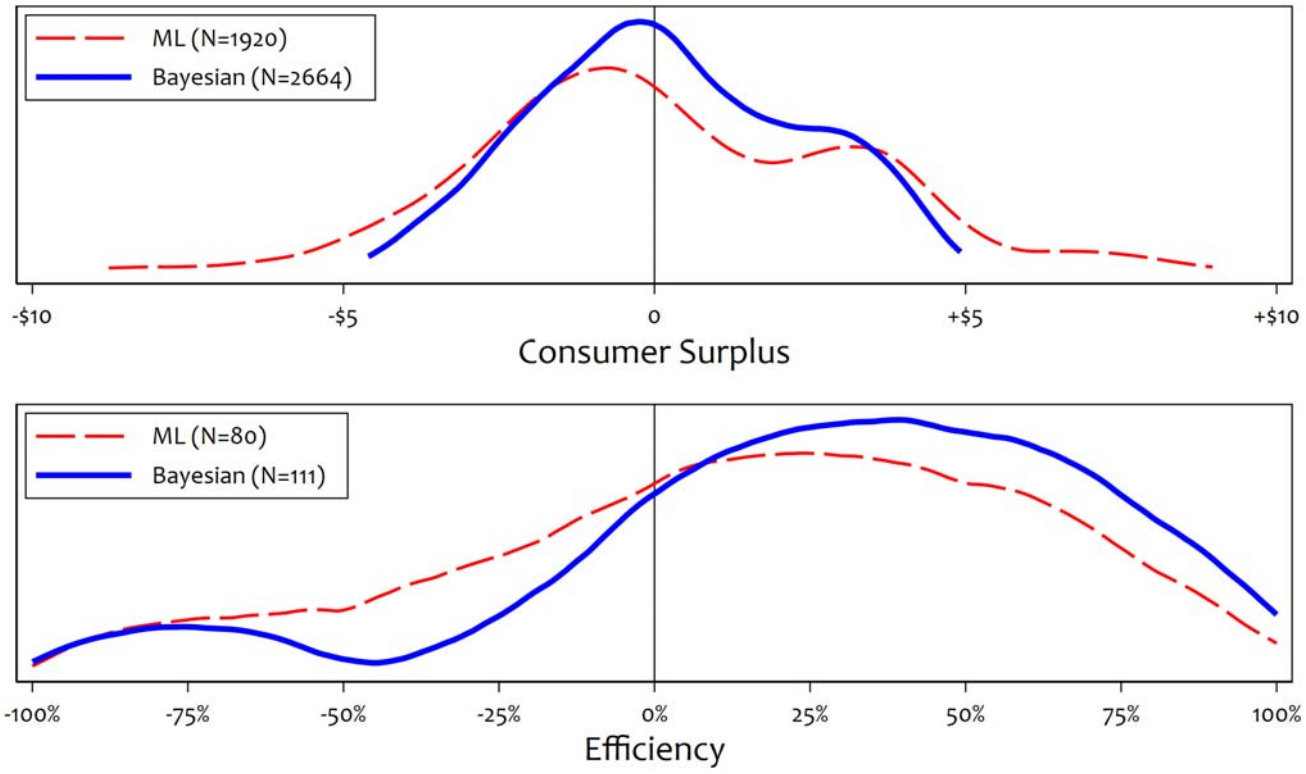


Figure 8: Scatter Plot of Effect of Inferences About Risk on Welfare

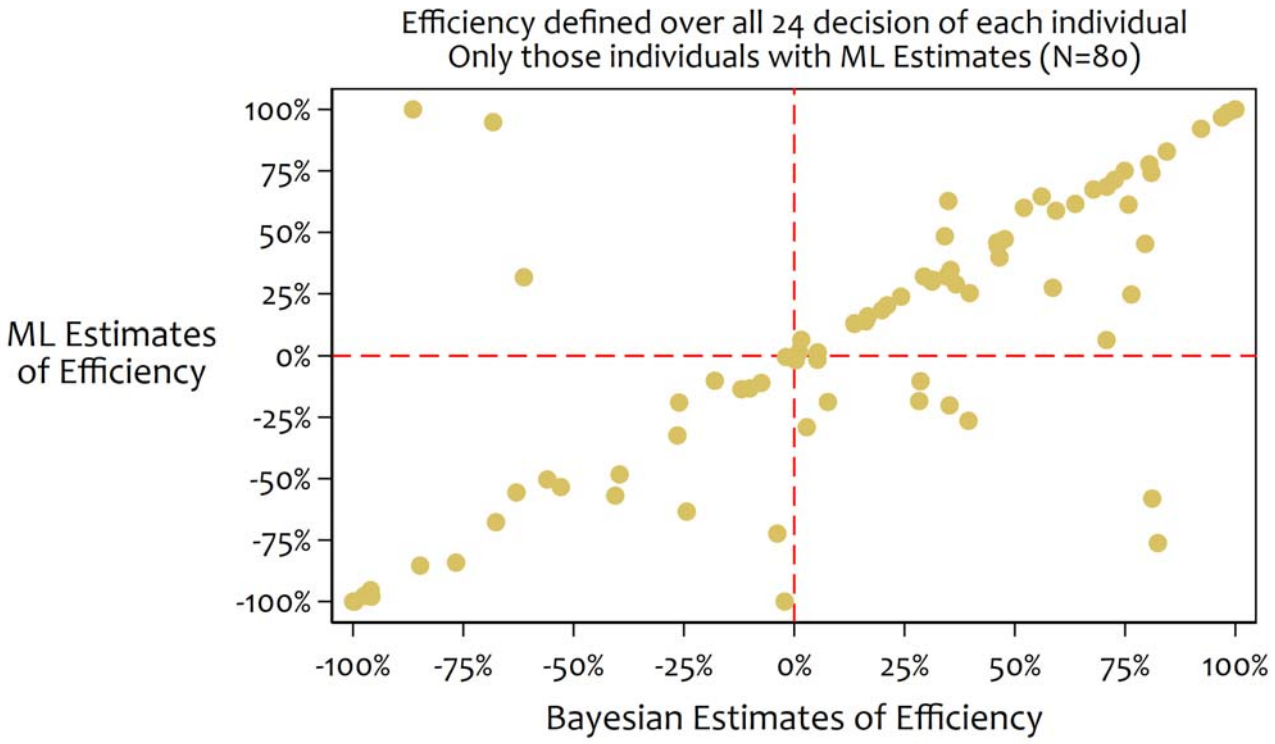


Figure 9: Comparison of Bayesian Hierarchical Estimates of Individual Risk Preference Parameters with Sample Sizes of 20 and 80 For Each Subject

