Behavioral Welfare Economics and Risk Preferences:

A Bayesian Approach

by

Xiaoxue Sherry Gao, Glenn W. Harrison and Rusty Tchernis †

December 2021

Forthcoming, *Experimental Economics*

ABSTRACT.

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.

Keywords: Behavioral Welfare Economics, Bayesian Analysis, Risk Preferences, Insurance
JEL Codes: D6, C11, D81, G40

† Department of Resource Economics, University of Massachusetts Amherst (Gao); Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University (Harrison); and Department of Economics, Andrew Young School of Policy Studies, Georgia State University (Tchernis). Harrison is also affiliated with the School of Economics, University of Cape Town, and Tchernis is also affiliated with IZA and NBER. E-mail contacts: xiaoxuegao@umass.edu, gharrison@gsu.edu and rtchernis@gsu.edu. We are grateful to Andre Hofmeyr, two referees and the editor 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, rigorously and elegantly. We propose, and constructively illustrate, how to undertake a Bayesian analysis in this way for applications in behavioral welfare economics.[1] We

[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] 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.

focus initially 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 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]

We deliberately refer to the hierarchical prior as if it refers to a "representative agent," to better connect what we are doing to popular classical approaches to estimating risk preferences. The expression "representative agent" is just a short-hand for viewing the behavior of many agents as if they were one agent. In the usual case this involves unconditionally pooling choices from many individuals and assuming that they all have the same risk preferences, even if one does allow for clustered statistical errors in the usual classical sense. Sometimes it involves pooling choices from many individuals and then modeling risk preferences conditional on a handful of observable demographic characteristics, such as gender and age. And sometimes it involves pooling choices from many individuals and conditioning on latent data-generating processes, such as EUT or Prospect Theory, to account for individual or subject choices. Harrison and Rutström [2008][2009] review classical methods for pooling behavior and estimating risk preferences in these unconditional or conditional ways.

The classical approach has been to either undertake pooled analyses, with or without conditioning, *or* to estimate using only the data for each individual. Although one could combine these methods,[5] doing so in a Bayesian hierarchical manner does so more systematically and elegantly because of the way in which priors can be incorporated. For example, Kruschke [2015;

---

[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] For example, by estimating a pooled model with covariates and predicting risk preference parameters for each subject, with standard errors. The predictions for each subject then, at least, utilize the exact values of those covariates for that subject. Then one could use those predictions in lieu of estimates with individual subject data for those subjects whose ML estimates fail to converge or violate some *ad hoc* priors as to what constitute sensible estimates.

§9.3] reminds us, by construction, that hierarchical models can be set up and estimated using maximum likelihood models, and exhibit some of the *qualitative* "shrinkage prior" effects that one finds with a Bayesian Hierarchical Model. The Bayesian approach just allows for more control over the manner in which data from the whole sample is pooled with data from an individual. Hence we should be clear that it is not the use of priors *per se* that is the advantage of Bayesian methods, nor are we saying that Bayesian approaches are the only feasible way to include priors. We just see them as the most elegant way to include and evaluate alternative priors.

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, illustrating why a general audience should be interested in using Bayesian methods. We first show 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 then show how the Bayesian approach provides rich distributional information on the welfare impacts of individual choices by each subject, expanding the informational base for further policy interventions to improve individual welfare. And we show that a Bayesian approach lends itself naturally to real-time "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.[6]

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

[6] 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 [2022], 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.

that involves a negative CS.[7] In addition, this approach provides a quantification of the CS, whether gained or lost, from the observed choices.

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. We assume ε_{it} follows a logistic distribution and is independently and identically distributed across individuals and decisions. Assume that subject i selects the left lottery in lottery pair t whenever the latent index y_{it}^{*} is greater or equal to 0:

$$\text{Prob}\,(y_{it} = 1) = \text{Prob}(\nabla EU_{it}(r_i, \mu_i) \geq 0\,) = \Lambda(\nabla EU_{it}(r_i, \mu_i)) , \tag{3}$$

[7] 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].

where $\Lambda(.)$ is the logistic function (the inverse of the logit 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:

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

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]:

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

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

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

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.

Collect subject i's decisions in $y_i = (y_{i1}, \ldots, y_{i80})$. The likelihood of observing y_i is then

$$p(y_i \,|\, r_i, \eta_i, \varphi_i, \mu_i) = \prod_{t=1,80} [\, \Lambda(\nabla\text{RDU}_{it}(r_i, \eta_i, \varphi_i, \mu_i)) \,]^{y_{it}} [\, 1 - \Lambda(\nabla\text{RDU}_{it}(r_i, \eta_i, \varphi_i, \mu_i)) \,]^{(1-y_{it})} \quad (8)$$

The likelihood of observed all of the decisions of all subjects, $y = (y_1, ..., y_{111})$ is then

$$p(y \mid r, \eta, \varphi, \mu) = \prod_{i=1,111} \prod_{t=1,80} [\, \Lambda(\nabla RDU_{it}(r_i, \eta_i, \varphi_i, \mu_i)) \,]^{y_{it}} [\, 1 - \Lambda(\nabla RDU_{it}(r_i, \eta_i, \varphi_i, \mu_i)) \,]^{(1-y_{it})} \quad (9)$$

where $r = (r_1, ..., r_{111})$, $\eta = (\eta_1, ..., \eta_{111})$, $\varphi = (\varphi_1, ..., \varphi_{111})$ and $\mu = (\mu_1, ..., \mu_{111})$. Since EUT is nested in RDU, (8) and (9) also define the likelihood for the EUT model.

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."[8]

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

- 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 $8 + (4 \times N)$ 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.[9] There is a well-understood tradeoff between the two

[8] Appendix A documents the template used for our Bayesian estimation of risk preferences.. Appendix B provides details of convergence diagnostics for the core model of Section 1.

[9] 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

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*:

$$r_i \sim N(m_r, \sigma^2_r), \tag{10}$$

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

$$m_r \sim N(0, 100), \tag{11}$$

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

$$\sigma^2_r \sim IG(0.001, 0.001). \tag{12}$$

The essential idea is that there is an informative, non-diffuse prior specified in (10), 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 (11) and (12). We can restate (10) in conditional form as

$$r_i \mid m_r, \sigma^2_r \sim N(m_r, \sigma^2_r), \tag{10′}$$

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, (10′) reminds us that it is as if one Bayesian model was estimated for the pooled data just assuming the diffuse priors (11) and (12), 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.

(Yaari [1987]).

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*

$$\ln(\eta_i) \sim N(m_{\ln\eta}, \sigma^2_{\ln\eta}) \tag{13}$$

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

$$m_{\ln\eta} \sim N(0, 100), \tag{14}$$

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

$$\sigma^2_{\ln\eta} \sim IG(0.001, 0.001). \tag{15}$$

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

$$\ln(\varphi_i) \sim N(m_{\ln\varphi}, \sigma^2_{\ln\varphi}) \tag{16}$$

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

$$m_{\ln\varphi} \sim N(0, 100), \tag{17}$$

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

$$\sigma^2_{\ln\varphi\mu} \sim IG(0.001, 0.001). \tag{18}$$

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

$$\ln(\mu_i) \sim N(m_{\ln\mu}, \sigma^2_{\ln\mu}) \tag{19}$$

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

$$m_{\ln\mu} \sim N(0, 100), \tag{20}$$

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

$$\sigma^2_{\ln\mu} \sim IG(0.001, 0.001). \tag{21}$$

We assume that r_i, η_i, φ_i and μ_i are independently distributed.

In effect, all that these priors are saying is that we let the pooled sample data determine the posterior distribution for the representative agent, 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.

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 [2015; 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].

[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.

There are alternative shrinkage priors to the ones we use. Some aspects of our choices are guided by theoretical considerations. For example, η_i, φ_i and μ_i should be positive according to theory, so we choose the log-normal prior to implement these constraints. This is not a completely diffuse prior, and is referred to by Bayesians as a "weakly informative" prior to reflect the fact that the only prior is the "weak" qualitative prior of being positive. The choice of the shrinkage prior can affect the posterior distributions of the parameters of each individual to some extent. However, when the priors are diffuse enough, the effects might not be expected to be great. Consider the prior distribution of r_i as an example. We choose a symmetric normal distribution as the shrinkage prior for this parameter, which is estimated to have a standard deviation of 0.35. When we look at the distribution of the posterior *means* of r_i, that distribution is asymmetric, indicating that the individual choice data are informative enough to pull the individual estimates away from the symmetric normal prior distribution.

Because of the shrinkage effect of the priors, the hierarchical Bayesian approach should be used with caution when such effects are inappropriate. Although our data are informative enough to allow r_i to break away from the symmetric prior, we can imagine cases where it could fail to do so when the individual choice data are too noisy or the sample is too small. Another example of this potential sensitivity arises if the actual parameters have a multi-modal distribution, but we assume a singled peaked prior. This choice of priors could introduce bias into the posterior estimates through the prior as well. In these cases, one can still use the hierarchical approach, but might choose a bimodal shrinkage prior and introduce latent types with a mixture model, requiring a categorical prior over those types.[11]

[11] Harrison and Rutström [2019] apply the idea of mixture models to the estimation of risk preferences for pooled data, and Kruschke [2015; §10.2] illustrates how one can use categorical priors in a Bayesian hierarchical model to include latent types.

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 choices chosen carefully to allow estimates of models of risk preferences such as these, standard numerical methods can simply fail to converge.[12] 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[13] 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

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

[13] 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.

magic, for the 31 = 9 + 22 subjects abandoned by ML. There could be two reasons for a failure to obtain ML estimates for these 31 subjects, and Bayesian estimation solves these in different ways. The first cause could be that these subjects' choices are actually informative, but ML fails to converge due to computational failures. 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 likelihood function evaluations in order to derive the posterior distribution. The second cause could be that the subjects' choices are not informative at all, and the likelihood function is globally flat. Of course, if this is the case then the posterior will just be a replica of the prior in the Bayesian approach, but 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, we find that the posterior point estimates of the parameters for most of these subjects are not at all close to the mean of the shrinkage prior, indicating that the choice data contain enough information, such that the estimates are not just replicas of the pooled prior: 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.[14] Of course this

[14] The sample size is 10,000 for the MCMC sample and 2,500 for the burn-in sample, and we do not use thinning. For convergence criteria we mainly reviewed trace plots and autocorrelations of the sample with different lags of iteration number. We generally find excellent mixing and low autocorrelations as the lag

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 *modes* of the posterior distributions for the representative agent.[15] 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.

A more complete comparison of ML and Bayesian estimates should take into account confidence intervals of the former and full distributions of the latter. Using 95% confidence intervals, we find that only 8.9% of the Bayesian posterior distribution for all parameters of the representative agent overlaps with the 95% confidence intervals of the ML estimates. The "cuplrit" here is the estimate of r.[16]

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," the mode, but in this instance it is a full posterior distribution *for*

---

increases. In addition, we also compared the kernel density of each parameter when we use all, the first half and the second half of the posterior sample for that parameter, and find consistent distributions.

[15] In Bayesian analysis it is common to report the mean of the posterior distribution, or occasionally the median. The mode is appropriate in this specific case since it the most directly comparable statistic of the posterior distribution to the classical maximum likelihood estimate.

[16] The ML confidence intervals for r, η and φ, respectively, are (0.49, 0.78), (0.63, 1.00) and (1.14, 1.31). The 95% Bayesian highest posterior density intervals for these parameters are (0.31, 0.53), (0.83, 1.05) and (1.08, 1.31). The Bayesian posterior overlaps with 10.3%, 90.0% and 80.4% of the corresponding ML confidence interval for these individual parameters. For η and φ jointly, the Bayesian posterior overlaps with 72.9% of the corresponding ML confidence intervals.

each and every individual. So Figure 1 is based on one mode for the parameter r, for instance, whereas the left panel of Figure 2 is a histogram constructed from 111 modes of the parameter r_i for each subject i. 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 hierarchical 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.[17] 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.

Figure 3 undertakes this comparison of individual ML and Bayesian estimates more systematically, by summarizing the percent of the Bayesian distribution that is defined by the 95%

[17] Again, this is an average of the 111 modes reflected in these histograms.

ML confidence interval for each of the 80 subjects with ML estimates.[18] These displays are the individual-level counterpart of the percentages reported in relation to Figure 1 for the representative agent estimates. For the individual estimates, the probability weighting parameters are now the "cuplrit" leading to differences.[19] It turns out that the interaction of the three risk parameters is what leads to lower similarity in the joint comparisons displayed in panel D of Figure 3. This is particularly true for the interaction of η and φ.

Two individual examples demonstrate the contrasts between ML and Bayesian estimates. Figure 4 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 4 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 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

---

[18] Imbens [2021] provides an exposition of the general value in economics of using Bayesian approaches to assess the "statistical significance" of inferences.

[19] In fact, the distribution of 80 percentages in panels B and C overstate the similarity of ML and Bayesian estimates: for those subjects that are well characterized under EUT, the estimates of η and φ should be roughly the same.

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 6 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. 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. As explained above, we have 80 subjects with ML estimates, and 111 subjects with Bayesian estimates.[20]

[20] 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.

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 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 allows us to see that the differences in CS shown in Panel A of Figure 7 do indeed amount to significant differences. The percent of the ML 95% confidence intervals on these CS estimates that are within the Bayesian posterior is very low, averaging less than 50%.[21]

Figure 9 shows a scatter plot of Efficiency outcomes to allow a literal "head to head"

[21] The same type of comparison does not apply for the Efficiency measures, since these are categorical (0% and 100%) posterior distributions at each observed choice, and real-valued for the ML estimates. The evidence for CS is sufficient to make the general case for significantly different welfare evaluations with the Bayesian approach.

comparison of the effects of using Bayesian estimates rather than ML estimates.[22] 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). For the 6 outliers with very high/low efficiency evaluated with ML estimates, but low/high efficiencies under Bayesian approach, we find that the ML estimates of their parameters are all much more extreme than the Bayesian estimates. For instance, for the 3 outliers in the top left corner, the estimates of r_i are in the range of −3 to −4.5 with the ML approach, in the range of −0.1 to − 0.4 with the Bayesian appraoch.

We make no formal inferences about the effects of using Bayesian estimates instead of ML estimates on *average* CS or average Efficiency. We could, from inspection of Figures 7 and 9, 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. We extend our analysis and explore these distributions further in §3.B.

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. Although some extensions discussed here can possibly be done with the ML approach, the Bayesian approach allows us to do so naturally and more elegantly. Here we focus on discussing how these extensions can be done easily using the Bayesian approach, rather than

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

pitching the outcomes of these extensions under the two approaches against each other.

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 relevant to the target choice for normative evaluation.[23] 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 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

[23] 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.

preference elicitation.[24] Hence time is a critical 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 10 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). The risk preference parameters are, again, based on the posterior means of each parameter for each individual,[25] while Efficiency is calculated using the full posterior distribution of the risk preference parameters (we defer a detailed discussion on the implications of this change of method to Section 3.B). Remarkably, the Pearson correlation ϱ for Efficiency, the target or normative evaluation, is 0.79 in this instance.[26] 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.

---

[24] 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, intertemporal risk aversion, and possibly even social preferences.

[25] Earlier we used the posterior mode because it is a more comparable measure to ML estimates. However, here we have shifted our focus away from ML, and aim to compare Bayesian estimates with different sample size, so we use the more common measure of posterior mean. Alternatively, one could also use the median of the posterior distributions.

[26] We also always report the Kendall rank correlation τ. The *rank* correlation is robust to the effect of outliers, but of course uses less information. In general the results for ϱ and τ are qualitatively consistent, with $\tau < \varrho$ in all cases.

