Risk Preferences, Gender Effects and Bayesian Econometrics

Jessica Alam[*] Konstantinos Georgalos[†] Harrison Rolls[‡]

June 22, 2022

Abstract

Gender differences in decision making is a topic that has attracted much attention in the literature and the debate seems to be inconclusive. A method that is often used in the economics literature to account for gender effects is by estimating econometric structural models and testing the significance of the estimated parameters. In this paper we focus on estimations of preference models and we show how omitting to account for behavioural heterogeneity can lead to failures in identifying potential differences. Using data from risky choice experiments, we compare the traditional representative agent Maximum Likelihood Estimation approach against two more flexible inference methods that allow for heterogeneity at the individual level, the Maximum Simulated Likelihood Estimation, and the Hierarchical Bayesian modelling. We show how ignoring heterogeneity may lead to failures capturing gender differences and we suggest the use of Bayesian modelling to effectively estimate the underlying parameters.

JEL classification: C11, C51, C52, D81, D91

Keywords: Gender differences; Risk preferences; Loss aversion, Rank-dependent utility; Prospect Theory; Maximum likelihood; Hierarchical Bayesian modelling

[*] Department of Economics, Lancaster University Management School, LA1 4YX, Lancaster, U.K.
✉ j.y.alam@lancaster.ac.uk,

[†] Department of Economics, Lancaster University Management School, LA1 4YX, Lancaster, U.K.
✉ k.georgalos@lancaster.ac.uk,

[‡] Department of Economics, Lancaster University Management School, LA1 4YX, Lancaster, U.K.
✉ h.rolls@lancaster.ac.uk

We are grateful to Hong Il Yoo for providing helpful comments. We would also like to thank the Guest Editor Lionel Page and an anonymous referee for their very helpful comments that led to significant improvements in both the analysis and the presentation. The support of The High End Computing (HEC) facility at Lancaster University is gratefully recognised.

1 Introduction

There is no doubt that risk preferences play a central role in every aspect of economic life. Gender differences in risk preferences is a much debated topic and it has often been argued that these differences might provide a possible explanation of the observed differences between the two genders in various aspects of economic life such as financial decision making, hold of front office roles, or entrepreneurship, to name but a few. Nevertheless, there is little agreement on whether there is a universal pattern of differences between the two genders. Early surveys from the economic literature (Eckel and Grossman, 2008; Croson and Gneezy, 2009) provide mostly supporting evidence of women being less willing to accept risks. Recently, Filippin and Crosetto (2016) conducted an extensive meta-analysis on gender differences and risk attitudes, using data from 7000 subjects and 54 replication studies of the Holt and Laury (2002) risk elicitation task. One of their main findings is that:

> “[..] gender differences appear in less than 10% of the studies and are significant but negligible in magnitude once all the data are pooled.”

and they conclude that:

> “[..] the structural model seems to confirm that significant gender differences are detected in the HL task when merging all the observations. The reason is to be found in the sky-rocketing increase of the statistical power of the test, which drives fairly close to zero the likelihood of observing a false negative when data are merged.”

The above statement indicates that in order to be in place to detect any potential gender effects, one needs to recruit an extremely large sample of subjects, for the standards of economic experimentation, a task which seems prohibiting given all the time, financial and practical constraints that a researcher may face. In this paper, motivated by the conclusions of Filippin and Crosetto (2016), we investigate how one can increase the extracted information from small sample datasets, and what are the implications of omitting to do so.

One of the most common approaches to explore potential differences between genders is to assume a particular preference functional, pool all the data together, and estimate a representative agent model, using demographic dummy variables to control for heterogeneity (see Harrison and Rutström, 2008, Xie et al. (2017), Vieider et al., 2015, Bouchouicha et al., 2019). Parameters are then obtained by using either Maximum Likelihood Estimation techniques (MLE) or Non-Linear Least Squares estimation methods, and the statistical significance of the dummies defines the existence and the size of potential differences. While the representative approach is attractive, due to its simplicity , it comes with a serious limitation. By ignoring individual heterogeneity, the estimated preferences may not be representative for any of the subjects. Consider an extreme scenario where out of 100 subjects, 50 are male and risk neutral, 25 female and risk seeking with a risk coefficient of -0.50 (assuming a power utility function as in Holt and Laury, 2002 and later in our analysis of the form $x^{1-r}/(1-r)$) and the remaining 25 subjects are females and risk averse, with a coefficient of 0.50. Pooling all the data together and fitting a representative agent model to this dataset, including a control variable to capture potential gender differences, will return an estimated risk aversion very close to zero, implying risk neutrality, and the coefficient of gender effects to be insignificant[1]. The main conclusion that a researcher could draw from a similar analysis is that the observed population has risk neutral preferences and there are no gender effects. Consider now a policy maker who aims to identify the risk seeking women in a population. By conducting a similar analysis, the policy maker will reach the conclusion that no risk seeking women exist in this sample and no action needs to be taken. While this example is extreme and perhaps improbable, it is used to highlight the impact of ignoring potential behavioural heterogeneity in identifying preferences and differences based on demographic criteria.