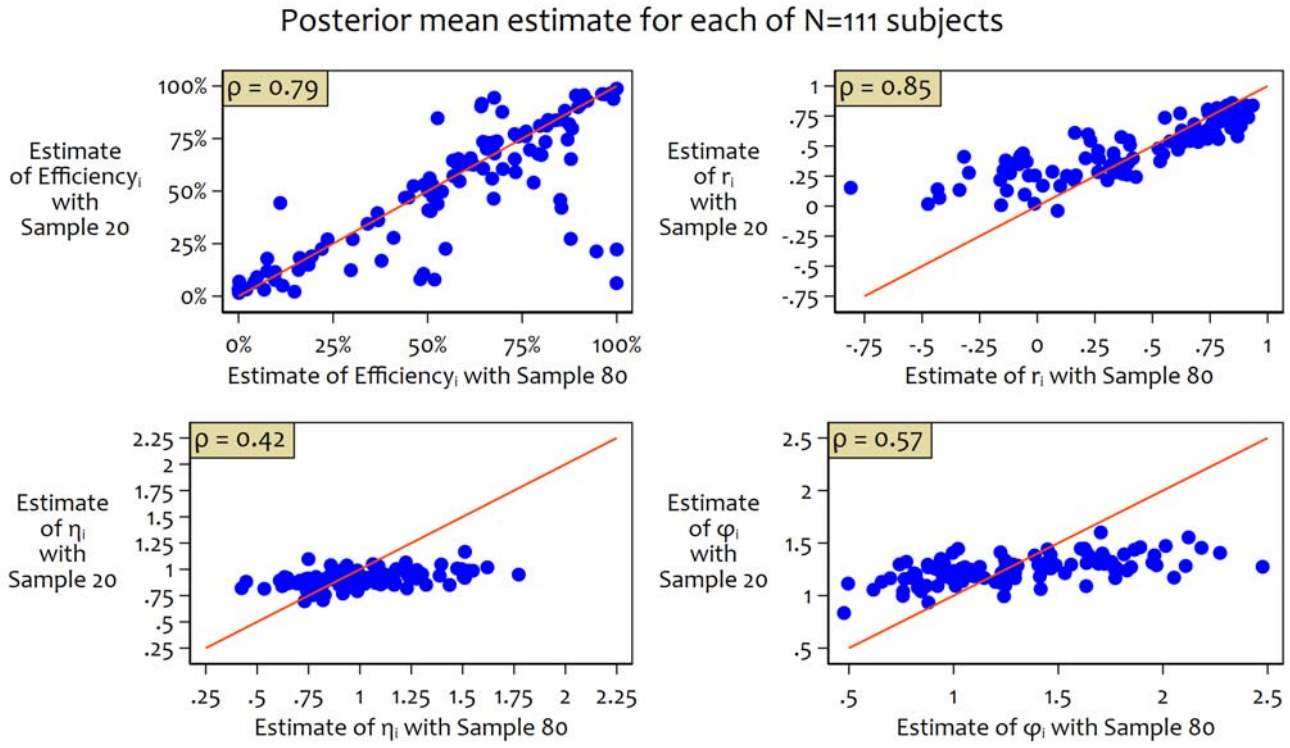


Figure 10: Comparison of Bayesian Hierarchical Estimates of Individual Risk Preference Parameters with Sample Sizes of 40 and 80 For Each Subject