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 11 provides a comparable display to Figure 10, but with samples of 40 instead of 20. Now we achieve a Pearson 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 Pearson correlation with Efficiency of 0.97, and 0.97, 0.90 and 0.94 for the r, η or φ risk preference parameters, respectively.[27]

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.[28] 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.[29] Again, field settings are natural here: one might have a large-scale survey

---

[27] The corresponding rank correlations are 0.87, 0.86, 0.74 and 0.80, respectively.

[28] An alternative approach is to exploit the ability to sequentially update based on *data solely obtained from one individual*, by explicitly designing the "next" experimental task in a Bayesian (or classical) manner. Examples applied to eliciting risk preferences include Cavagnaro et al. (2013a)(2013b), Chapman et al. (2018), Ray et al. (2019) and Toubia et al. (2013). These methods place some "real-time" computational burden on the software generating the experimental interface, but these burdens are becoming minor as hardware and software improve. And one could certainly link these to Bayesian Hierarchical Models, providing informative priors for the individual prior to any dynamic optimization based on accumulating choices by the individual.

[29] 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, 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

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 in earlier analyses. Under the Bayesian approach, the correct inferences should take into account the fact that these are *full* posterior distributions.[30] 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).[31] 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*. We can then report that distribution as a kernel density, or select some measure of central tendency such as the mean or median.

---

impact Bayesian analyses.

[30] 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.

[31] 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 consider the mean of the posterior predictive distribution of Efficiency for each individual. Figure 12 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).[32] 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 [2021] 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. Recall that the order of insurance products, differentiated by their actuarial parameters, was randomly assigned to

[32] The left and middle panels of Figure 12 are the same as the top left panels of Figures 10 and 11. We replicate them here to draw attention on the effects of sample sizes on inference about welfare.

each subject.[33] Figure 13 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.[34] This probability is presented in Figure 13, 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 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

[33] 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.

[34] 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]).

purchase lottery by the subject should properly reflect her risk preferences.

Consider comparable evaluations for four individuals from our sample in Figure 14. 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 13.

The remaining subjects in Figure 14 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.[35] 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 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.[36]

---

[35] 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.

[36] The intertemporal risk aversion of a subject 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.

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. And the case for using a Bayesian approach in normative evaluation obviously extends beyond the (canonical) example of insurance purchase decisions, and beyond just the elicitation of atemporal risk preferences. Our emphasis has been on the utility of the Bayesian approach when it comes to *prediction*, as required by *normative* evaluations. There are already many excellent sources on the advantages of the Bayesian approach for *testing* hypotheses with various *descriptive* structural models in economics.

There are also more general reasons for wanting to adopt a Bayesian approach, to make explicit the role that priors have when making normative evaluations. Again, the point is not the use of priors *per se*, but the ability to incorporate them elegantly and transparently as a central part of the overall analysis.

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.[37] In general we need to be able to pool disparate sources of data, even observational

[37] 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]

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.[38]

A second general reason for a Bayesian approach is that researchers who do not regard themselves as Bayesian often do in fact use priors but are unaware of doing so, and end up including priors in an *ad hoc* manner that often makes it hard to see what inferences are being driven by the priors and what are being driven by the data. This is the entire point of Leamer [1978], so exhaustively examines a myriad of popular "specification search" procedures used by non-Bayesians, so as to examine what priors are being inadvertently used. A common theme is that when the implied prior is exposed by a Bayesian light, the specification search is not the best way to include that prior. So even if some non-Bayesian claims that they have no interest in priors or "related estimates," and hence have no need for Bayesian methods, the methods they *practice* suggest otherwise.

A final 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

---

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 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. Harrison [2021] provides a more extensive review of these issues from a Bayesian perspective, with implications for experimental design in economics.

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

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 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
Bayesian Estimates are the mode of the posterior distribution
ML 95% confidence intervals for r, η and φ are 8.9% of the Bayesian posterior

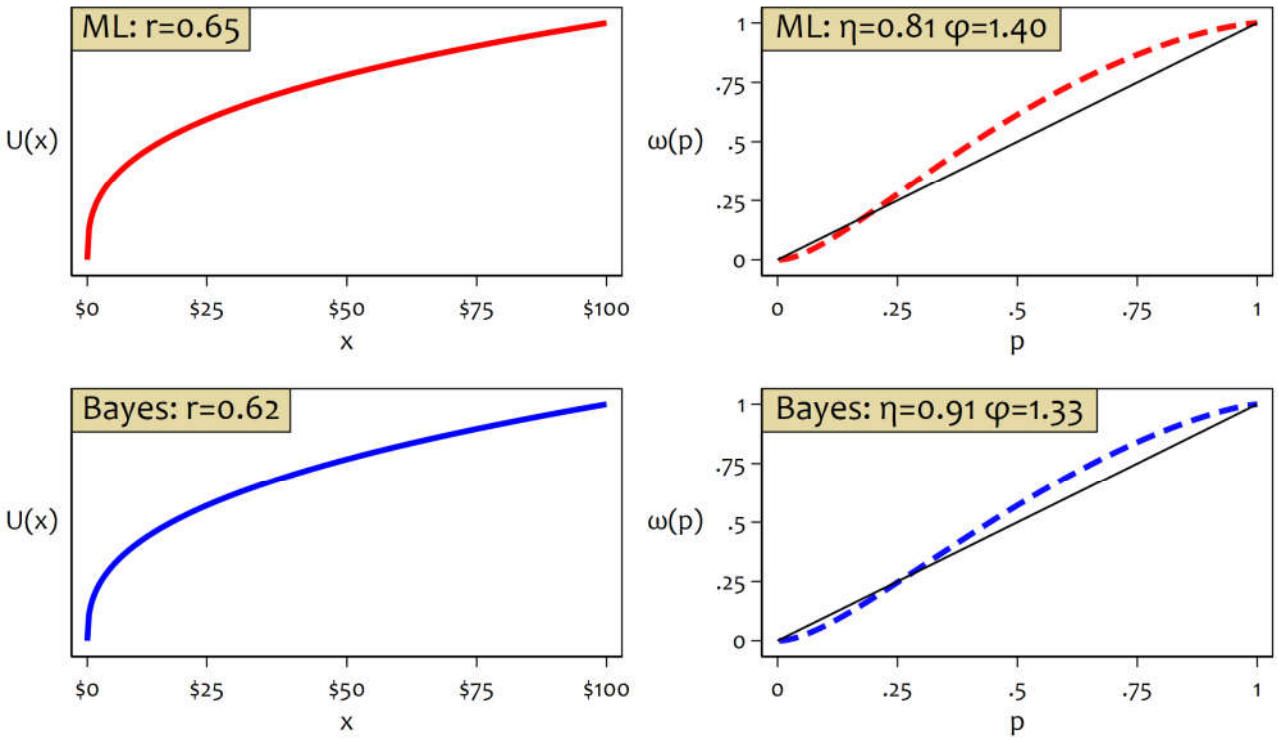


Figure 2: Distributions of Individual Risk Preference Parameters from Hierarchical Bayesian Model

Posterior modes for each of N=111 subjects
Dashed lines indicate posterior averages

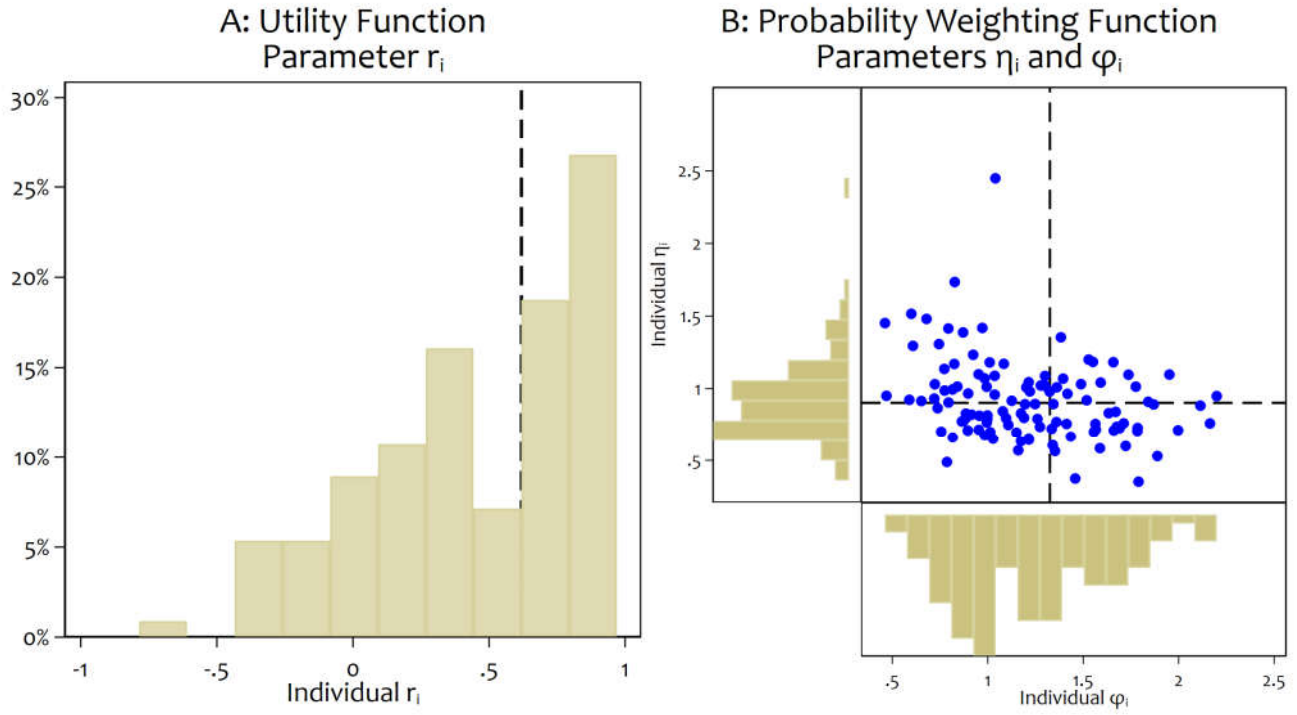


Figure 3: Comparisons of Individual Risk Preference Parameters from Maximum Likelihood and Bayesian Models

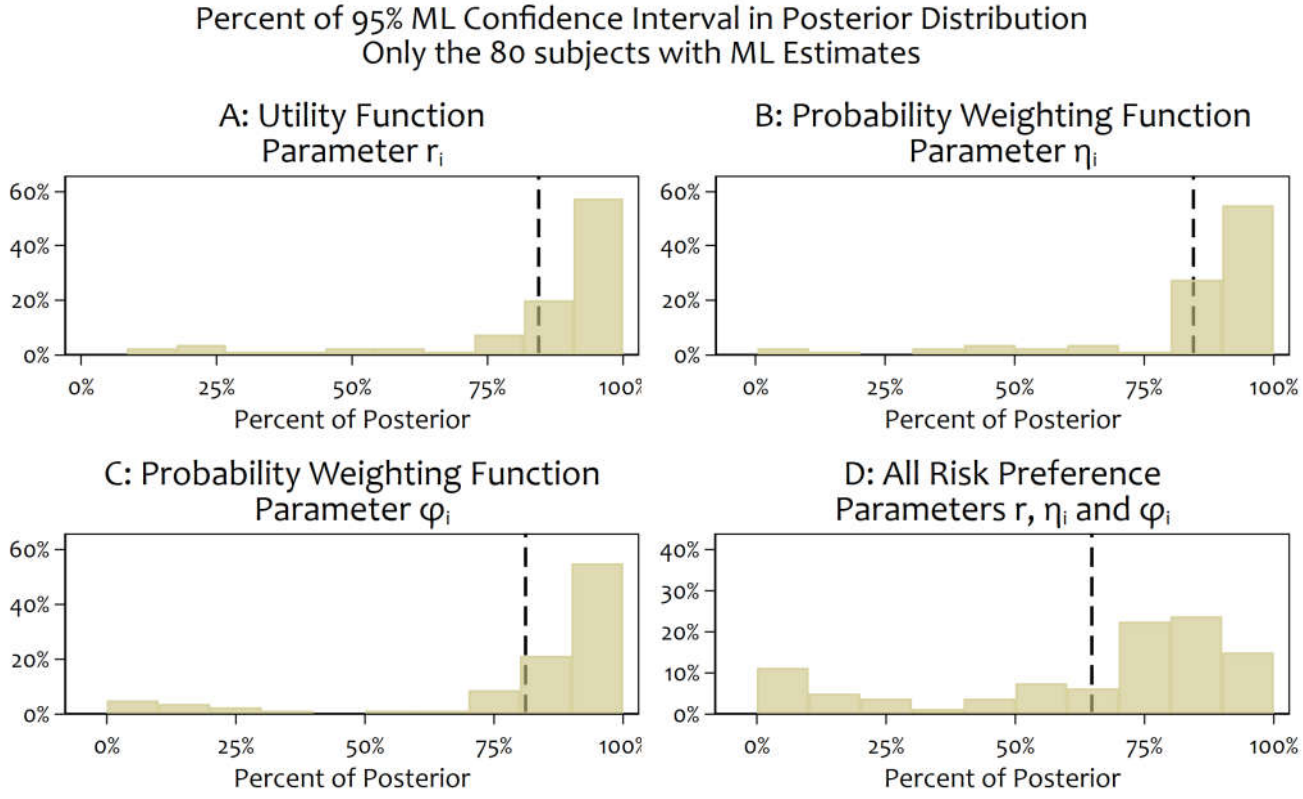


Figure 4: Risk Preferences for One Subject

Maximum Likelihood *versus* modal Bayesian Estimates for the same subject
ML 95% confidence intervals for r, η and φ are 5.1% of the Bayesian posterior

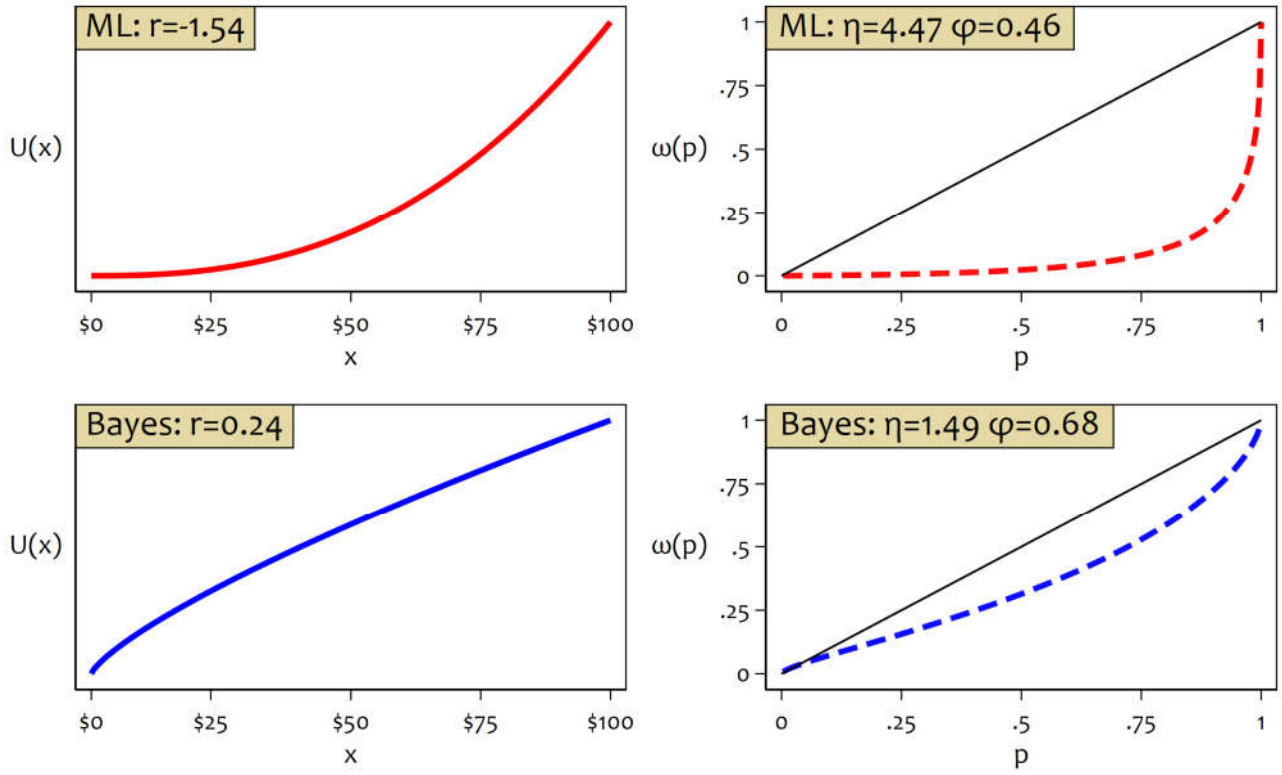


Figure 5: Risk Preferences for a Second Subject

Maximum Likelihood *versus* modal Bayesian Estimates for the same subject
ML 95% confidence intervals for r, η and φ are 53.5% of the Bayesian posterior