On the other end of the spectrum, one could estimate preference functionals at the individual subject-level (see Hey and Orme, 1994, Stott, 2006). While this approach takes into

[1] We indeed executed a similar simulation exercise where this result was confirmed. Details are available on request.

consideration the individual characteristics of each subject, a large amount of data points is required in order to obtain robust and reliable estimates. This comes at a high cost for the researcher, as larger number of decision tasks would mean longer sessions that could potentially lead to boredom and eventually to more noisy data.

In the present study we compare the representative agent modelling approach, to two more flexible and informative methods of parameter estimation that allow one to simultaneously make inferences at both the individual subject and the experimental population level. In particular, we compare the frequentist and the Bayesian methods, by analysing the data using Maximum Simulated Likelihood Estimation techniques (MSLE), as well as Hierarchical Bayesian (HB) econometric modelling. We use data from three prominent studies of decision making under risk. First, we use the original data from the Holt and Laury (2002) experiment, and assuming Expected Utility preferences, we first show how all three inference methods are able to capture gender effects. Then, we extend our analysis to non-Expected Utility preferences, and particularly to Rank Dependent Utility, since our focus in on risky choice in the gains domain. Using the dataset from Baillon et al. (2020), we show that taking into consideration individual heterogeneity, improves the inference, while the MLE representative agent model fails to identify the existence of gender differences. Finally, we focus on the domain of losses, and adopting a Cumulative Prospect Theory framework, we explore the differences between the two genders, across all the components of risk preferences, namely utility curvature, probability weighting and loss aversion. We show how MLE fails to capture gender differences and we also focus on the differences between the MSLE and the HB methods in capturing these differences.

Our results can be summarised as follows. When there is a small number of parameters to estimate, any of the inference methods will be able to detect the presence of gender differences in the key behavioural parameters. As the model complexity increases, and therefore the number of parameters along with their collinearity, more flexible methods that take into consideration individual heterogeneity, provide more robust inference when the focus is on the difference between two populations. We complement our study with an extensive Monte Carlo simulation to compare the three inference methods, and we show that while all MLE, MSLE and HB methods are able to successfully recover the mean values of the simulated parameters, frequentist methods are more prone to ignore statistical significance due to overfitting, compared to Bayesian methods.

The rest of the paper is organised as follows: section 2 briefly introduces the idea of Hierarchical Bayesian modelling, section 3 focuses on the Holt and Laury (2002) risk elicitation task and presents, along with the task and the data, the econometric specification for both MLE and HB, assuming Expected Utility preferences (EU), section 4 relaxes the hypothesis of EU and introduces Rank Dependent Utility preferences using data that allow the estimation of such preferences, and finally, section 5 focuses on the domain of losses, introducing a Cumulative Prospect Theory model and loss aversion. In section 6 we report the results of the simulation. We then conclude.

2 Frequentist Vs Bayesian Parameter Estimation

The most common approach to estimate structural decision making models is by either pooling all data together and fit a representative agent model, or by assuming complete independence and fit subject-level models, using maximum likelihood estimation techniques (MLE). Fitting a representative agent model ignores much of individual behavioural heterogeneity and generates estimates which potentially, are not representative of any individual subject in the sample. A simple way to introduce heterogeneity to the representative model is to condition the parameters to a set of observable demographics and assume that subjects that belong to the same demographic group share the same behavioural parameters (see for example Harrison and Rutström, 2008, Bouchouicha et al., 2019). An alternative way to introduce heterogeneity, within the frequentist framework, is to use a *random-coefficients* model, a popular method to model unobserved heterogeneity, on top of the observed one (e.g. through demographics). In this kind of modelling, it is assumed that each behavioural parameter in the model is characterised by an underlying distribution across the population. Using MLE techniques and simulation, it is possible to combine estimates of the population distribution (mean and standard deviation) with individual choices, and make inferences at both the population and the subject level (for applications see von Gaudecker et al., 2011; Conte et al., 2011; Moffatt, 2016). Nevertheless, it is known that MLE is susceptible to overfitting and may generate noisy and unreliable estimates when there is a lack of a large number of observations (see Bishop, 2006, pp. 166, Nilsson et al., 2011). An alternative method to introduce heterogeneity and mitigate these drawbacks is to adopt Hierarchical Bayesian estimation techniques (see Balcombe and Fraser, 2015; Ferecatu and Önçüler, 2016 and Baillon et al., 2020 for some recent applications of hierarchical models for choice models under risk and Stahl, 2014 for ambiguity models.). The key aspect of hierarchical modelling is that even though it recognises individual variation, it also assumes that there is a distribution governing this variation (individual parameter estimates originate from a group-level distribution). As Baillon et al. (2020) highlight, Hierarchical Bayesian modelling is a compromise between a representative agent and subject-level type estimation. It estimates the model parameters for each subject separately, but it assumes that subjects share similarities and draw their individual parameters from a common, population level distribution. In that way, individual parameter estimates inform each other and lead to a *shrinkage* towards the group mean that reduces biases in parameter estimates. The latter leads to more efficient and reliable estimates compared to those estimated using frequentist methods. One of the most crucial aspects of Bayesian inference, is the way uncertainty is incorporated in the econometric model in the form of probability distributions. A researcher can use her subjective beliefs or objective knowledge and form a prior distribution which summarises all the available knowledge regarding a particular parameter, before observing any data. In Bayesian inference, the estimation of a parameter of interest corresponds to the calculation of the probability distribution over the parameter, given the observed data and the prior beliefs. Another aspect of the Hierarchical model is that it is applied in an hierarchical form providing both within decision unit analysis (subject level) and across unit analysis (population level). Both the way the Bayesian model incorporates uncertainty and its Hierarchical structure, allows it generate precise estimation of preferences, even when the available data are limited.