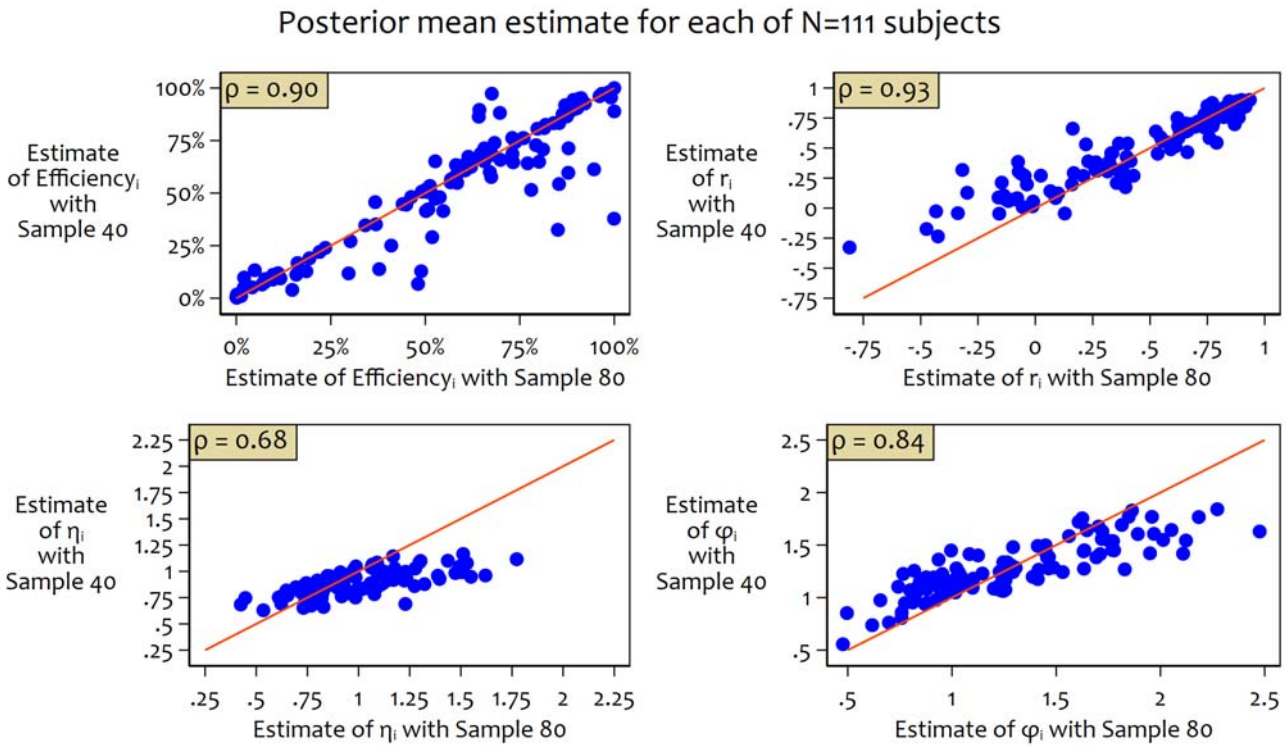


Figure 11: Comparison of Bayesian Posterior Predictive Estimates of Efficiency with Sample Sizes of 24, 40 or 60 For Each Subject