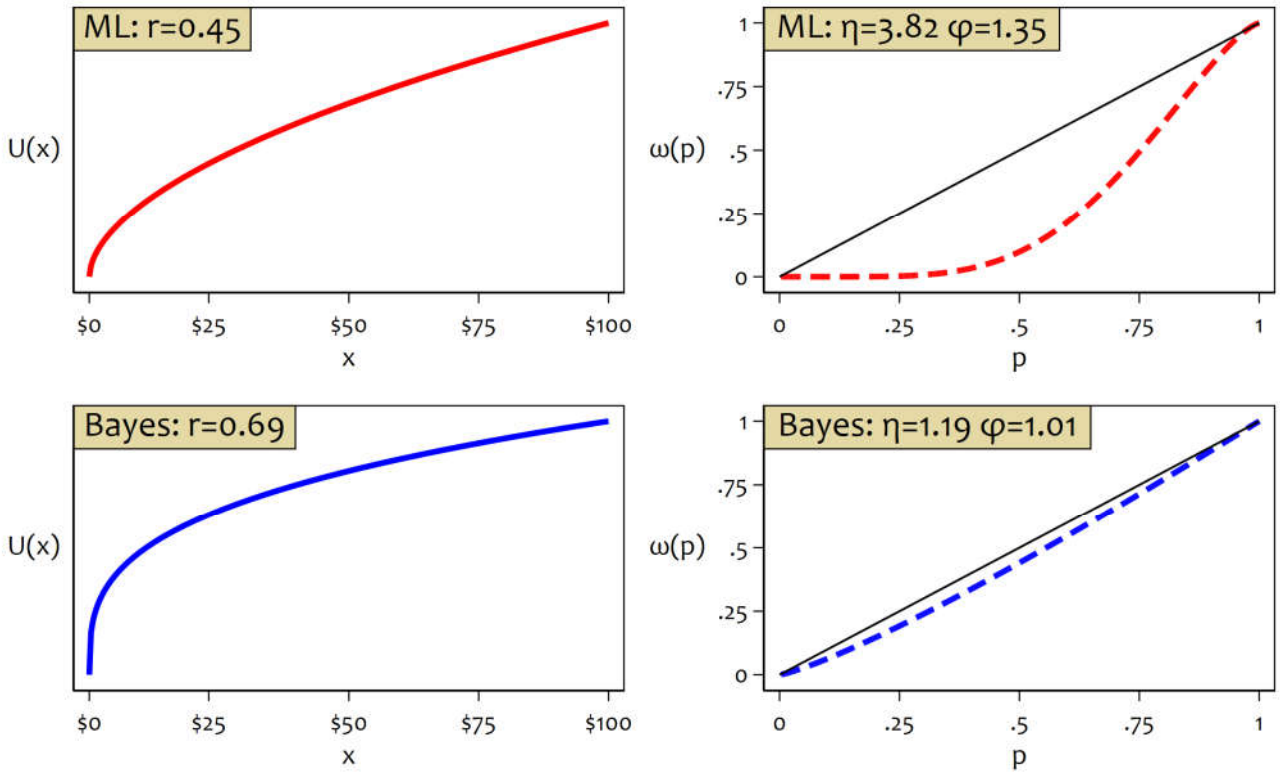


Figure 6: Comparison of ML and Bayesian Estimates of Risk Preference Parameters for Each Individual

Extreme ML estimates are excluded for r and η

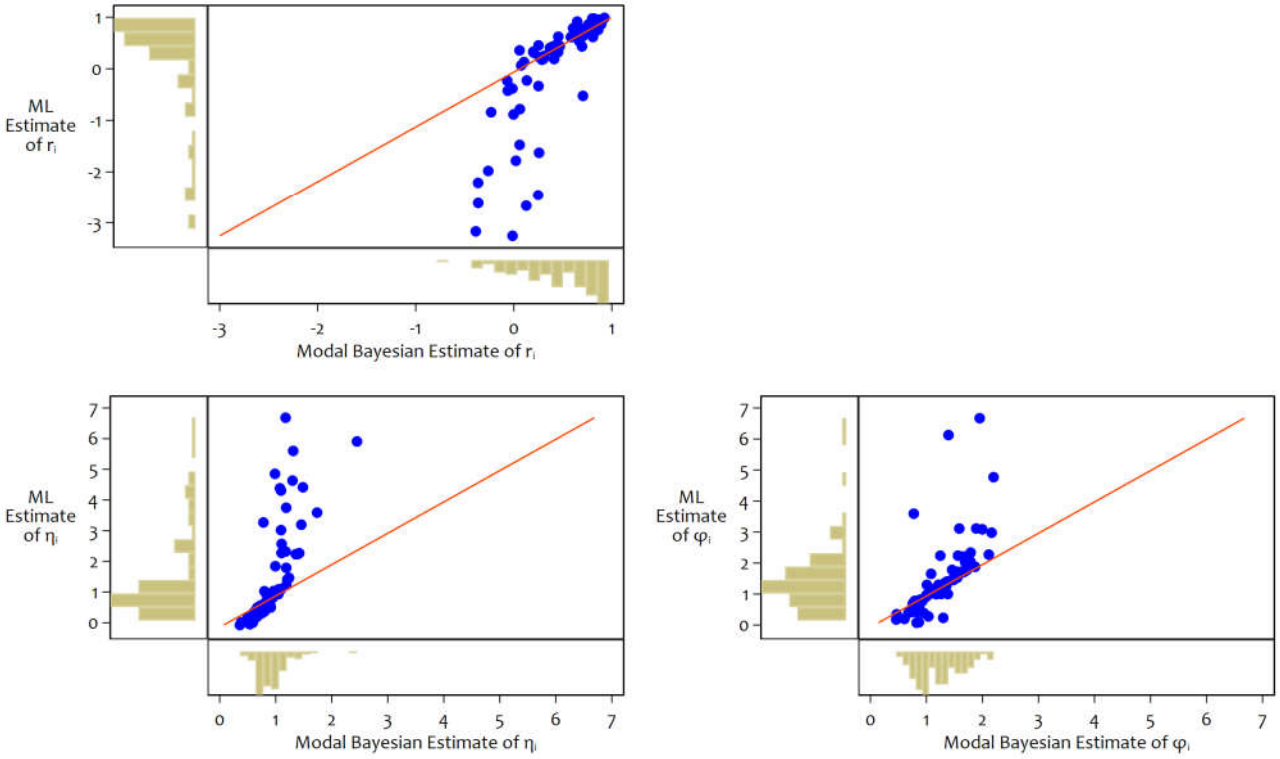


Figure 7: Effects of Inferences About Risk on Welfare

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

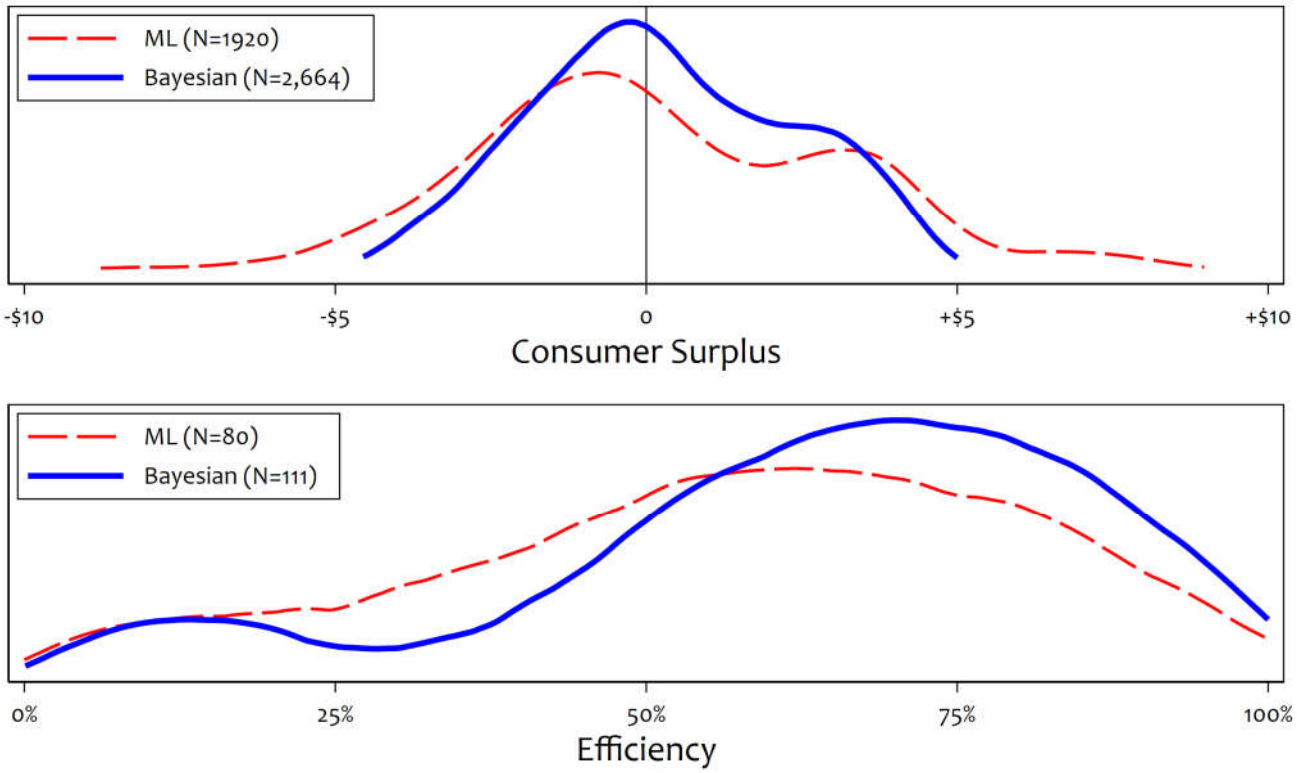


Figure 8: Comparison of Individual Expected Consumer Surplus Estimates from Maximum Likelihood and Bayesian Models

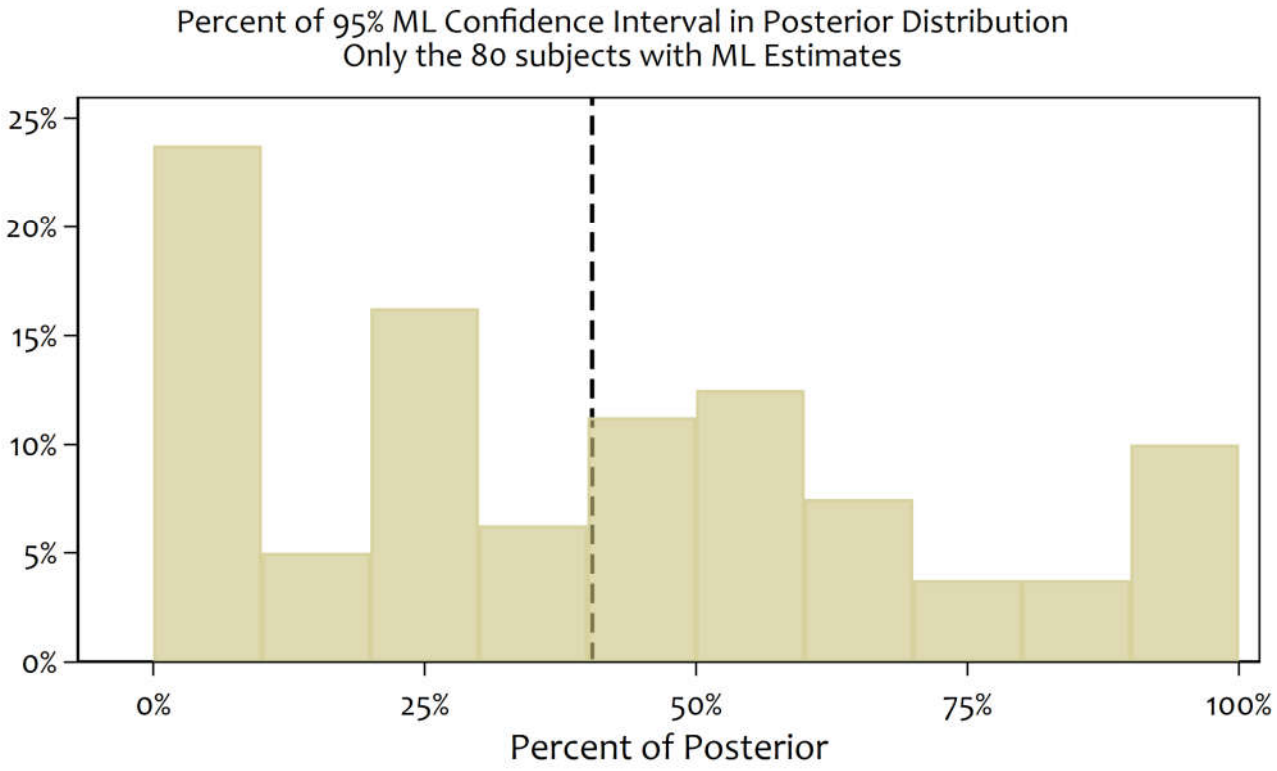


Figure 9: Scatter Plot of Effect of Inferences About Risk on Welfare

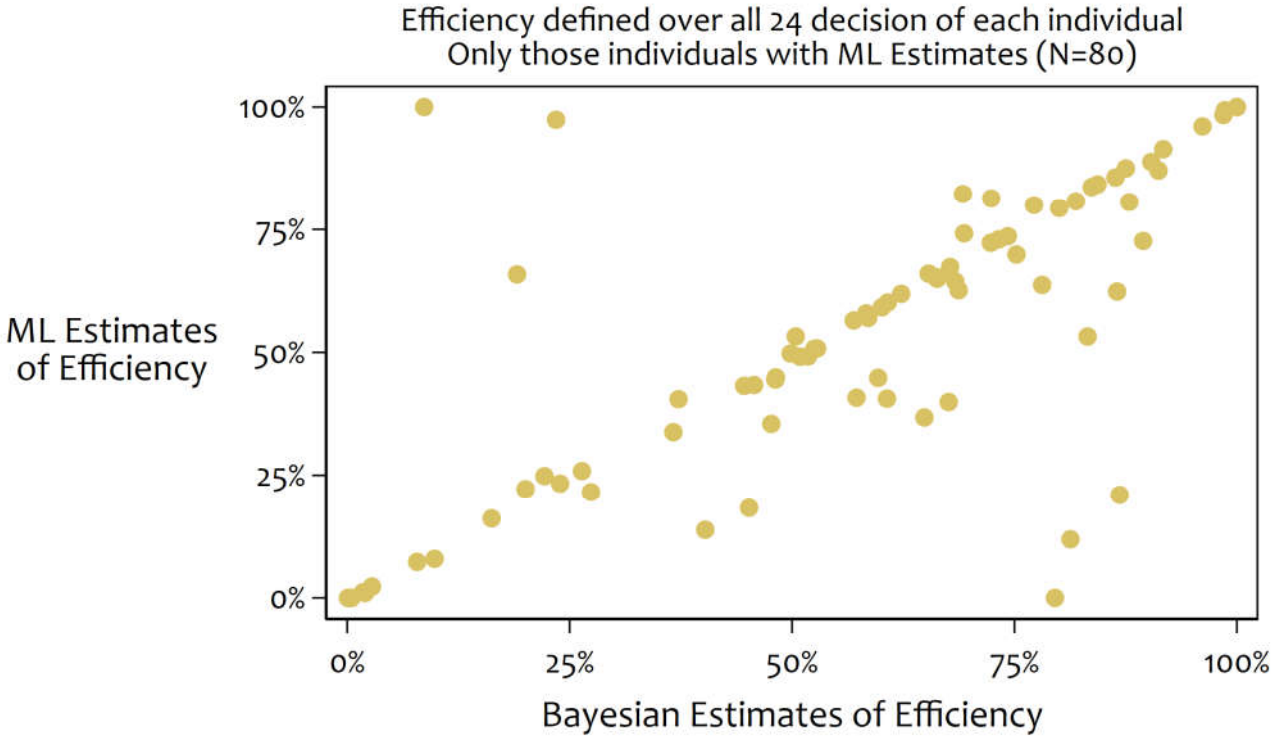


Figure 10: Comparison of Bayesian Hierarchical Estimates of Individual Risk Preference Parameters with Sample Sizes of 20 and 80 For Each Subject