Jacquement and L'Haridon (2018, p. 247) provide a comparison between the frequentist and Bayesian methods, highlighting the most important differences, namely the way each method interprets each parameter, the nature of the point estimation, the way intervals for statistical significance are estimated, and; the way hypothesis testing can be done. For the frequentist method the parameter is an unknown constant while for the Bayesian a random variable. Similarly, the point estimation will be the value of the estimator in the former, while a posterior summary in the latter (e.g. the mode of the distribution). For statistical significance, the frequentist method requires the estimation of confidence intervals, compared to the credible intervals in the Bayesian inference. As Huber and Train (2001) point out, in the presence of small samples, the two procedures can provide numerically different results, due to the different way of treating uncertainty in the parameters of the population distribution. In what follows, we compare the three different inference methods (MLE, MSLE and HB) in their capacity to detect gender differences, focusing on three representative examples of decision making under risk.

3 Risk Preferences and Expected Utility

Gender differences in risky decision making has been the topic of numerous studies. Eckel and Grossman (2008) and Croson and Gneezy (2009) summarise the literature, finding that female subjects tend to be more risk averse. Charness and Gneezy (2012) and Holt and Laury (2014) discuss how the risk elicitation task affects the inference on differences, while Filippin and Crosetto (2016) challenge the early evidence by finding that the observed effects are negligible in magnitude. In this section we focus on one of perhaps the most common elicitation methods that has been used in the literature, the Holt and Laury (2002) task.

3.1 Decision Task and Data

For the analysis, we use the data from the original Holt and Laury (2002) study. Each subject is presented with the 10 choice tasks, as shown in Table 1. Each task consists of a choice between two paired lotteries A and B. The payoffs for lottery A are fixed to \$2.00 and \$1.6, while for lottery B, the payoffs are 3.85 and \$0.10. Since lottery A is characterized by less variable payoffs, one can label A as the *safe* option and B the *risky* one. In the first choice task, the probability of getting the high payoff is equal to 10% for both lotteries, and it increases as one moves down the table. At the first row, only the extremely risk seeking subjects are expected to choose lottery B. A risk neutral person is expected to choose lottery A for the first 4 tasks (since the expected value of lottery A is greater) and then switches to lottery B for the remaining tasks. Holt and Laury (2002), assuming a particular form of risky preferences, provide a mapping between then number of safe choices and the value of risk coefficient of a subject (the higher the degree of risk aversion, the higher the number of safe choices).

There are data from 212 subjects (95 females) from 4 treatments, an incentivised low-payoff treatment (LOW_1), with payoffs as those in Table 1, a hypothetical treatment (HYP), with the payoffs scaled up by 20, 50 or 90, an incentivised high-payoff treatment ($HIGH$), with payoffs scaled up by 20, and finally, a low-payoff treatment (LOW_2), identical to the first one. For our purposes, we use only the data from the low-payoff treatment (LOW_1).

Table 1: The 10 Lotteries from Holt and Laury (2002).

	Option A				Option B			
Task	p_{A1}	x_{A1}	p_{A2}	x_{A2}	p_{B1}	x_{B1}	p_{B2}	x_{B2}
1	0.1	2.00$	0.9	1.60$	0.1	3.85$	0.9	0.10$
2	0.2	2.00$	0.8	1.60$	0.2	3.85$	0.8	0.10$
3	0.3	2.00$	0.7	1.60$	0.3	3.85$	0.7	0.10$
4	0.4	2.00$	0.6	1.60$	0.4	3.85$	0.6	0.10$
5	0.5	2.00$	0.5	1.60$	0.5	3.85$	0.5	0.10$
6	0.6	2.00$	0.4	1.60$	0.6	3.85$	0.4	0.10$
7	0.7	2.00$	0.3	1.60$	0.7	3.85$	0.3	0.10$
8	0.8	2.00$	0.2	1.60$	0.8	3.85$	0.2	0.10$
9	0.9	2.00$	0.1	1.60$	0.9	3.85$	0.1	0.10$
10	1.0	2.00$	0	1.60$	1.0	3.85$	0	0.10$

3.2 Theoretical Framework and Econometric Specification

We assume that the agent holds Expected Utility preferences and receives utility from income according to a Constant Relative Risk Aversion (CRRA) utility function of the form:

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

where x is the monetary payoff, and r is the risk coefficient with $r > 0$ indicating a concave utility for gains (risk aversion), $r < 0$ a convex utility (risk seeking) and $r = 0$ a linear utility (risk neutrality). For $r = 1$ the function collapses to the logarithmic function. A lottery is evaluated by the weighted sum of the utilities of the payoffs, therefore, the expected utility of lottery A, for a particular task, is given by

$$EU_A = p_{A1}\frac{{x_{A1}}^{1-r}}{1-r} + (1-p_{A1})\frac{{x_{A2}}^{1-r}}{1-r} \tag{2}$$

To account for the stochastic nature in choices, we assume a *logit* link function. Thus, the probability of choosing lottery A is given by:

$$P(A) = \frac{\exp(1/\xi EU_A)}{\exp(1/\xi EU_A) + \exp(1/\xi EU_B)} \tag{3}$$

with ξ a precision parameter to be estimated. According to the above assumptions, the log-likelihood function is given by:

$$LL(\theta) = \sum_{n=1}^{N}\sum_{i=1}^{I} y_{ni}\ln(P_{ni}(A_i)) + (1-y_{ni})\ln(1-P_{ni}(A_i)) \tag{4}$$

where N is the total number of subjects, I is the number of tasks, $y_{ni} = 1(0)$ is an indicator function denoting the choice of lottery A(B) for subject n in task i, and θ is the vector of behavioural parameters to be estimated. Therefore, there are 2 parameters to estimate, the risk coefficient r and the precision parameter ξ. To introduce gender effects, we introduce a dummy variable y_{FEMALE} which takes the value 1 is the subject is female, otherwise it is equal to 0. For each parameter θ_n in our model, with $\theta_n \in \{r, \xi\}$ we specify

$$\theta_n = \theta_0 + \theta_{FEMALE} \times y_{FEMALE} \tag{5}$$

Since we consider a stochastic model which takes into consideration the errors of the decision maker, we include in the analysis the observations of all the subjects (rather than focusing only on subjects without multiple switches). There are in total 4 parameters to estimate, the risk coefficient, the precision parameter and the two parameters that capture gender effects. [2]

[2] For the estimation we use a general nonlinear augmented Lagrange multiplier optimisation routine that allows

For the HB estimation, we follow Rouder and Lu (2005) and Nilsson et al. (2011) set-up. Each subject n made a series of I binary choices in a given dataset and the observed choices vector is denoted by $D_n = (D_{n1} \cdots D_{nI})$. Every subject is characterised by its own parameter vector $\Theta_n = (r_n, \xi_n)$, and we assume that both the utility curvature r_n and the sensitivity parameter ξ_n are normally distributed ($\theta_n \sim N(\mu_\theta, \sigma_\theta)$), while for the hyper-parameters we assume normal priors for the mean μ_θ and uninformative priors (uniform) for σ_θ. We also follow the standard procedure and transform all the parameters to their exponential form to ensure that they lie within the appropriate bounds (see Balcombe and Fraser, 2015). To capture gender differences, we condition the mean of all parameters to a female covariate. For each subject n, each parameter θ_n is assumed to be drawn from a normal distribution of the form: $\theta_n \sim N(\theta + \theta_{FEMALE} \times y_{FEMALE}, \sigma^2_\theta)$, with y_{FEMALE} a female dummy variable. That is, the mean between the two groups differs by θ_{FEMALE}. In what follows, we use either a normal or a log-normal distribution, depending on whether there are constraints for a parameter to be strictly positive.

The likelihood of subject's n choices is given by:

$$P(D_n|\Theta_n) = \prod_{i=1}^{I} P(D_{n,i}|\Theta_n)$$

where $P(D_{n,i}|\Theta_n)$ is given by:

$$LL(\theta) = \sum_{i=1}^{I} y_{ni} \ln(P_{ni}(A_i)) + (1 - y_{ni}) \ln(1 - P_{ni}(A_i)) \tag{6}$$

Combining the likelihood of the observed choices and the probability distribution of all the

for random initialisation of the starting parameters as well as multiple restarts of the solver, to avoid local maxima. The estimation was conducted using the *R* programming language for statistical computing (The *R* Manuals, version 3.6.1. Available at: http://www.r-project.org/).

behavioural parameters, the posterior distribution of the parameters is given by:

$$P(\Theta|D) \propto P(D|\Theta) \times P(\Theta)$$

with $P(D|\Theta)$ being the likelihood of observed choices over all the subjects and $P(\Theta)$ the priors for all parameters in the set Θ.

Monte Carlo Markov Chains (MCMC) were used to estimate all the specifications. The estimation was implemented in JAGS (Plummer, 2017). The posterior distribution of the parameters is based on draws from two independent chains, with 50,000 MCMC draws each. Due to the high level of non-linearity of the models, there was a burn-in period of 25,000 draws, while to reduce autocorrelation on the parameters, the samples were thinned by 10 (every tenth draw was recorded). Convergence of the chains was confirmed by computing the $\hat{R}$ statistic (Gelman and Rubin, 1992).