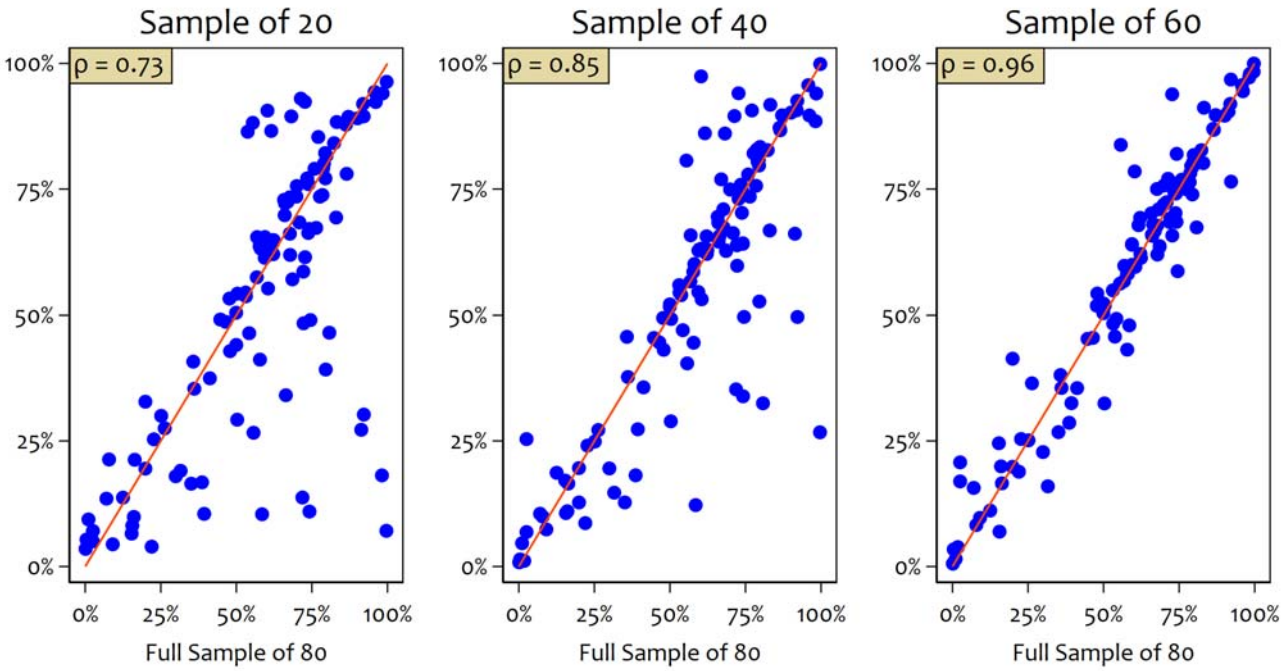


Figure 12: Adaptive Welfare Evaluations for Subject #1