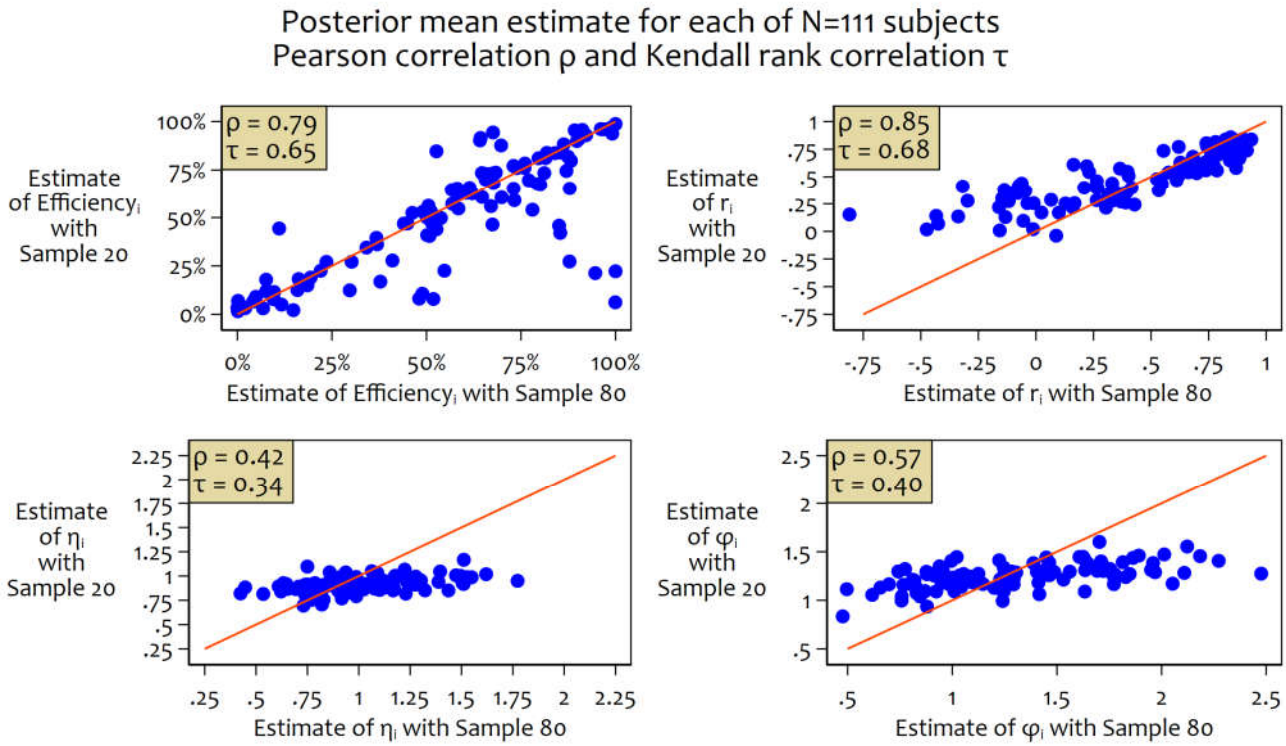


Figure 11: Comparison of Bayesian Hierarchical Estimates of Individual Risk Preference Parameters with Sample Sizes of 40 and 80 For Each Subject

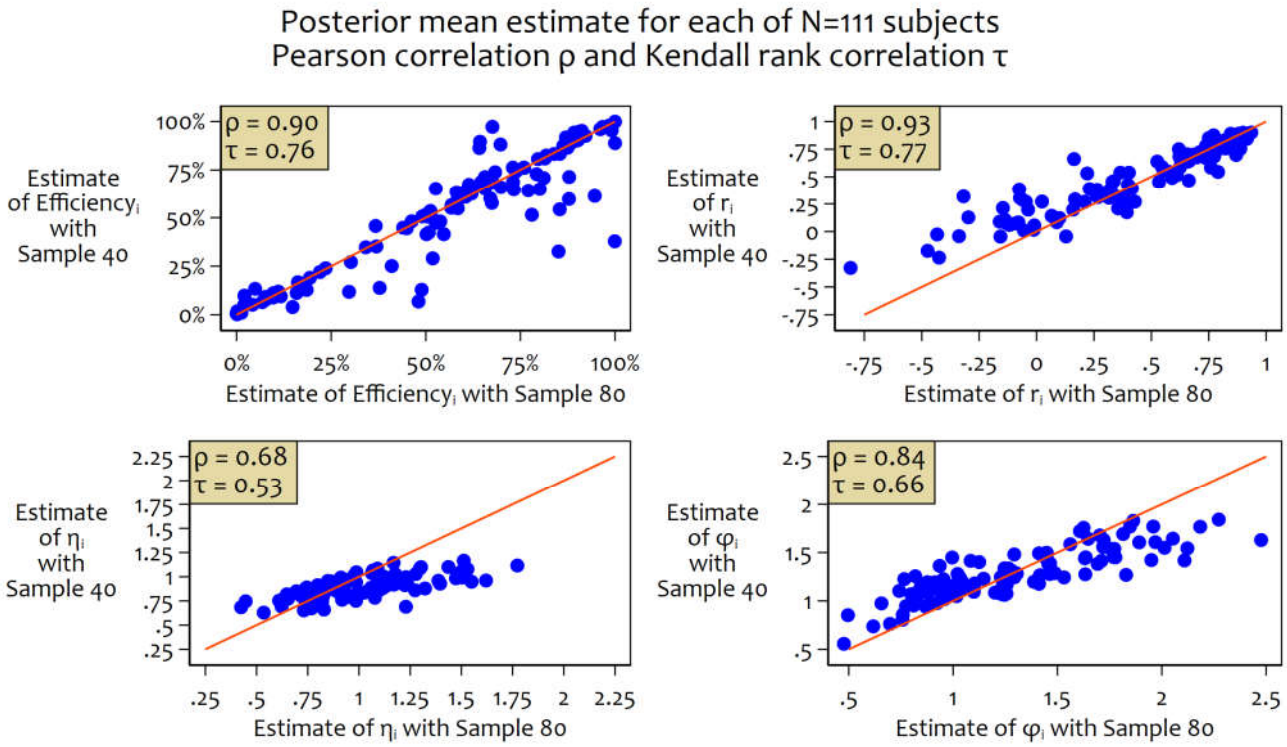


Figure 12: Comparison of Bayesian Posterior Predictive Estimates of Efficiency with Sample Sizes of 20, 40 or 60 For Each Subject

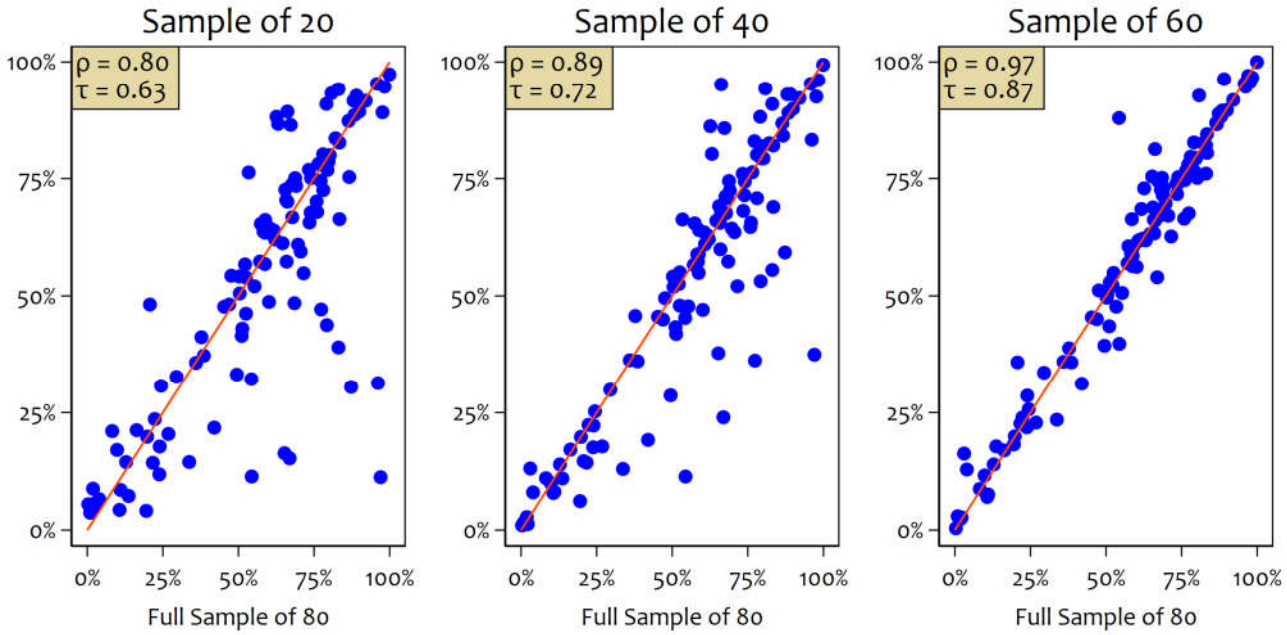


Figure 13: Adaptive Welfare Evaluations for Subject #1

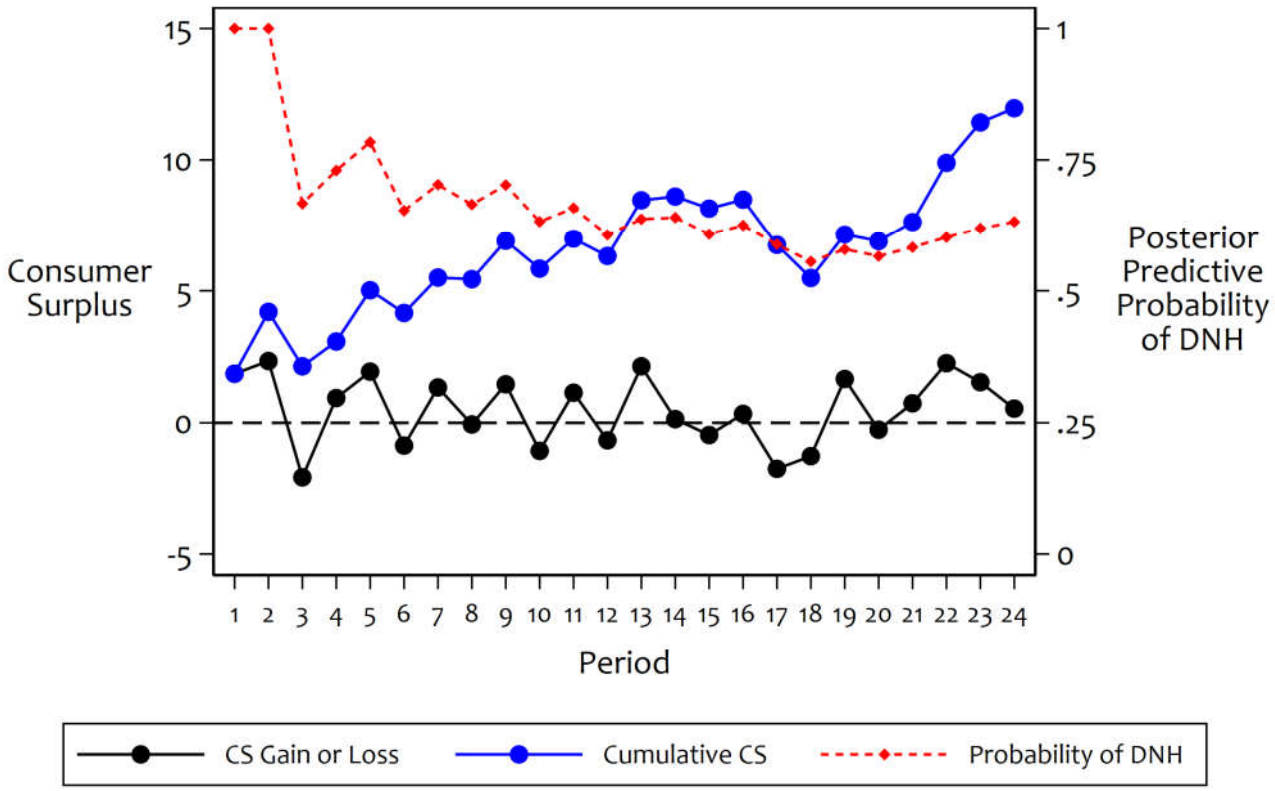
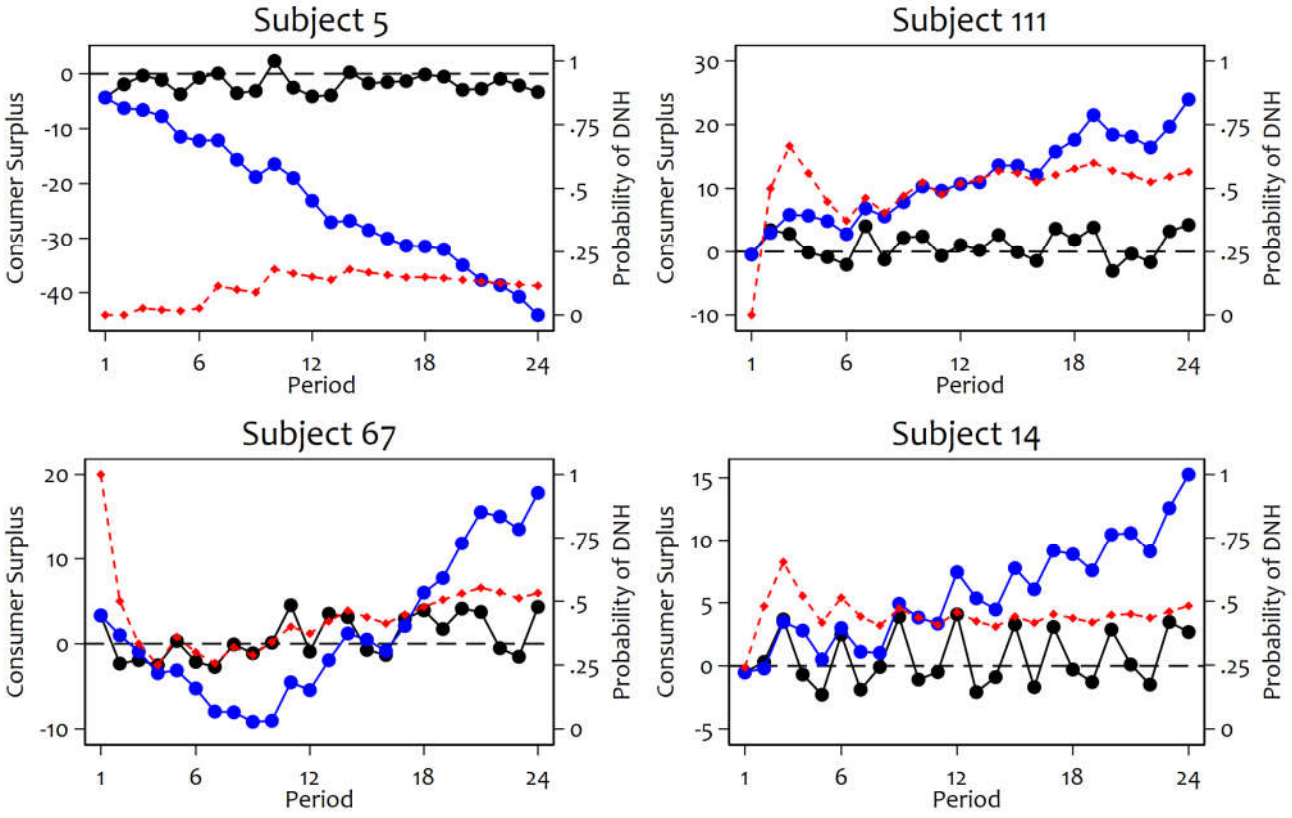


Figure 14: Individual Adaptive Welfare Evaluations for Four Subjects



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 Teytelboym, 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/

Cavagnaro, Daniel R.; Pitt, Mark A.; Gonzalez, Richard, and Myung, Jay I., "Optimal Decision Stimuli for Risky Choice Experiments: An Adaptive Approach," *Management Science*, 59(2), 2013a, 358-375.