Finally, for the MSLE we follow Train (2009) and Moffatt (2016) and we estimate the models with the help of simulation. As mentioned before, in this random-coefficient model, the behavioural parameters for a given subject are fixed and they vary across the experimental population according to a distribution (usually assumed Normal). Assuming that a parameter θ is drawn from a distribution with density $g(\theta)$, for a set of I choices, the likelihood of subject's n choices is given by:

$$LL(\theta) = \int \left[\prod_{i=1}^{I} P_{ni}(A_i)^{y_{ni}} \times (1 - P_{ni}(A_i)^{1-y_{ni}}) g(\theta) \right] d\theta \tag{7}$$

and the total log-likelihood is given by the sum of the logarithm of (7) across all subjects. The parameter θ is distributed over subjects according to the density function $g(\theta)$, and is known as the subject-specific random effect. The variation in θ captures the between-subject heterogeneity. When there are more than one parameters θ, the distribution $g(\theta)$ is a multivariate distribution and the integral is multidimensional. Therefore, the challenge for the estimation method is how to evaluate the integral in (7), since there is no analytical solution. In our analysis, we resort to simulation to approximate the integral, using Maximum Simulated Likelihood Estimation techniques. We use 100 Halton draws per subject. Following Conte et al. (2011), we assume the stochastic parameter ξ to be constant. For the Expected Utility model we therefore estimate 5 parameters, the mean and standard deviation of the risk coefficient r, the precision parameter ξ and the gender effects for both parameters.

3.3 Results

Table 2 reports the estimates from the three inference methods. The first column reports the results from the MLE, the middle from the MSLE and the last one from the HB model. For each parameter θ, we report the point estimate for the MLE, the mean of the distribution μ_θ for the MSLE, and the mode of the posterior distribution for the HB. The standard errors are reported in the Table, with the exception of the HB model where the standard deviation of the posterior distribution of each parameter is reported instead. The statistical significance is based on the respective confidence intervals (credible intervals for the HB).

Table 2: Estimates using the Holt and Laury (2002) data.

	MLE	MSLE	HB
r	0.289^{***}	0.265^{***}	0.292^{***}
s.e.	0.022	0.021	0.033
r_{FEMALE}	0.103^{***}	0.103^{***}	0.103^{**}
s.e.	0.034	0.021	0.050
σ_r	-	0.192	-
s.e.	-	0.000	-
ξ	0.253^{***}	0.184^{***}	0.086^{***}
s.e.	0.014	0.016	0.014
ξ_{FEMALE}	0.050^{**}	0.136	0.041
s.e.	0.023	0.193	0.333

The Table reports estimates from all three inference methods: Maximum Likelihood Estimation (MLE), Maximum Simulated Likelihood Estimation (MSLE) and Hierarchical Bayesian (HB). For each parameter θ, the Table reports the point estimate for the MLE, the mean of the distribution μ_θ for the MSLE, and the mode of the posterior distribution for the HB. Standard errors are reported (standard deviation for the HB). *p<0.1; **p<0.05; ***p<0.01

In all cases, the risk coefficient is positive and statistically significant, indicating risk averse preferences for all subjects. The coefficient of risk aversion ranges between 0.265 and 0.292 between the three inference methods, what Holt and Laury (2002) characterise as "slightly risk averse". Focusing on the gender effects parameters, the coefficient is positive and statistically significant in all three cases, and remarkably at the same magnitude of 0.103. Finally, focusing on the precision parameter ξ, the effect of introducing more flexible inference methods to its magnitude, is apparent. The estimate of ξ using MLE is equal to 0.253 which is quite large compared to the other two methods. Since there is an inverse relationship between the size of ξ and the estimated noise (the lower the ξ the higher the precision) a larger estimate of ξ indicates issues with overfitting. As the inference methods become more flexible, the estimate of ξ

becomes smaller, indicating more precise and less noisy estimates. The main conclusion from this analysis, is that by ignoring the between-subject heterogeneity, and estimating a model assuming a basic level of heterogeneity, as in the case of the MLE estimation, it is possible to detect the existence of gender differences, regardless of which estimation method is adopted. In what follows, we explore whether this result can be generalised when the complexity of the model increases. Filippin and Crosetto (2016) extend their analysis and investigate whether relaxing the expected utility assumption, has an effect to the inferred gender differences. By introducing a probability weighting function and non-expected utility preferences, they estimate a structural specification, using MLE, and show that the gender differences in the risk coefficient disappear, and they appear in the probability weighting parameter. As the original Holt and Laury (2002) task was not developed with non-expected utility preferences in mind, in the next section we repeat the same analysis as above, using data from an experiment which was particularly developed to identify risk preferences, stemming from both the curvature of the utility function and the shape of the probability weighting function.