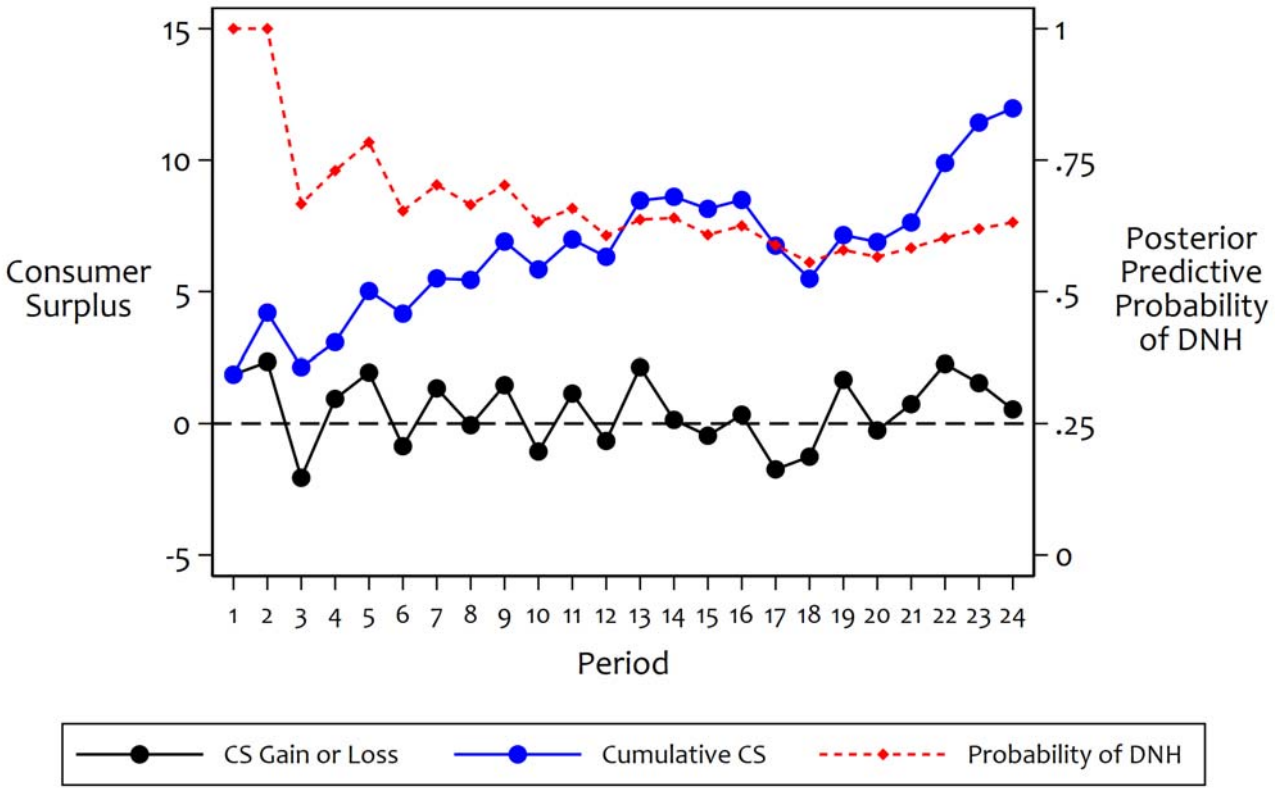
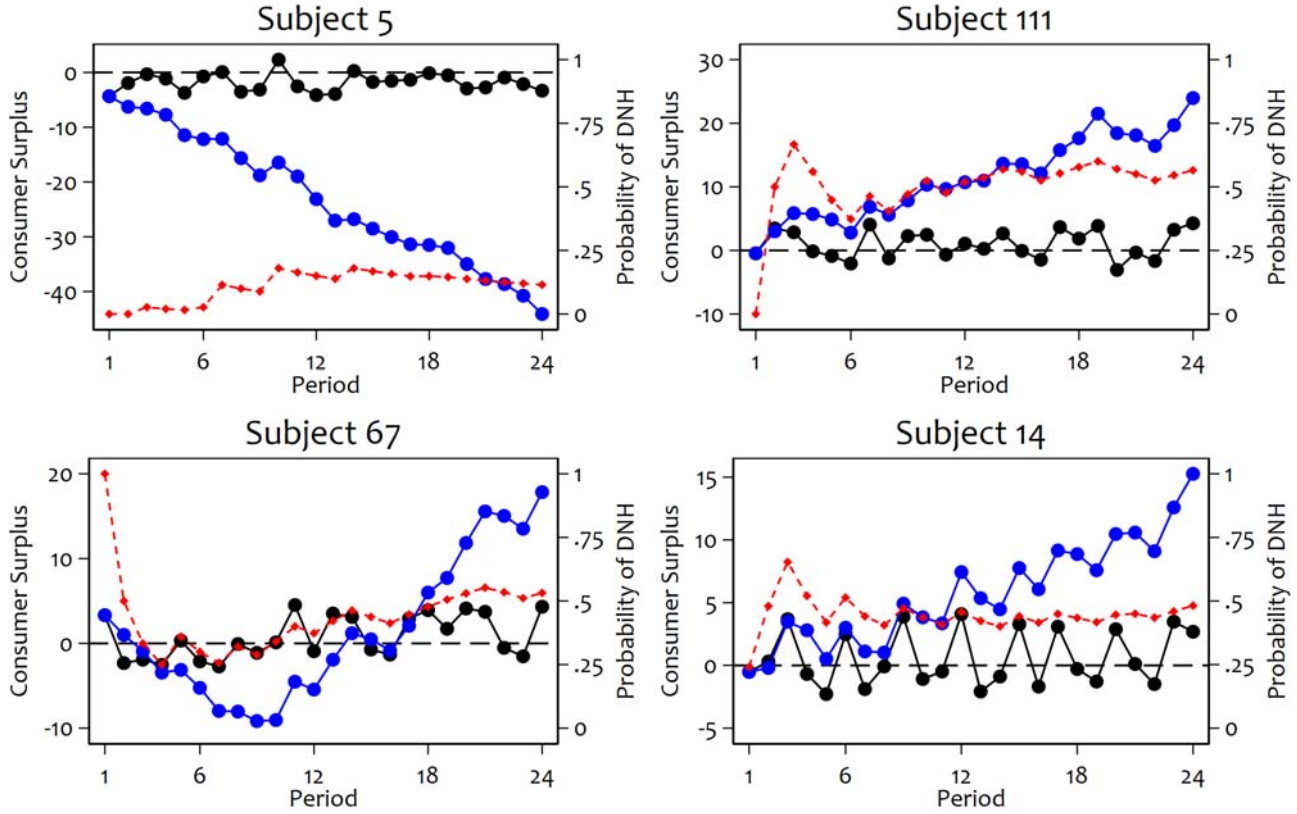