Cavagnaro, Daniel R.; Pitt, Mark A.; Gonzalez, Richard, and Myung, Jay I., "Discriminating Among Probability Weighting Functions using Adaptive Design Optimization," *Journal of Risk and Uncertainty*, 47(3), 2013b, 255-289.

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

Chapman, Jonathan; Snowberg, Erik; Wang, Stephanie, and Camerer, Colin, "Loss Attitudes in the U.S. Population: Evidence from Dynamically Optimized Sequential Experimentation

(DOSE)," *NBER Working Paper 25072*, September 2018.

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," *CEAR Working Paper 2020-15*, 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., "Experimental Design and Bayesian Interpretation," in H. Kincaid and D. Ross (eds.), *Modern Guide to the Philosophy of Economics* (Cheltenham, UK: Elgar, 2021).

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).

Harrison, Glenn W., and Rutström, E. Elisabet, "Expected Utility *And* Prospect Theory: One Wedding and a Decent Funeral," *Experimental Economics*, 12(2), June 2009, 133-158.

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

Imbens, Guido W., "Statistical Significance, *p*-Values, and the Reporting of Uncertainty," *Journal of Economic Perspectives*, 35(3), 2021, 157–174.

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

Kasy, Maximilian, and Sautmann, Anja, "Adaptive Treatment Assignment in Experiments for Policy Choice," *Econometrica*, 89(1), 2021, 113-132.

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., *Specification 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, 2022 forthcoming).

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.

Ray, Debajyoti; Golovin, Daniel; Andreas Krause, Andrew, and Camerer, Colin, "Bayesian Rapid Optimal Adaptive Design (BROAD): Method and Application Distinguishing Models of Risky Choice," *OSF Preprints*, November 2019, doi:10.31219/osf.io/utvbz.

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).

Toubia, Olivier; Johnson, Eric; Evgeniou, Theodoros, and Delquié, Philippe, "Dynamic Experiments for Estimating Preferences: An Adaptive Method of Eliciting Time and Risk Parameters," *Management Science*, 59(3), 2013, 613-640

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.

Appendix A: Template Code (FOR ONLINE PUBLICATION)

We provide the codes for the estimation of the hierarchical Bayesian Rank Dependent Utility (RDU) model specified in §1.B and §1.C. We use a combination of the Metropolis Hastings algorithm and Gibbs sampler to obtain samples from the posterior distributions of the parameters, using the **bayesmh** Bayesian estimation procedures in *Stata*. These procedures requires only an input of the likelihood evaluators based on equations (4)-(9) as a user-defined function, and automatically applies the Gibbs sampler to parameters from the hierarchical structures in (10)-(21). We consider this close connection of programing syntax to the econometric formalization to be an advantage of the *Stata* package for novice users of Bayesian econometric methods, although it is no doubt shared in comparable software such as *R*, *Stan*, *JAGS* and *WinBUGS*.

For models that are larger than ours, particularly with respect to the number of subjects included, one would want to complement the Metropolis Hastings algorithm and Gibbs sampler currently available in *Stata* with the Hamiltonian MCMC methods available in *Stan*. To this end, we are developing *Stan* templates for the models estimated here with *Stata*, allowing calls to *Stan* from within *Stata*.[39]

A.1 Data and Variables

The data are saved in a *Stata* dataset with each row of observation recording all the observables of subject i in lottery pair t, including a subject ID, the prizes and probabilities of the lottery pair, and the subject's choice of left or right lottery from the pair. Each variable is defined as

[39] The "call to *Stan*" involves the real-time compilation of a C++ program, which can then be run as an executable program to provide results back to *Stata*. The fixed time costs of this compilation are quickly offset by the greater efficiency of the Hamiltonian MCMC methods in practical settings of 100 or more subjects. Excellent discussions of the connection between *Stata* and *Stan* by John Thompson can be found at https://staffblogs.le.ac.uk/bayeswithstata/2015/05/01/stan-with-stata-part-1-a-plan-of-action/ and https://staffblogs.le.ac.uk/bayeswithstata/2015/07/03/stan-vs-openbugs-controlled-from-stata/.

follows:

- **probkL**: the probabilities of the left lottery, where k = 1, . . . , 4;
- **prizekL**; the prizes of the left lottery, where k = 1, . . . , 4;
- **probkR**: the probabilities of the right lottery, where k = 1, . . . , 4;
- **prizekR**: the prizes of the right lottery, where k = 1, . . . , 4;
- **endowment**: monetary endowment the subject receives;
- **sid**: sequentially coded subject ID; and
- **choiceL**: choice of the subject in lottery pairs, equal to 1 if the left lottery is chosen.

We collect relevant variables in a global **Rdata** to allow for a more succinct presentation of the main syntax:

```
global Rdata "prob1L prob2L prob3L prob4L prob1R prob2R prob3R prob4R prize1L,
      prize2L, prize3L prize4L prize1R prize2R  prize3R prize4R endowment"
```

We use a "hanging indent" to flag that this is a single line of code in *Stata*. Text files that show the code in raw form make this clear, and are available on request.

A2. The RDU Model in *Stata*

The RDU model specified in equations (4)-(9) needs to be written in a user-defined function referred to as "user-defined likelihood evaluator" in *Stata*. Akin to the evaluator functions written for classical ML estimation in *Stata*, it is essentially a program that takes the data of the choice and assumed values of the parameters, and returns the likelihood of the observed choice. The "assumed" values of the parameters are generated by algorithms. In classical MLE these are typically gradient-based algorithms.

To allow for flexibility in the specifications, we use several globals: the **utype** global specifies the specific form of the CRRA utility function in (6), the **contextual** global specifies whether to use ν_{it} to normalize utilities in (7), and the **cdf** global specifies whether to use a Probit or Logit link

between the latent index and the observed choice in (3). To replicate the hierarchical RDU model in its exact form in the main text, we define these globals as follows:

```
global utype "1-r"
global contextual "y"
global cdf "normal"
```

The user-defined likelihood evaluator is then:

```
program probitRDUprelecLN

	args lnf r LNeta LNphi LNmu
	tokenize $MH_extravars
	local h = 0
	foreach par in prob prize {
		forvalues i=1/4 {
			local h = `h'+1
			local `par'`i'L ``h''
		}
		forvalues i=1/4 {
			local h = `h'+1
			local `par'`i'R ``h''
		}
	}
	local h = `h'+1
	local endowment ``h''
	tempvar lnfj
	tempvar euL euR eudiff m1L m2L m3L m4L m1R m2R m3R m4R u1L  u2L u3L u4L u1R u2R u3R u4R
	tempvar a_prob1L a_prob2L a_prob3L a_prob4L a_prob1R  a_prob2R a_prob3R a_prob4R
	tempvar pw_prob1L pw_prob2L pw_prob3L pw_prob4L pw_prob1R  pw_prob2R pw_prob3R pw_prob4R
	tempvar dw_prob1L dw_prob2L dw_prob3L dw_prob4L dw_prob1R  dw_prob2R dw_prob3R dw_prob4R
	tempvar eta phi mu
	tempvar low high

	quietly {

	* transform parameters
		generate double `phi'    = exp(`LNphi')
		generate double `eta'    = exp(`LNeta')
		generate double `mu'    = exp(`LNmu')
	* add in endowments
		foreach x in L R {
			forvalues i=1/4 {
				generate double `m`i'`x'' = `endowment' + `prize`i'`x''
			}
		}
	* generate the utility function
	foreach x in L R {
		forvalues i=1/4 {
			if "$utype" == "2" {
				generate double `u`i'`x'' = (`m`i'`x''^(1- `r'))/(1-`r')
			}
			else {
				generate double `u`i'`x'' = `m`i'`x''^`r'
			}
		}
	}
	* generate the decumulative probabilities for each lottery
	foreach x in L R {
		generate double `a_prob4`x'' = `prob4`x''
		generate double `a_prob3`x'' = `prob3`x'' +  `a_prob4`x''
		generate double `a_prob2`x'' = `prob2`x'' +  `a_prob3`x''
		generate double `a_prob1`x'' = `prob1`x'' +  `a_prob2`x''
```

```
            * generate the weighted probabilities
            forvalues i=1/4 {
                  generate double `pw_prob`i''`x'' = `a_prob`i''`x''
                  replace `pw_prob`i''`x'' = exp((-`eta')*(-  ln(`a_prob`i''`x''))^`phi') if
                        `a_prob`i''`x''>0 &  `a_prob`i''`x''<1
            }
            * Generate the decision weights
            generate double `dw_prob4`x'' = `pw_prob4`x''
            generate double `dw_prob3`x'' = `pw_prob3`x''- `pw_prob4`x''
            generate double `dw_prob2`x'' = `pw_prob2`x''- `pw_prob3`x''
            generate double `dw_prob1`x'' = `pw_prob1`x''- `pw_prob2`x''
      }
      * evaluate the RDU of each lottery (called "eu" here due to sloth)
      generate double `euL' = 0
      generate double `euR' = 0
      foreach x in L R {
            forvalues i=1/4 {
                  replace `eu`x'' = `eu`x'' + `dw_prob`i''`x''*`u`i''`x''
            }
      }
      * get the Fechner index
      if "$contextual" == "y" {
            generate double `low' = `u1L'
            generate double `high' = `u1L'
            forvalues i=1/4 {
            foreach s in L R {
                  replace `low' = `u`i''`s'' if `u`i''`s'' < `low' &  `prob`i''`s'' > 0
                  replace `high' = `u`i''`s'' if `u`i''`s'' > `high'  & `prob`i''`s'' > 0
            }
            }
            generate double `eudiff' = ((`euL' - `euR')/(`high'-`low'))/`mu'
      }
      else {
            generate double `eudiff' = (`euL' - `euR')/`mu'
      }
      * construct the likelihood contribution
      generate double `lnfj' = ln($cdf( `eudiff')) if $MH_y1 == 1  & $MH_touse
      replace `lnfj' = ln($cdf(-`eudiff')) if $MH_y1 == 0 &  $MH_touse
      summarize `lnfj', meanonly

      * end of "qui" block
      }
      * check that the required evaluations are done
      if r(N) < $MH_n {
            scalar `lnf' = .
            Exit
      }
      scalar `lnf' = r(sum)

end
```

It is relatively easy to see the "economics" at work in this syntax, as stressed for comparable ML evaluators by Harrison and Rutström [2008; Appendix E]. It is also relatively easy to see how to adapt this template for simpler models, such as EUT, or more complex models.

A.3 Main Syntax

The main *Stata* command for the application of the Metropolis-Hastings algorithm, and that allows for the user-defined likelihood function, is **bayesmh**. The comparable command for classical ML estimation in *Stata* is **ml**. The template for **bayesmh** is provided below. In this command we first define all the individual parameters r_i, η_i, φ_i, μ_i in line 1. In line 1 we also call in the user-defined likelihood evaluator introduced in A.2 with option **llevaluator( )**, which tells *Stata* how the likelihood are evaluated at given values of these parameters. In addition, we also tell *Stata* to input the variables in **extravars( )**, which are the variables saved in a *Stata* data file and introduced in A.1. The command parses these variables into temporary variables when evaluating the likelihood.

In lines 2 through 13 we specify the hyper distributions in equations (10)-(21) of the main text and ask *Stata* to use the Gibbs sampler for hyperparameters m_r, σ_r^2, m_lnη, $\sigma_{ln\eta}^2$, $m_{ln\varphi}$, $\sigma_{ln\varphi}^2$, $m_{ln\mu}$ and $\sigma_{ln\mu}^2$. In lines 14 through 19 we specify options for the size of the MCMC and burn-in samples, display of progress, saving the MCMC samples for later use, adaptation parameters for the adjustment of proposal steps in the MH algorithm, initial values, and so on. Further documentation is provided in the *Stata* manual for the **bayesmh** command: StatCorp [2019; p. 112-275]. The line numbers on the left are solely for exposition, and not used in the actual *Stata* command line.

```
1 bayesmh (r: choiceL i.sid) (eta:choiceL i.sid) (phi:choiceL  i.sid) (mu: choiceL i.sid),
      llevaluator(probitRDUprelecLN, extravars($Rdata)) ///
2     prior({r:i.sid}, normal({rMean:constant},{rVar})) block({r:i.sid}, split) ///
3     prior({rMean:constant}, normal(0, 100)) block({rMean:constant},gibbs) ///
4     prior({rVar}, igamma(0.001, 0.001))block({rVar},gibbs) ///
5      prior({eta:i.sid}, normal({etaMean:constant},{etaVar})) block({eta:i.sid}, split) ///
6     prior({etaMean:constant}, normal(0, 100)) block({etaMean:constant},gibbs) ///
7     prior({etaVar}, igamma(0.001, 0.001)) block({etaVar},gibbs) ///
8     prior({phi:i.sid}, normal({phiMean:constant},{phiVar})) block({phi:i.sid}, split) ///
9     prior({phiMean:constant}, normal(0, 100)) block({phiMean:constant},gibbs) ///
10    prior({phiVar}, igamma(0.001, 0.001)) block({phiVar},gibbs) ///
11    prior({mu:i.sid}, normal({muMean:constant},{muVar})) block({mu:i.sid},split)  ///
12    prior({muMean:constant}, normal(0, 100)) block({muMean:constant},gibbs) ///
13    prior({muVar}, igamma(0.001, 0.001)) block({muVar},gibbs) ///
14    rseed(54321) mcmcs(10000) burnin(2500) ///
15    adapt(every(10) alpha(0.75) beta(0.8) gamma(0.0001) maxiter(1250)) ///
16    nomleinitial nocons initial({r:i.sid} 0 {eta:i.sid} 0  {phi:i.sid} 0 ///
17    {mu:i.sid} 0 {rMean:constant} {etaMean:constant} {phiMean:constant} ///
18    {muMean:constant} 0 {rVar} {etaVar} {phiVar} {muVar} 1) initsummary blocksummary ///
19    saving(estimates/Choice_SpecrduPR, replace)  dots(1,every(10)) notable
```

Additional References

StataCorp, *Stata Bayesian Analysis Reference Manual: Release 16* (College Station, TX: Stata Corporation, 2019).