4 Risk Preferences and Rank Dependent Utility

Motivated by the Allais paradox, a vast theoretical and experimental literature emerged, challenging the assumption of expected utility preferences (see Starmer, 2000 for a review of non-EU theories; Camerer, 1995 for an early discussion of the experimental work; and Hey, 2014 for a more recent review). In this section, we focus on one of the most influential alternatives to EU, the Quiggin (1982) Rank Dependent Utility model (RDU) which later led to the modification of the Original Prospect Theory model and the development of the Tversky and Kahneman (1992) Cumulative Prospect Theory model (which we explore in the next section). In the RDU model, attitudes towards risk are characterised by both the curvature of the utility function, and the shape of the probability weighting function, while there is evidence that the two components are not strongly correlated (Qui and Steiger 2011; Toubia et al. 2013). Therefore, given the extensive empirical evidence of the existence of non-EU preferences, it is crucial to take both components into consideration, when one investigates the existence of gender differences in risk preferences. We do so by using the data from Baillon et al. (2020).

4.1 Decision Task and Data

Objective of this experiment was to identify the reference point that subjects are using when they make choices under risk. Each experimental task involved a choice between two paired lotteries again, A and B. An optimal design was employed to construct the questions of the experiment in a way that they would satisfy the following 5 criteria: (1)the questions must be diverse in terms of number of outcomes and magnitudes of probabilities involved, (2)the questions within each choice must have nonmatching maximal or minimal outcomes, (3) the questions must be diverse in terms of relative positioning in the outcome space, (4) they must have similar expected value to avoid trivial or statistically noninformative choice situations, and; (5) they must be "orthogonal" in some sense to maximise statistical efficiency. The number of the outcomes within each lottery varied between tasks, from 2 to 4 outcomes, all in the gains domain (strictly positive). An example of a task is provided below:

$$A = \begin{cases} 135, & \text{with probability } 0.55 \\ 290, & \text{with probability } 0.35 \\ 329, & \text{with probability } 0.10 \end{cases} \quad B = \begin{cases} 159, & \text{with probability } 0.05 \\ 259, & \text{with probability } 0.55 \\ 359, & \text{with probability } 0.10 \\ 409, & \text{with probability } 0.30 \end{cases}$$

The order of the tasks was randomised, and there was a total of 70 tasks per subject, with varying payoff and probability levels, generating a rich dataset for structural estimations. There are in total data from 139 subjects (49 females).[3] The experimental population consisted of

[3] For our analysis we use the data from 136 subjects as there were missing data on the gender of 2 subjects, and 1 subject had missing data.

students in Moldova, and the payoffs were expressed in the local currency. To incentivise the experiment, each subject had a one-third chance to be selected among all the subjects, to play out one of their choices for real. The experiment involved high stakes with payoffs up to a week's salary.

4.2 Theoretical Framework and Econometric Specification

As mentioned before, the RDU model consists of two components, the utility function and the probability weighting function, which transform every objective probability p to the decision weight $w(p)$ in the interval $[0,1]$. We again assume a CRRA utility function, while for the probability weighting function, we assume the widely used Tversky and Kahneman (1992) function of the form:

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}} \tag{8}$$

where γ is the probability weighting parameter. The form of the function is inverse-S shaped for $\gamma < 1$, indicating overweighting of low probabilities and underweighting of moderate and high probabilities. To evaluate the RDU of a lottery, we first need to rank the outcomes of the lottery from the best to the worst, such that $x_1 \geq x_2, \cdots, \geq x_n$. The decision weight associated with each outcome is given by:

$$\begin{aligned}
\pi(x_1) &= w(p_1) \\
\pi(x_2) &= w(p_1 + p_2) - w(p_1) \\
\ldots \\
\pi(x_n) &= 1 - w(p_1 + p_2 + \cdots + p_n)
\end{aligned}$$

The RDU of lottery A is then given by

$$RDU(A) = \sum_{n=1}^{N} \pi(x_n) \frac{{x_n}^{1-r}}{1-r} \tag{9}$$

We assume the same stochastic function as in Equation 3, by replacing the expected utility with the corresponding Rank Dependent Utility, and we form the log-likelihood function as in Equation 4 for the MLE estimation. As there are no multiple treatments, we control only for gender differences by introducing a gender dummy for all the parameters (r, γ, ξ, giving in total 6 parameters to estimate).

For the HB model, on top of the specifications for r and ξ, which are exactly the same as in the EU case, we need an additional specification for the γ parameter. This parameter must be positive, with a lower bound equal to 0.279 to ensure the monotonicity of the function. For the MSLE estimation, we need to estimate the parameters of the two distributions for r and γ, namely the means μ_r and μ_γ and their standard deviations σ_μ and σ_γ[4].

4.3 Results

Table 3 reports the results from all the three inference methods. The results are quite similar to what is usually observed in this literature. The estimated risk coefficient r is between 0.360 and 0.480, indicating moderate risk averse preferences, while the estimate for the probability weighting function is equal to 0.586 and 0.621, indicating an inverse-S shape of the function. While these results are quite uniform and the estimates look quite close in terms of magnitude and statistical significance, there are contradictory results regarding the presence of gender effects. Assuming heterogeneity only at the gender level (MLE) fails to capture any kind of effects for any of the parameters, while a same pattern is observed when MSLE is used to

[4] In the framework of MSLE, the coefficient vector θ is assumed to be normally distributed, across the population, with mean equal to a vector b and covariance matrix W. To maintain a manageable number of parameters, we assume that the off-diagonal elements of W are equal to zero and estimate the variance of each distribution. Allowing for correlation between the parameters led to worse performance of the model.

estimate the model. Nevertheless, when HB is used, one can infer that there is a significant difference between males and females in the way objective probabilities are transformed. With an estimate of γ equal to 0.705 (compared to 0.621 for men), it seems that women tend to exhibit lower probability distortion. Again, the effect of the different estimation methods on the precision parameter ξ is similar as in the Expected Utility case (the noise in the estimates decreases when more flexible inference methods are introduced).