Figure 13: Individual Adaptive Welfare Evaluations for Four Subjects



References

Allenby, Greg M., and Gintner, James L., "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32, November 1995, 392-403.

Allenby, Greg M., and Rossi, Peter E., "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89, 1999, 57-78.

Andersen, Steffen; Fountain, John; Harrison, Glenn W., and Rutström, E. Elisabet, "Estimating Subjective Probabilities," *Journal of Risk & Uncertainty*, 48, 2014, 207-229.

Andersen, Steffen; Harrison, Glenn W.; Lau, Morten Igel, and Rutström, E. Elisabet, "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), May 2008, 583-618.

Andersen, Steffen; Harrison, Glenn W., Lau, Morten I., and Rutström, E. Elisabet, "Multiattribute Utility Theory, Intertemporal Utility, and Correlation Aversion," *International Economic Review*, 59(2), May 2018, 537-555.

Armitage, Paul, "The Search for Optimality in Clinical Trials," *International Statistical Review*, 53(1), 1985, 15-24.

Bartlett, Robert H.; Roloff, Dietrich W.; Cornell, Richard G.; Andrews, Alice French; Dillon, Peter W., and Zwischenberger, Joseph B., "Extracorporeal Circulation in Neonatal Respiratory Failure: A Prospective Randomized Study," *Pediatrics*, 76(4), October 1985, 479-487.

Berry, Donald A., "Comment: Ethics and ECMO," *Statistical Science*, 4(4), 1989, 306-310.

Berry, Donald A., and Fristedt, Bert (eds.), *Bandit Problems: Sequential Allocation of Experiments* (New York: Springer, 1985).

Caria, Stefano; Gordon, Grant; Kasy, Maximilian; Quinn, Simon; Shami, Soha, and Teytelboyn, Alexander, "An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan," *Draft Working Paper*, Oxford University, May 2020; available at https://maxkasy.github.io/home/research/

Chaloner, Kathryn, and Verdinelli, Isabella, "Bayesian Experimental Design: A Review," *Statistical Science*, 10(3), 1995, 273-304.

Freedman, Benjamin, "Equipoise and the Ethics of Clinical Research," *New England Journal of Medicine*, 317(3), 1987, 141-145.

Gao, Xiaoxue Sherry; Harrison, Glenn W., and Tchernis, Rusty, "Estimating Risk Preferences for Individuals: A Bayesian Analysis," *Unpublished Manuscript*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2020.

Gelman, Andrew; Carlin, John B.; Stern, Hal S.; Dunson, David B.; Vehtari, Aki, and Rubin, Donald B., *Bayesian Data Analysis* (Boca Raton, FL, CRC Press, Third Edition 2013).

Glennerster, Rachel, "The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency," in Banerjee, A. and Duflo, E. (eds.), *Handbook of Field Experiments: Volume One* (Amsterdam: North-Holland, 2017).

Hadad, Vitor; Hirshberg, David A.; Zhan, Ruohan; Wager, Stefan, and Athey, Susan, "Confidence Intervals for Policy Evaluation in Adaptive Experiments, *Working Paper*, Stanford University, July 2020; available at https://arxiv.org/abs/1911.02768.

Harrison, Glenn W, "Risk Attitudes in First-Price Auction Experiments: A Bayesian Analysis," *Review of Economics & Statistics*, 72, August 1990, 541-546.

Harrison, Glenn W., "Experimental Methods and the Welfare Evaluation of Policy Lotteries," *European Review of Agricultural Economics*, 38(3), 2011, 335-360.

Harrison, Glenn W., "The Behavioral Welfare Economics of Insurance," *Geneva Risk & Insurance Review*, 44(2), September 2019, 137–175.

Harrison, Glenn W.; Morsink, Karlijn, and Schneider, Mark, "Do No Harm? The Welfare Consequences of Behavioural Interventions," *CEAR Working Paper 2020-12*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2020.

Harrison, Glenn W. and Ng, Jia Min, "Evaluating the Expected Welfare Gain from Insurance," *Journal of Risk and Insurance*, 83(1), 2016, 91–120.

Harrison, Glenn W. and Ng, Jia Min, "Welfare Effects of Insurance Contract Non-Performance," *Geneva Risk & Insurance Review*, 43(1), May 2018, 39-76.

Harrison, Glenn W. and Ross, Don, "Varieties of Paternalism and the Heterogeneity of Utility Structures," *Journal of Economic Methodology*, 25(1), 2018, 42–67.

Harrison, Glenn W., and Rutström, E. Elisabet, "Risk Aversion in the Laboratory," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Huber, Joel, and Train, Kenneth, "On the Similarity of Classical and Bayesian Estimates of Individual Mean Partworths," *Marketing Letters*, 12(3), 2001, 259-269.

Kass, Robert E., and Greenhouse, Joel B., "Comment: A Bayesian Perspective," *Statistical Science*, 4(4), 1989, 310-317.