Appendix B: Convergence Diagnostics (FOR ONLINE PUBLICATION)

We provide diagnostic analysis on the convergence of models parameters for the hierarchical RDU model presented in Section 1. Figures B1 through B8 presents diagnostic graphs for the population parameters from equations (10), (13), (16) and (19). Figures B9 though B12 present diagnostic graphs for the model parameters r_i, η_i, φ_i and μ_i of four individual subjects from the likelihood function in equation (8). We have a total of 111 subjects, so we present diagnostic graphs for selected subjects here as a representation, to save space. However, we checked convergence for all subjects' parameters, and observed similar convergence quality for all other subjects. Full results are provided in the online documentation of our data and code.

Focus initially on Figure B1 through B8 for the population parameters of the model. For each parameter there are four plots. In the top left quadrant is the **trace plot**, showing MCMC iteration number on the horizontal axis and simulated values for the parameter on the vertical axis. If the display exhibits roughly constant mean and variance, we say that the Markov chain is well-mixed. This is what we see for each of the 8 parameters. In the bottom left quadrant is the **autocorrelation plot**, providing insight into the efficiency of the MCMC sampling process. The vertical axis of the autocorrelation plot shows the estimated autocorrelation for each of the lags (in iterations) displayed on the horizontal axis. For maximal efficiency one would like to see these autocorrelations close to zero after just a few lags. Consistent with the use of Metroplis algorithms, we do not observe great efficiency for the parameters r and η, although there is some improvement in efficiency for the parameters φ and μ. Average efficiency of sampling over all parameters is only 0.04187, implying that the 10,000 MCMC samples generated approximately 419 independent observations to estimate these parameters.

In the top right quadrant the **histogram** displays the marginal posterior distribution of the parameter. The diagnostic interest in this display comes from checking if the distribution is consistent with the distributional assumptions made about the parameter. For the mean parameters we look to see

symmetric, unimodal Normal distributions, and for the variance parameters we look to see right-skewed, unimodal distributions consistent with the Inverse Gamma distribution. In all cases the histograms are consistent with these expectations. Finally, the bottom right quadrant displays several **kernel density** plots, which convey similar information to the histograms. In addition, we have plots of the kernel densities based only on the first half of the MCMC samples, and based only on the second half of the MCMC samples. If these three kernel density plots overlap, then there is further evidence for the Markov chain having converged and mixed well. We do see this in all cases for the parameters.

Figures B9, B10, B11 and B12 display comparable diagnostic plots for subjects 1, 20, 40 and 111, respectively. Here, of course, we deal with the parameters of final interest for our analyses, those for each subject. We observe good mixing in terms of the trace plots, the histograms, and the kernel densities. The autocorrelation plots show much more efficient sampling than for the population parameters. Average efficiency over all parameters is 0.0711, implying that the 10,000 MCMC samples generated about 711 independent observations to estimate these parameters. The relatively low efficiency here can be mitigated by moving to more powerful algorithms for sampling, such as the Hamiltonian MCMC method emodied in *Stan*.

These diagnostics examine convergence and efficiency of sampling within a chain. It is also useful to check for so-called **"pseudo-convergence,"** by examining the consistency of convergence across different chains. One concern that is addressed by this approach is for local convergence to what are actually multi-modal posterior distributions. This approach leads to the Gelman-Rubin statistic, which compares the within-chain variance of parameter estimates to the between-chain variance of parameter estimates, one parameter at a time. Good pseudo-convergence occurs when this ratio is below 1.2 across parameters, and sometimes a conservative threshold of 1.1 is used (Brooks and Gelman [1998; p. 444]). Employing four chains, the ratios for the 8 population parameters are all less

than 1.007, well below any of these thresholds. Over all 452 population and individual subject parameters the ratios are all less than 1.02, again well below any of these thresholds. In terms of this statistic, there are no convergence issues.

Additional Reference

Brooks, Stephan P., and Gelman, Andrew, "General Methods for Monitoring Convergence of Iterative Simulations," *Journal of Computational and Graphical Statistics*, 7, 1998, 434-455.