Table 3: Estimates using the Baillon et al. (2020) data.

	MLE	MSLE	HB
r	0.479^{***}	0.424^{***}	0.360^{**}
s.e.	0.032	0.038	0.029
r_{FEMALE}	-0.025	-0.030	-0.054
s.e.	0.111	0.203	0.099
σ_r	-	0.291^{**}	-
s.e.	-	0.122	-
γ	0.598^{***}	0.586^{**}	0.621^{***}
s.e.	0.016	0.251	0.029
γ_{FEMALE}	0.050	-0.060	0.084^{*}
s.e.	0.032	0.244	0.049
σ_γ	-	0.943	-
s.e.	-	0.821	-
ξ	0.121^{***}	0.076^{***}	0.083^{***}
s.e.	0.001	0.007	0.010
ξ_{FEMALE}	0.041	-0.003	0.026^{***}
s.e.	0.044	0.013	0.005

The Table reports estimates from all three inference methods: Maximum Likelihood Estimation (MLE), Maximum Simulated Likelihood Estimation (MSLE) and Hierarchical Bayesian (HB). For each parameter θ, the Table reports the point estimate for the MLE, the mean of the distribution μ_θ for the MSLE, and the mode of the posterior distribution for the HB. Standard errors are reported (standard deviation for the HB). *p<0.1; **p<0.05; ***p<0.01

The analysis above provides an example of the implications of ignoring heterogeneity between (as well as within) participants. While a basic MLE estimation provides no evidence of any kind of gender differences, allowing for a more informative approach reveals the existence of such differences. In the next section, we extend our analysis to one of the most important domains of decision theory under risk, that of loss aversion.

5 Risk Preferences and Loss Aversion

In this section we focus on the three components that characterise risk preferences in the losses domain, as these are articulated in the Tversky and Kahneman (1992) Cumulative Prospect Theory (CPT) model. The CPT model adopts a similar approach to the RDU model, on the way it handles monetary payoffs and probabilities, with the additional feature of *loss aversion*, the concept that "losses loom larger than gains". The results from the literature are mixed. Some studies find that women are more less loss averse (see Schmidt and Traub, 2002, Brooks and Zank, 2005), others that males are more loss averse (Booij et al., 2009), others that there is no difference (Harrison and Rutström, 2008), and others with a mixed result (Bouchouicha et al., 2019). As Bouchouicha et al. (2019) argue, currently, there is no consensus of what is the appropriate definition of loss aversion in the literature[5]. Nevertheless, for the sake of the example, we will focus on the CPT definition of loss aversion, while our approach can be extended to alternative definitions.

5.1 Decision Task and Data

To estimate a CPT specification when losses are present, we use the data from Bouchouicha et al. (2019) which is a subset from the data used in Vieider et al. (2015). There are in total observations of almost 3000 subjects, from 30 countries, on decision making under risk and ambiguity, in both the gains and losses domain. As our focus is on small samples, we use only the USA data. This set includes the choices of 95 subjects (47 females) in 12 choice tasks (6 in the gains domain, 5 in the losses domain, and 1 in the mixed domain to identify the loss aversion parameter). While there are available data on a larger set of risky tasks (28 tasks), we follow Bouchouicha et al. (2019) and use only the smaller subset for two reasons: (1) this set of

[5] See Schmidt and Zank (2005) for the various definitions of loss aversion.

tasks includes only 50:50 gambles, which allows the estimation of a functional-free probability weighting function, and; (2) estimating a structural model from a small set of observations per participant is one of the strengths of the Hierarchical approach, and this dataset allows to test the limits of this approach.

All tasks are in the form (x, y), representing the prospect of getting the monetary payoff x with probability 50% or y with the residual probability, with x and y being positive, negative or zero, depending on the task (Table 4 lists the 12 tasks). The subject had to express her *certainty equivalent* for each of the tasks. For the mixed domain prospect, the amount l was elicited, that would make the subject indifferent between a 50:50 gamble of (20,l) and the status quo of zero. The experiment was incentivised and an endowment equal to the largest possible loss was provides to the subject, to cover for potential losses.