Kasy, Maximilian, and Sautmann, Amja, "Adaptive Treatment Assignment in Experiments for Policy Choice, *Working Paper*, Oxford University, December 2019; available at https://maxkasy.github.io/home/research/, *Econometrica*, forthcoming.

Kitagawa, Toru, and Tetenov, Aleksey, "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86(2), 2008, 591-616.

Kruschke, John K., *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (Burlington, MA: Academic Press, Second Edition, 2015).

Kruschke, John K., and Liddell, Torrin M., "The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective," *Psychonomic Bulletin & Review*, 25, 2018, 178-206.

Kruschke, John K., and Vanpaemel, Wolf, "Bayesian Estimation in Hierarchical Models," in Busemeyer, J.R., Townsend, J.T., Wang, Z.J., and Eidels, A. (eds.) *Oxford Handbook of Computational and Mathematical Psychology* (Oxford, UK: Oxford University Press, 2015).

Leamer, Edward E., *Specifcation Searches: Ad Hoc Inference with Nonexperimental Data* (New York: Wiley, 1978).

Lindley, David V., *Bayesian Statistics: A Review* (Philadelphia, PA: Society for Industrial and Applied Mathematics, 1972).

McCulloch, Robert; Rossi, Peter E., and Allenby, Greg M., "Hierarchical Modeling of Consumer Heterogeneity: an Application to Targeting," in C. Gatsonis, J.S. Hodges, E.E. Kass and N.D. Singpurwalla (eds.), *Case Studies in Bayesian Statistics, Volume II* (New York: Springer, Lecture Notes in Statistics, vol 105).

Monroe, Brian, "The Welfare Consequences of Individual-Level Risk Preference Estimation," in G.W. Harrison and D. Ross (eds.), *Models of Risk Preferences: Descriptive and Normative Challenges* (Bingley, UK: Emerald, Research in Experimental Economics, 2021 forthcoming).

Murphy, Ryan O., and ten Brincke, Robert H.W., "Hierarchical Maximum Likelihood Parameter Estimation for Cumulative Prospect Theory: Improving the Reliability of Individual Risk Parameter Estimates," *Management Science*, 64(1), January 2018, 308-326.

Nilsson, Håkan; Rieskamp, Jörg, and Wagenmakers, Eric-Jan, "Hierarchical Bayesian Parameter Estimation for Cumulative Prospect Theory," *Journal of Mathematical Psychology*, 55, 2011, 84-93.

Peto, Richard, "Discussion of Papers by J.A. Bather and P. Armitage," *International Statistical Review*, 53(1), 1985, 31-34.

Prelec, Drazen, "The Probability Weighting Function," *Econometrica*, 66, 1998, 497-527.

Quiggin, John, "A Theory of Anticipated Utility," *Journal of Economic Behavior & Organization*, 3(4), 1982, 323-343.

Rossi, Peter E., and Allenby, Greg, M., "A Bayesian Approach to Estimating Household Parameters," *Journal of Marketing Research*, 30, May 1993, 171-182.

Rossi, Peter E.; Allenby, Greg, M., and McCulloch, Robert, *Bayesian Statistics and Marketing* (Chichester, UK: Wiley, 2005).

Royall, Richard, "Comment," *Statistical Science*, 4(4), 1989, 318-319.

Teele, Dawn Langan, "Reflections on the Ethics of Field Experiments," in Teele, D. (ed.), *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences* (New Haven, NJ: Yale University Press, 2014).

Train, Kenneth, *Discrete Choice Methods with Simulation* (New York: Cambridge University Press, Second Edition, 2009).

Ware, James H., "Investigating Therapies of Potentially Great Benefit: ECMO," *Statistical Science*, 4(4), 1989, 298-306.

Yaari, Menahem E., "The Dual Theory of Choice under Risk," *Econometrica*, 55(1), 1987, 95-115.

Yusuf, Salim; Peto, Richard; Lewis, John; Collins, Rory, and Sleight, Peter, "Beta Blockade During and After Mycoardial Infarction: An Overview of the Randomized Trials," *Progress in Cardiovascular Diseases*, 28(5), 1985, 335-371.