Figure B1: Convergence Diagnostics for Mean of Hyper-Parameter *r*

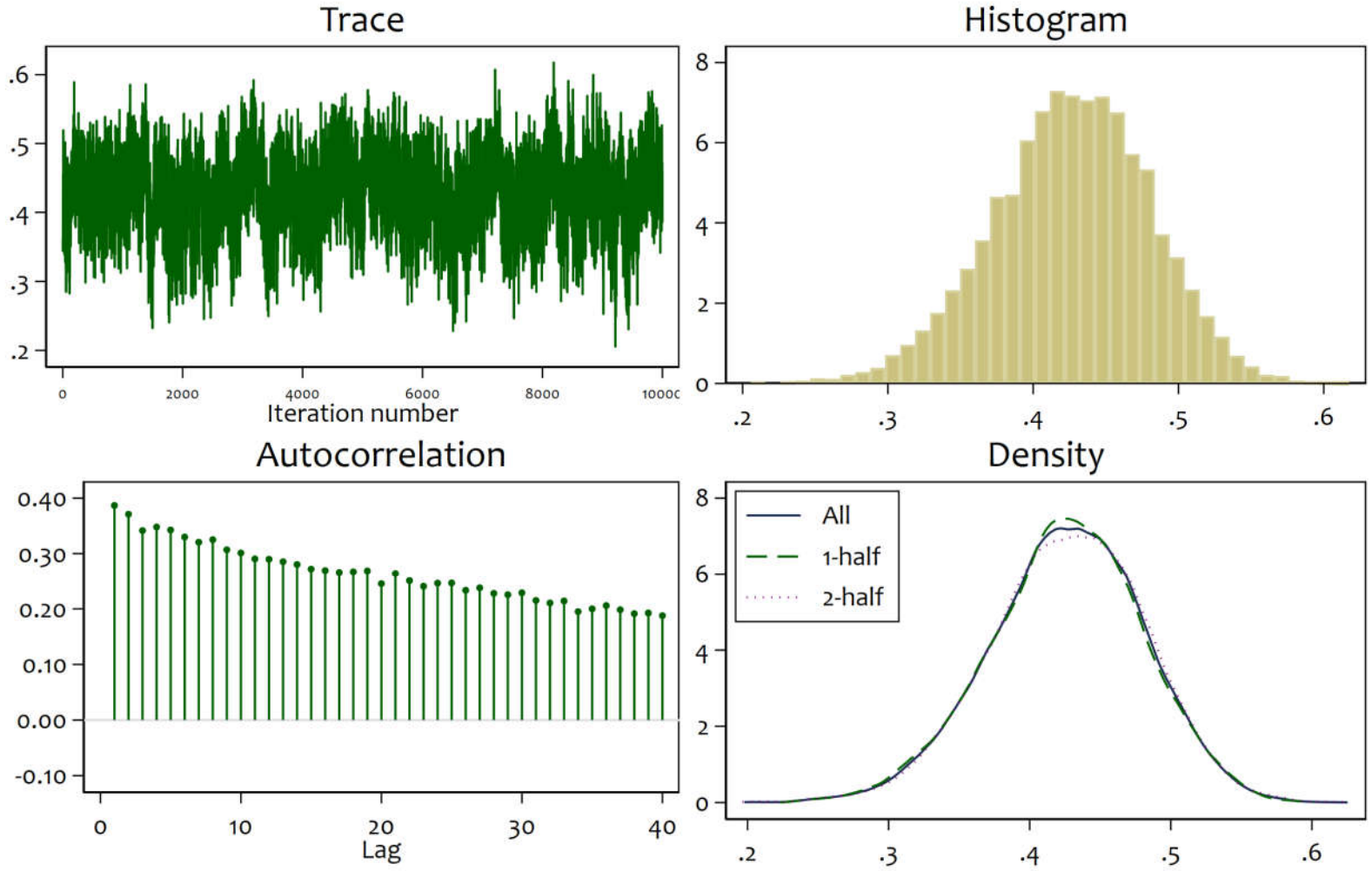


Figure B2: Convergence Diagnostics for Variance of Hyper-Parameter *r*

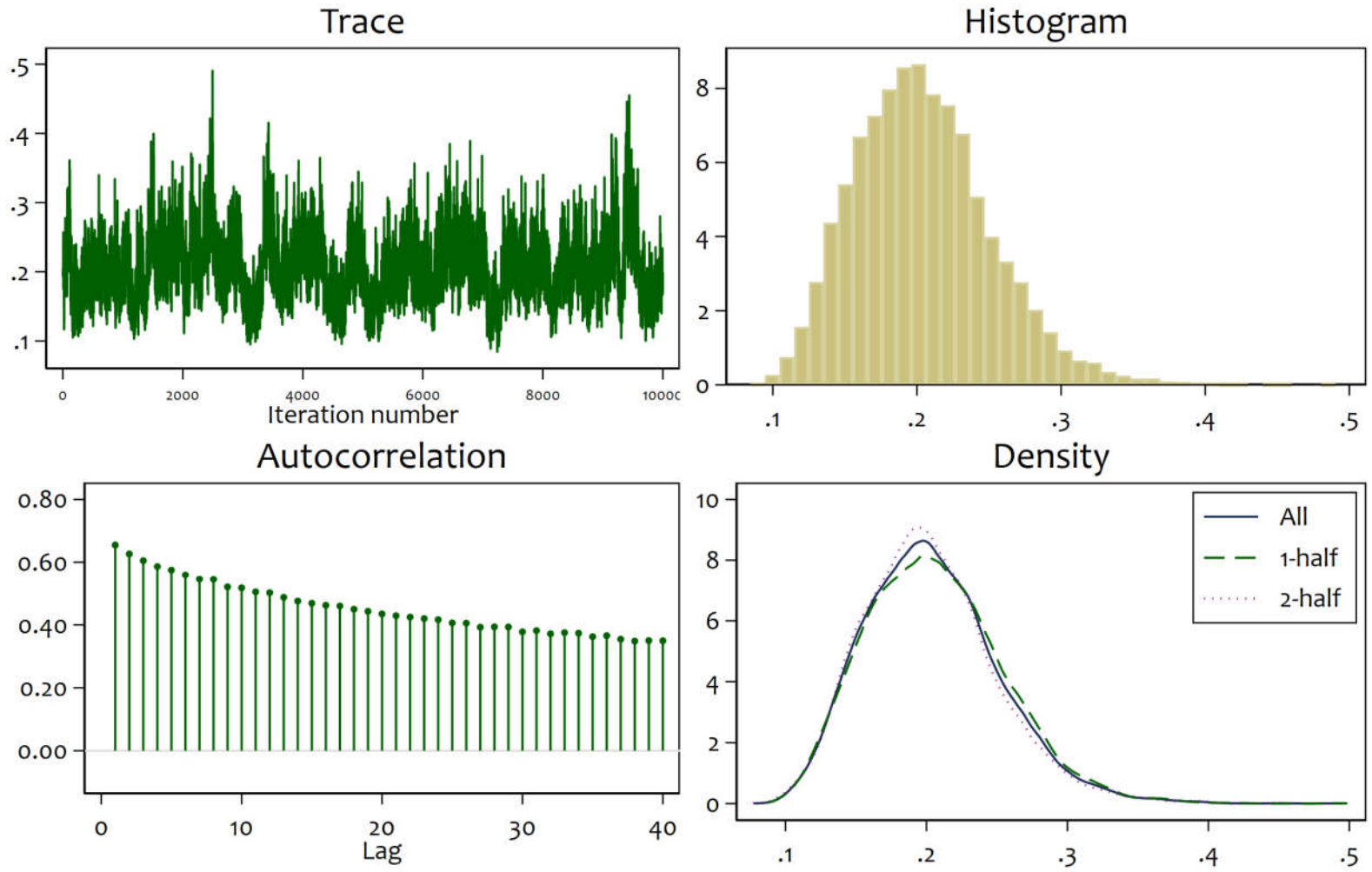


Figure B3: Convergence Diagnostics for Mean of Hyper-Parameter η

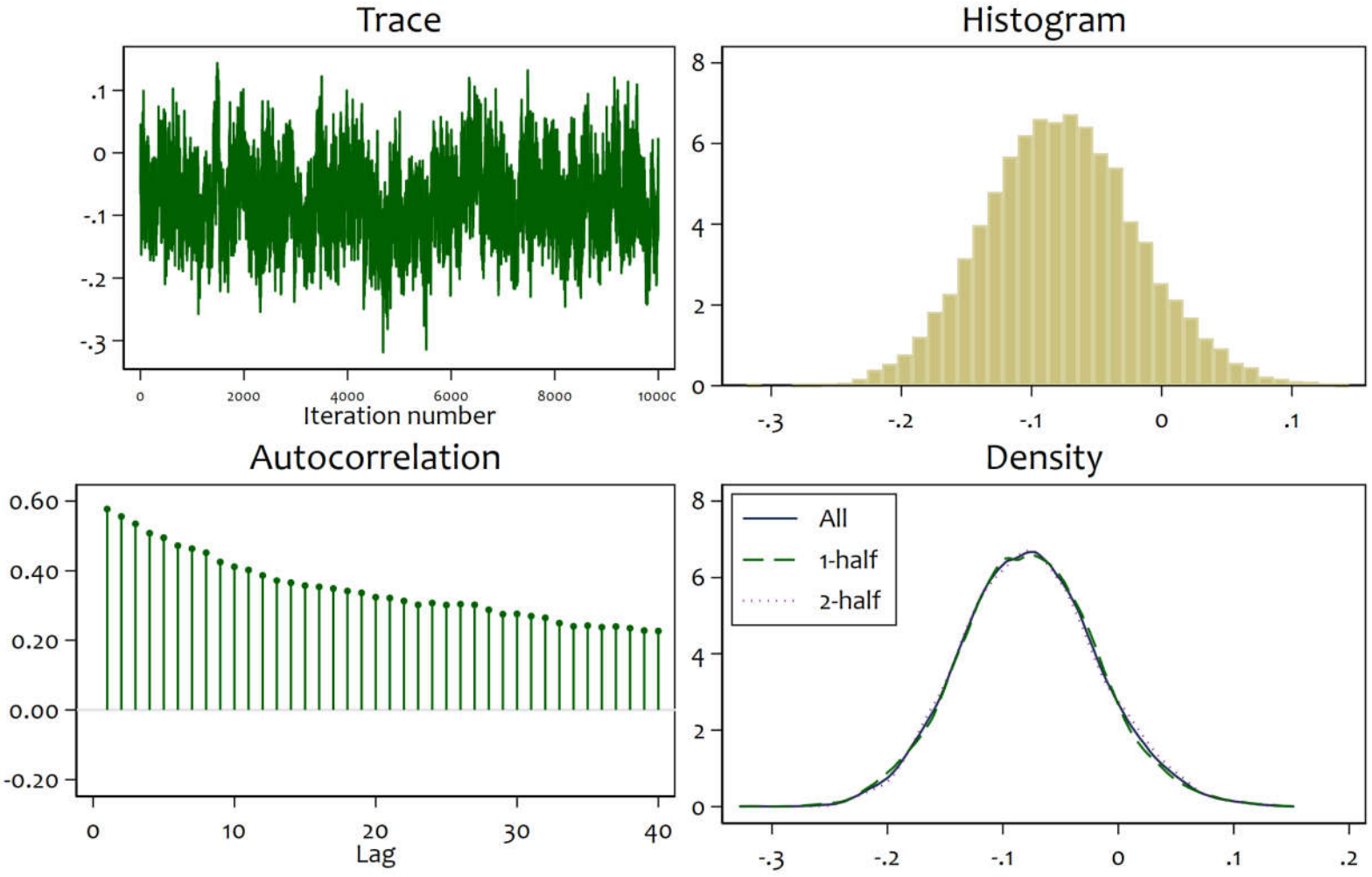


Figure B4: Convergence Diagnostics for Variance of Hyper-Parameter η

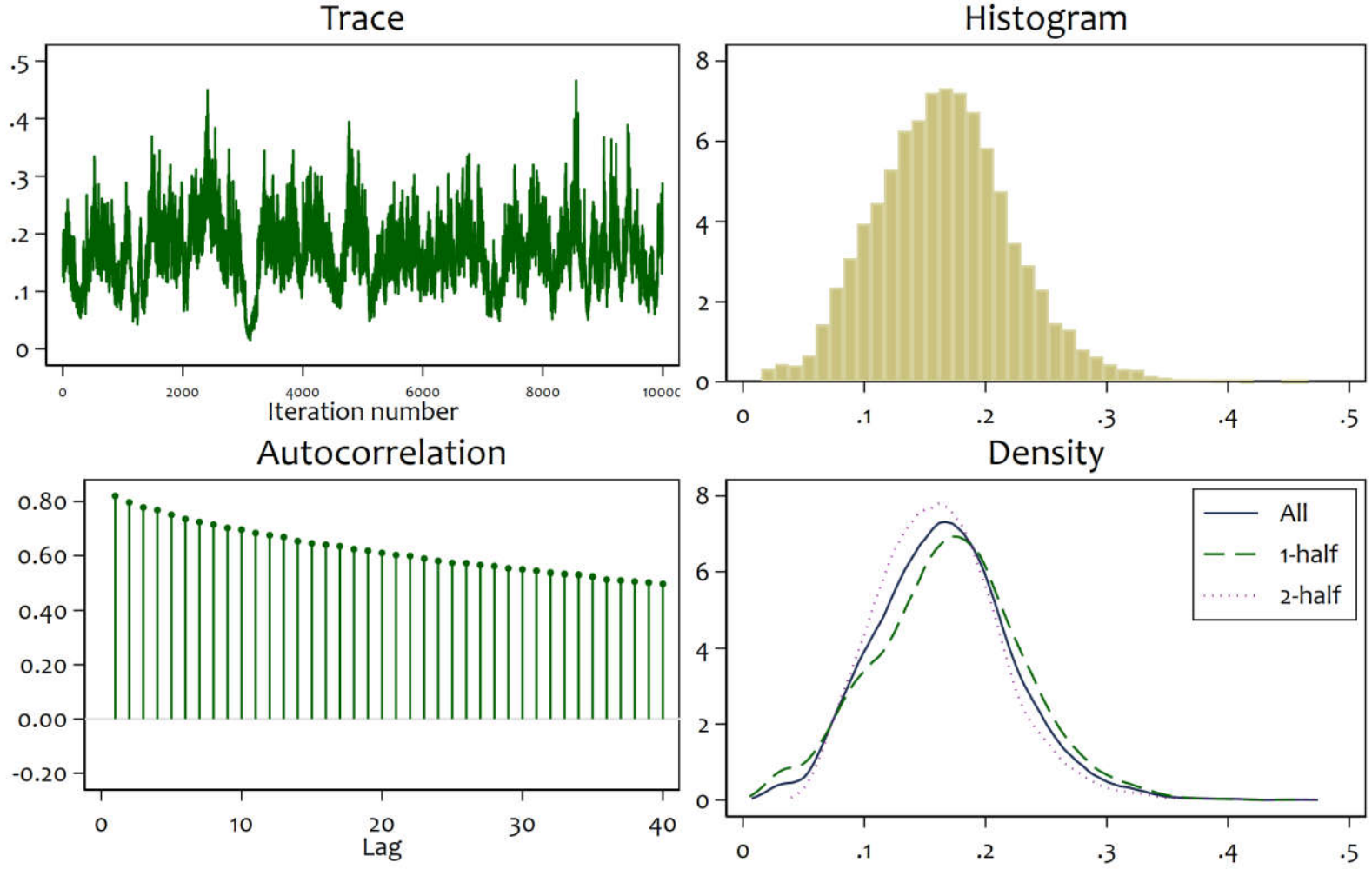


Figure B5: Convergence Diagnostics for Mean of Hyper-Parameter φ

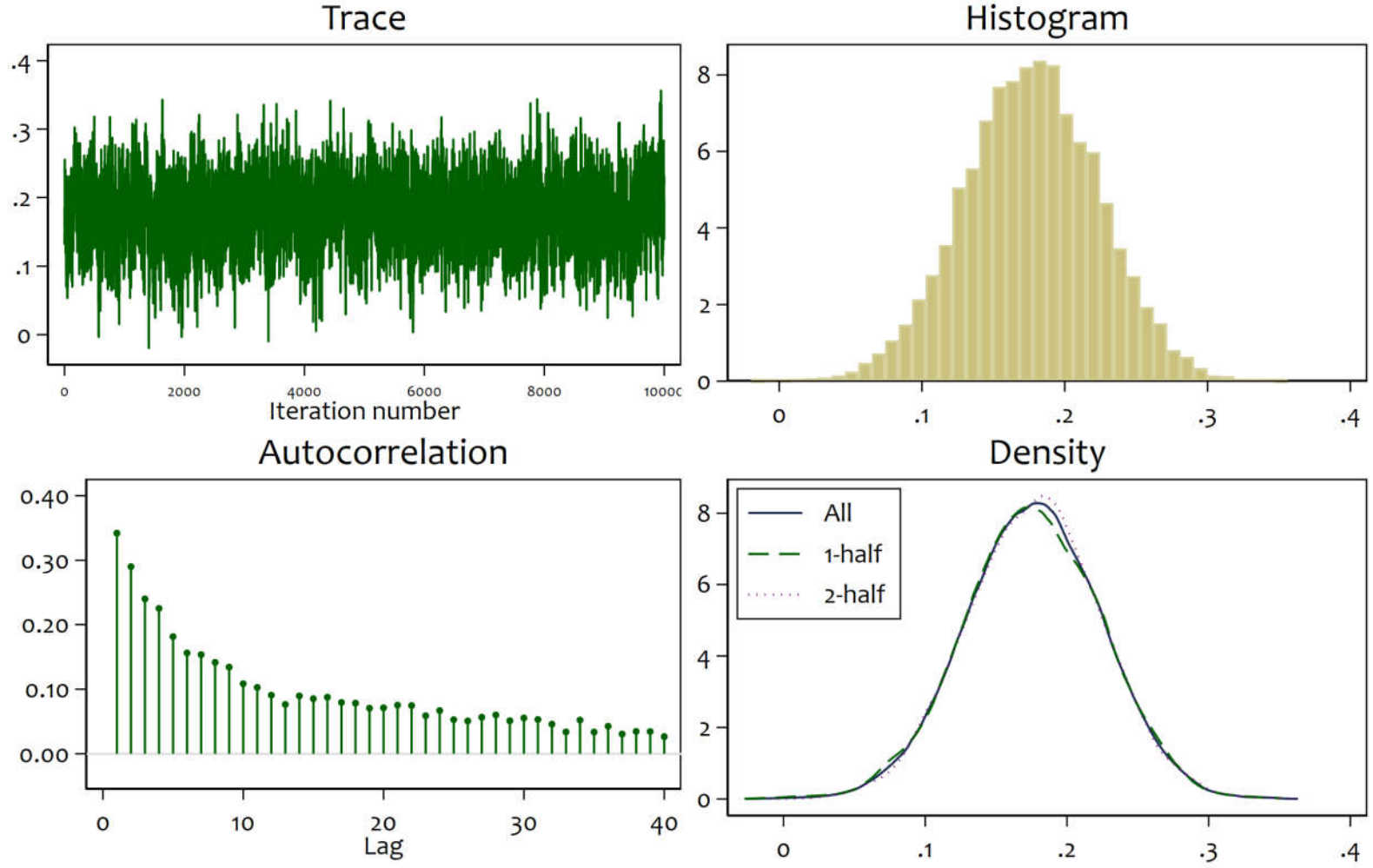


Figure B6: Convergence Diagnostics for Variance of Hyper-Parameter φ

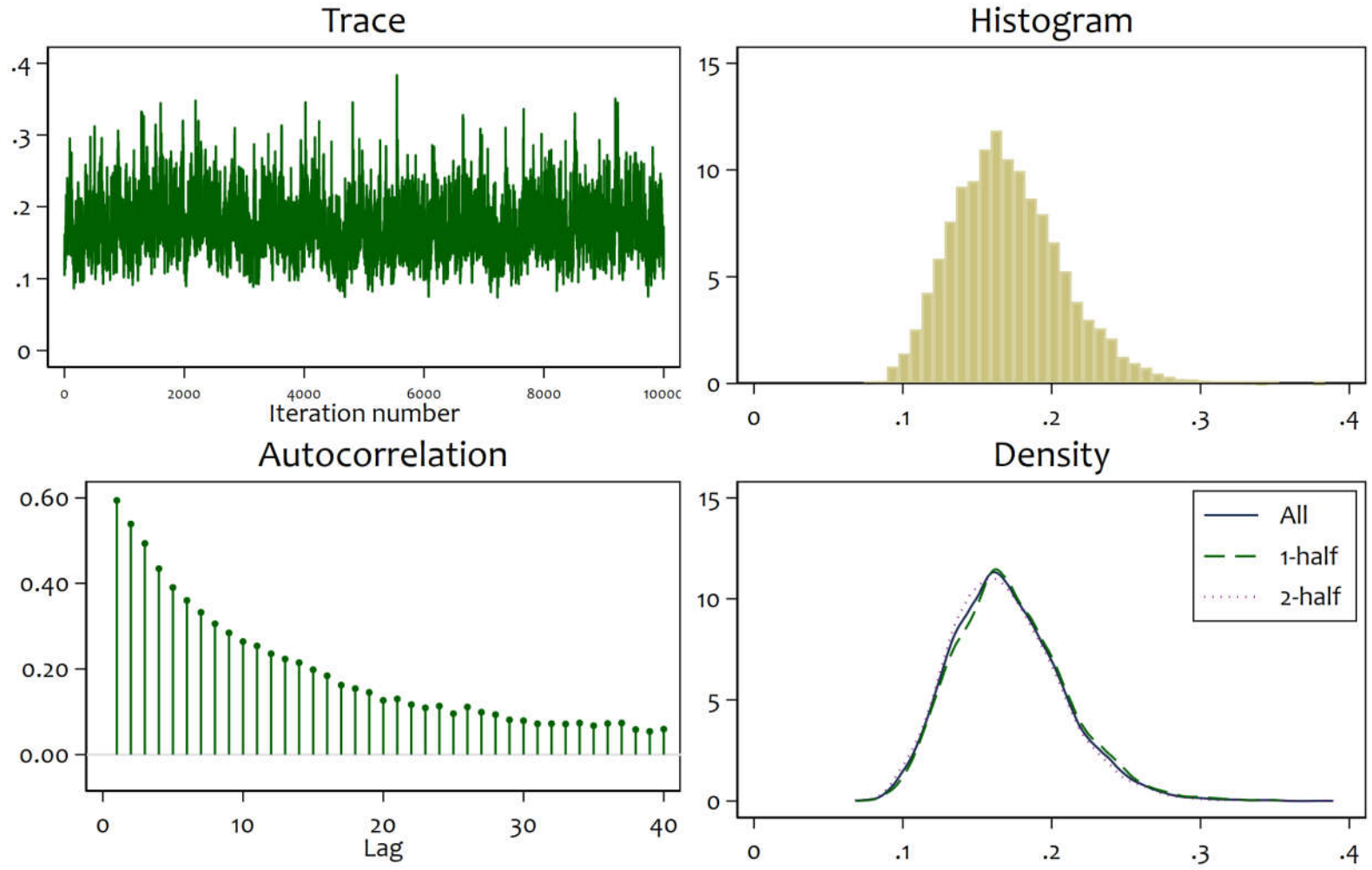


Figure B7: Convergence Diagnostics for Mean of Hyper-Parameter μ

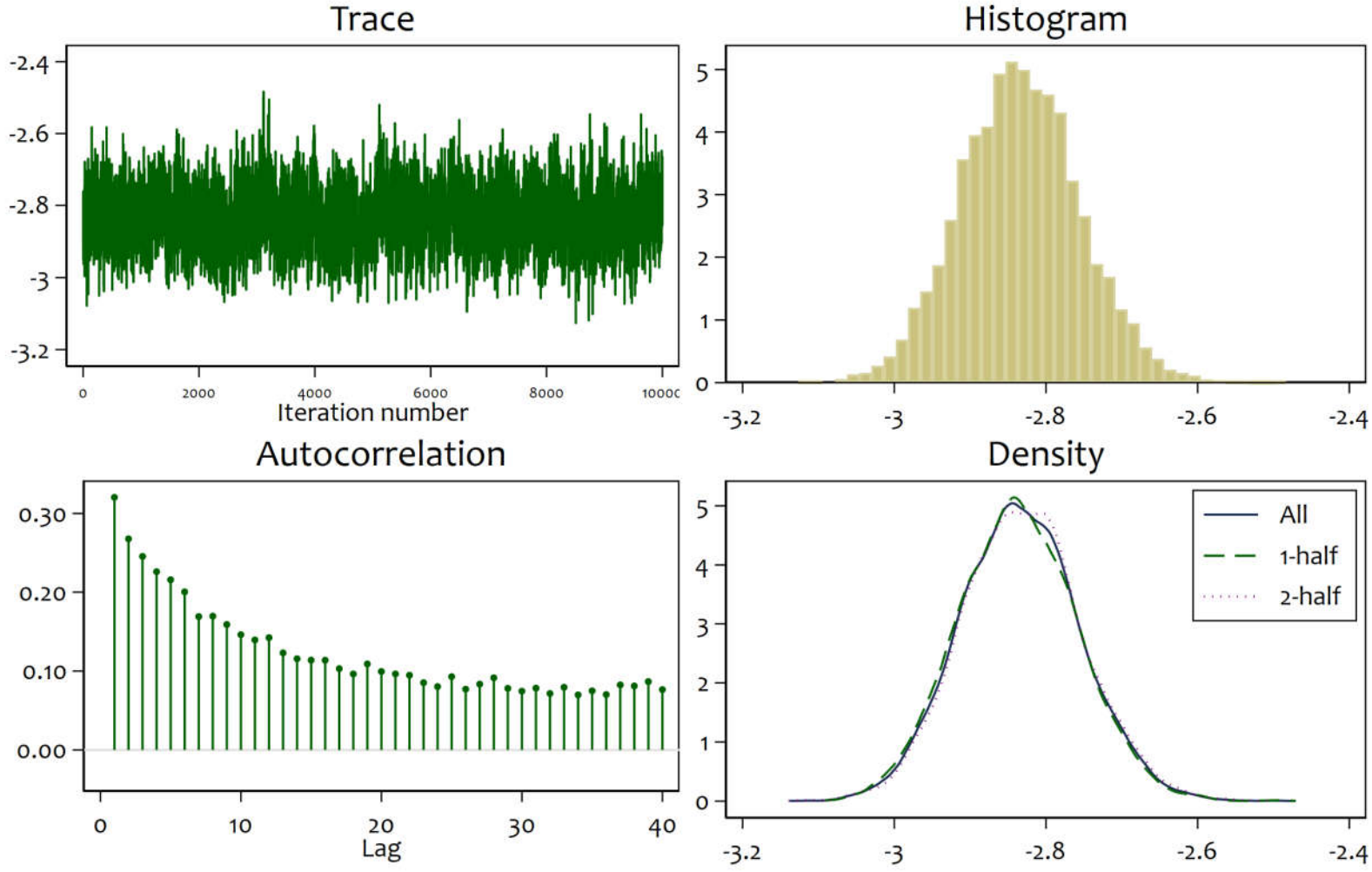


Figure B8: Convergence Diagnostics for Variance of Hyper-Parameter μ

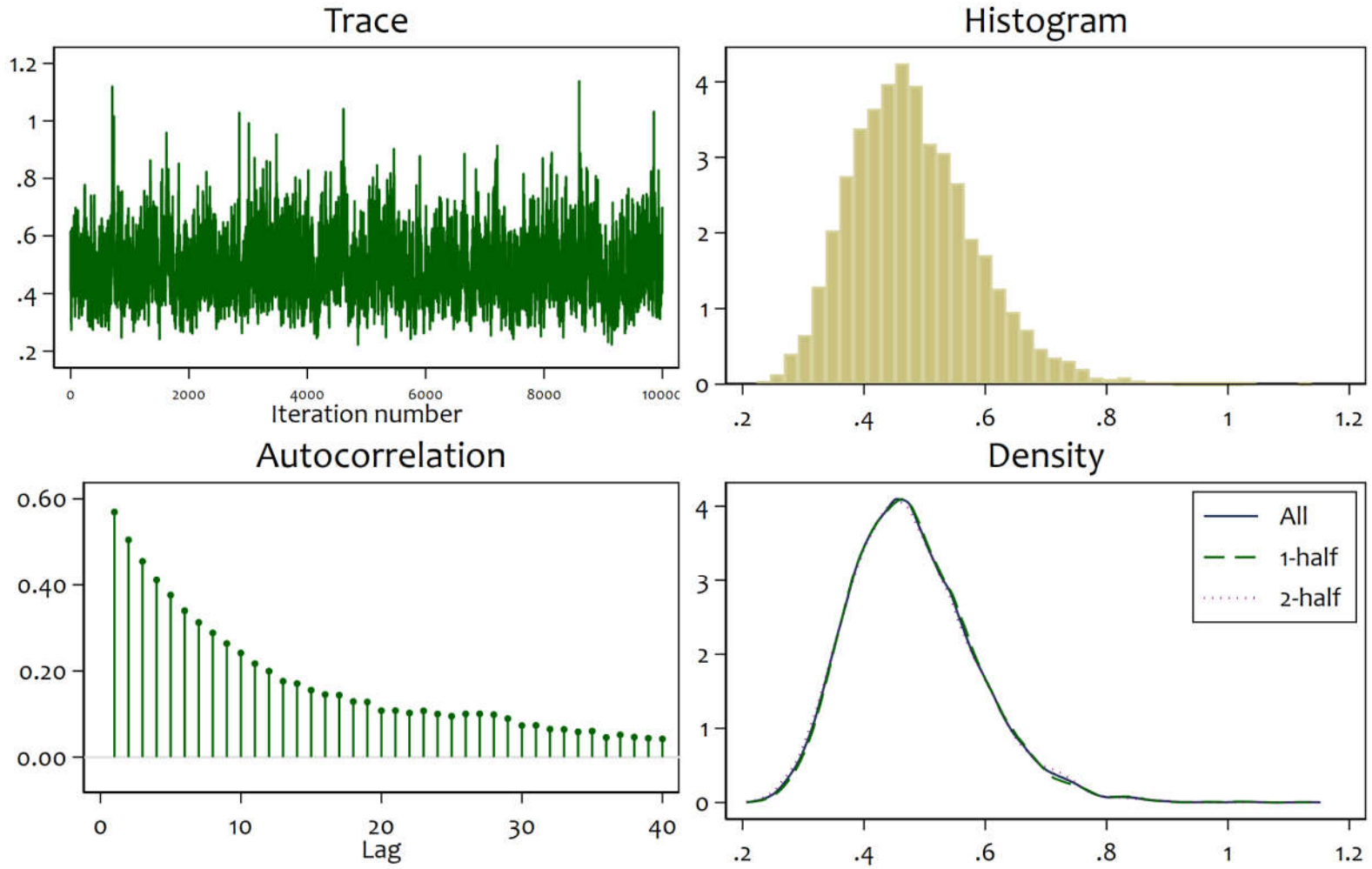


Figure B9: Convergence Diagnostics for Subject #1 Parameters

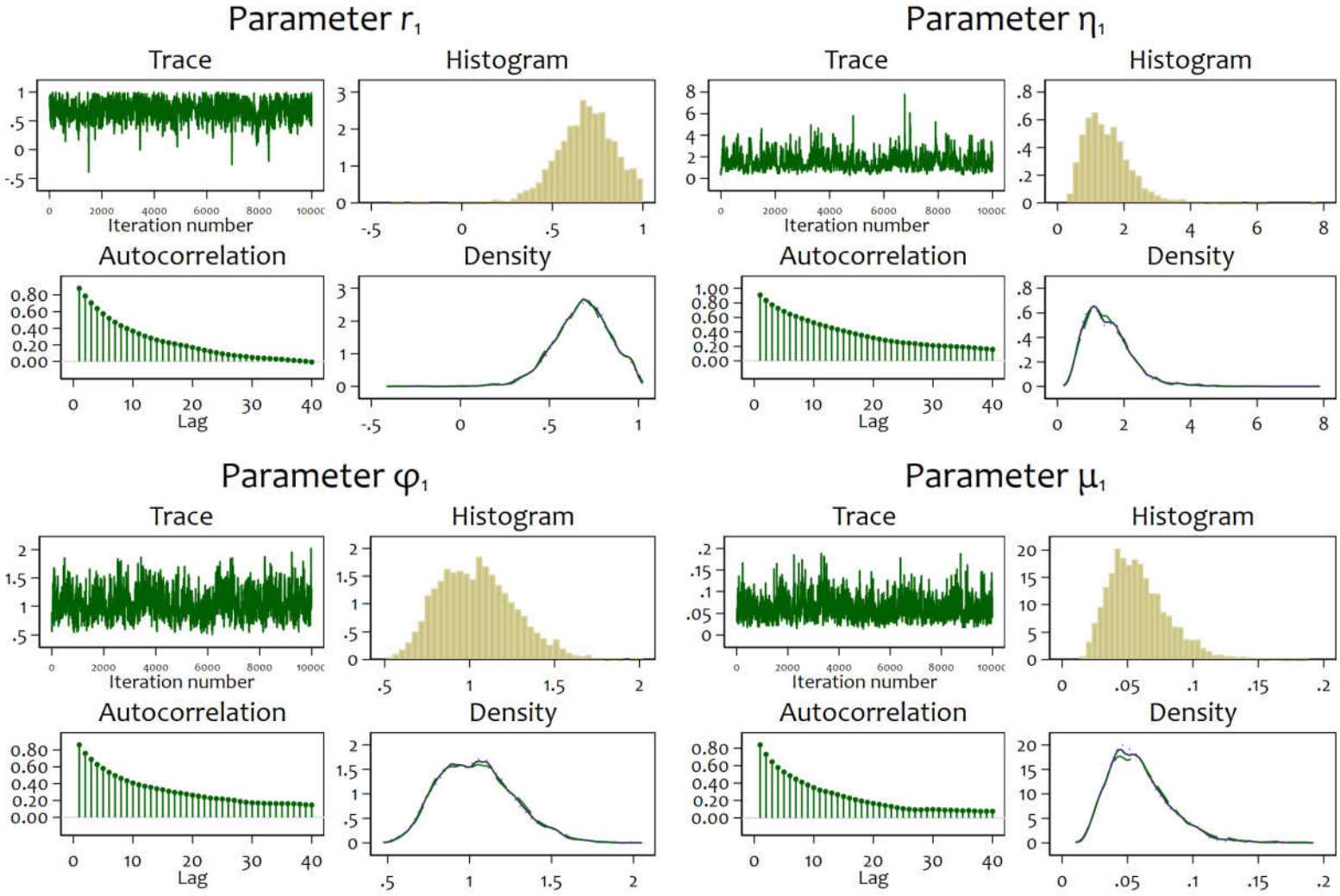


Figure B10: Convergence Diagnostics for Subject #20 Parameters

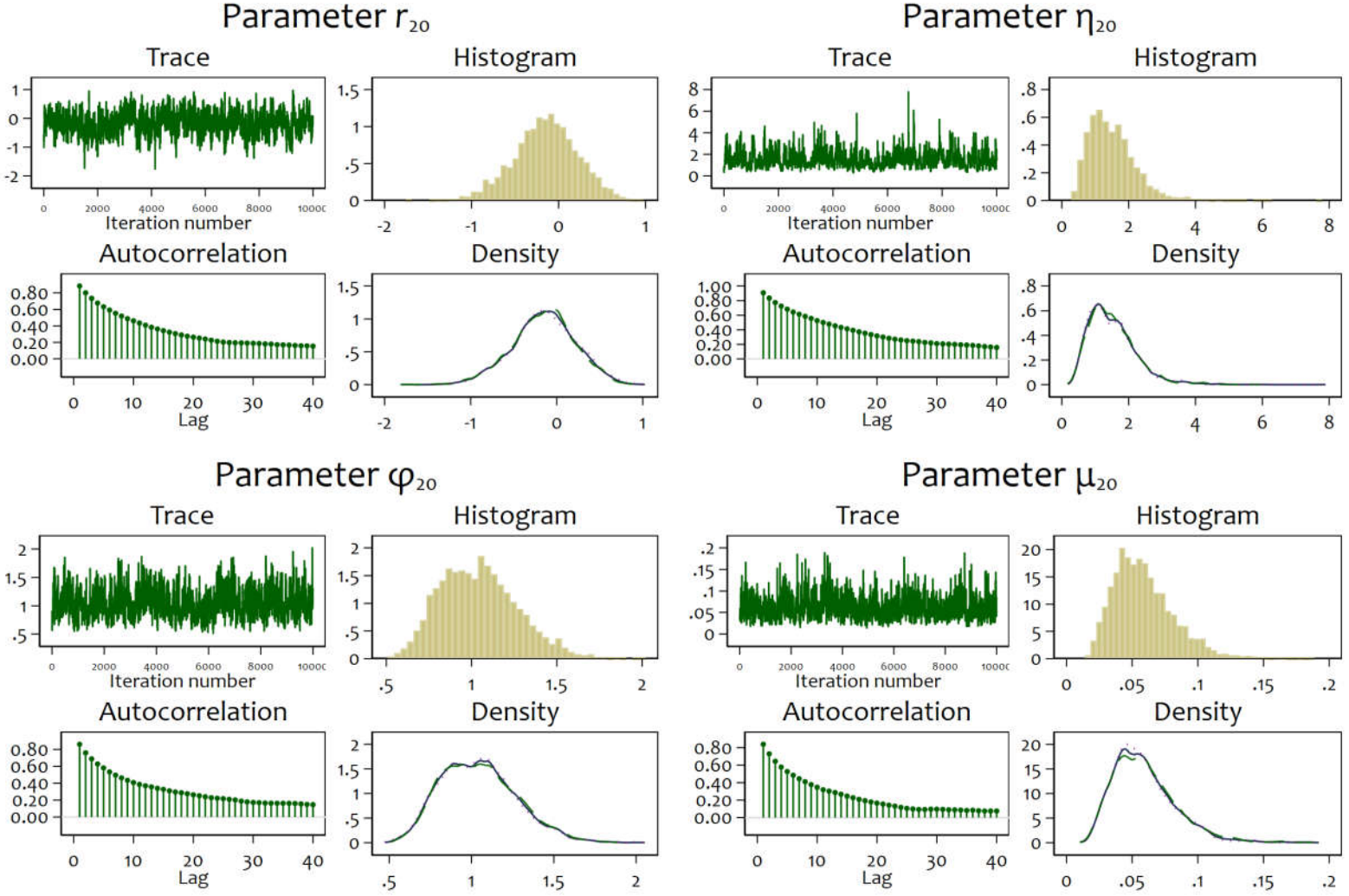


Figure B11: Convergence Diagnostics for Subject #40 Parameters

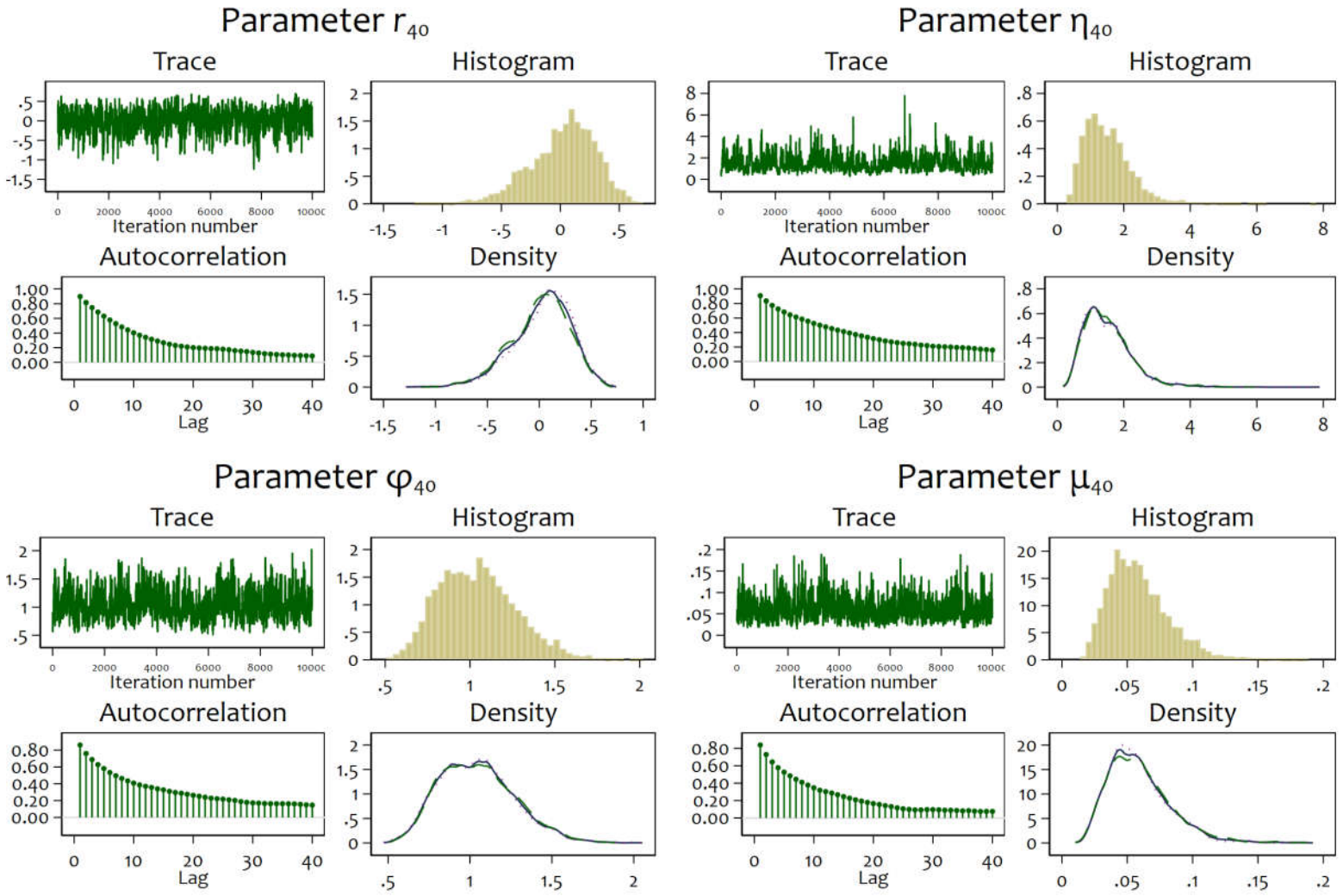


Figure B12: Convergence Diagnostics for Subject #111 Parameters