Table 4: The tasks from Bouchouicha et al. (2019)

Gains	Losses	Mixed
(5,0)	(-5,0)	(20,-l)
(10,0)	(-10,0)	
(20,0)	(-20,0)	
(30,0)	(-20,-5)	
(30,10)	(-20,-10)	
(30,20)		

5.2 Theoretical Framework and Econometric Specification

We model preferences assuming a CPT decision maker. We employ a power utility function as before, of the form:

$$u(x) = \begin{cases} \frac{x^{1-r}}{1-r}, & \text{if } x \geq 0 \\ -\lambda \frac{(-x)^{1-r}}{1-r}, & \text{if } x < 0 \end{cases}$$

with r the risk coefficient, and λ the parameter of loss aversion. The status quo of zero is assumed as a reference point. We assume a common parameter for r for gains and losses, for two reasons: (1) there is extensive empirical evidence of no difference between the two domains (see Fox and Poldrack 2009), and; (2) to avoid any potential identification issues of the loss aversion parameter (see Wakker, 2010). As mentioned before, since only 50:50 gambles are used in the analysis, there is no need to specify a functional form for the probability weighting function. Therefore, we introduce two parameters to estimate, w_g and w_l, which represent the probability weighting for gains and losses respectively. Summarising, a prospect $L = (x, y)$ can be evaluated as:

$$U(L) = w_s u(x) + (1 - w_s) u(y)$$

with $s \in \{g, l\}$, while for the mixed prospect $L = (x, l)$, the prospect is evaluated as:

$$U(L) = w_g u(x) + w_l u(l)$$

The certainty equivalent $\hat{ce}$ for a prospect L is then given by:

$$\hat{ce} = u^{-1}[w_s u(x) + (1 - w_s) u(y)]$$

To form the likelihood function we need a different approach to the one used in the previous sections. In particular we assume that a decision maker states her certainty equivalent with some noise. The observed certainty equivalent of a subject in a task i is equal to $ce_i = \hat{ce} + \varepsilon_i$, where $\hat{ce}$ is the theoretical optimal certainty equivalent, for a set of behavioural parameters, and $\varepsilon \sim \mathcal{N}(0, \xi^2)$ with ξ being the standard deviation of the Fechner error (see Hey and Orme, 1994). We assume that this error is domain-specific (for mixed gambles we use the error for losses) and we also take into consideration a *contextual* error Wilcox 2011 by making the parameter ξ to be dependent on the difference between the best and the worst outcome of each prospect. That is, $\xi_i = \xi|x_i - y_i|$. The loglikelihood function for N subjects and I tasks is then given by:

$$LL(\theta) = \sum_{n=1}^{N}\sum_{i=1}^{I} \ln[\psi(\theta_n, L_i)] \tag{10}$$

with θ a vector of behavioural parameters to be estimated, L_i a task i and ψ the contribution to the likelihood function given by:

$$\psi(\theta_n, L_i) = \phi\left(\frac{\hat{ce}_{ni} - ce_{ni}}{\xi_{nis}}\right)$$

where ϕ is the standard normal density function. For the MLE estimation, we follow Bouchouicha et al. (2019) and we assume heterogeneity of the parameters at the gender level, and the domain level for the decision weights and the precision parameters. We need to estimate 12 parameters in total ($r, \lambda, w_g, w_l, \xi_g, \xi_l$ along with the controls for gender).

For the HB model, the specification of the likelihood function remains the same as in the MLE case. We specify distributions for the six parameters as above, with the decision weights constrained to the interval $[0, 1]$ and the loss aversion parameter to the interval $[0, 10]$, while for the MSLE, we estimate the parameters of the distributions for the risk attitude, the loss aversion and the probability weighting for gains and losses.

5.3 Results

Table 5 reports the estimates from the three inference methods. Three points are worth to mention: (1) there is significant loss aversion in this sample with a λ parameter statistically significant ranging between 1.596 and 1.672, (2) the risk coefficient is not statistically different than zero for the MLE and the HB cases, indicating a linear utility function, (3) the probabilities in the gains domain are distorted more that the probabilities in the losses domain (for instance, the decision weight of 0.5 is estimated to be 0.426 for gains and 0.478 for losses in the MLE case), and; (4) the control coefficient for gender differences is insignificant for all the major parameters of interest, in the MLE case with the exception of the noise parameter. Once again, using MLE techniques, one can conclude that there are no gender differences in the way females and males perceive monetary outcomes, transform probabilities to decision weights or perceive losses. Focusing on the more flexible methods of MSLE and HB, two points are interesting. First, the estimates of the mean, for all the parameters, are remarkably close between the two methods reinforcing the result of Huber and Train (2001). Nevertheless, when gender effects are considered, while both methods find differences in the loss aversion parameters between the two groups, the MSLE methods fails to detect any gender effects in the key parameter or risk attitude. A potential explanation for this result could be the larger estimate of the precision parameter (a lower value indicates more precise estimates).

In this additional example, we provide further evidence that as the model complexity increases, by ignoring heterogeneity at the subject level, it may lead to incorrect inference regarding the difference between different demographic groups. Both methods that allow for this kind of heterogeneity (MSLE and HB) managed to detect the existence of such effects. Nevertheless, the results are not uniform. To identify which method is the most appropriate to use, in the next section we report the results of an extensive simulation exercise were we compare the performance of each of the methods.

6 Exploring the Advantages of HB Modelling

In the previous sections, we have shown that the identification on gender effects largely depends on the adopted inference method. We have provided a rigorous comparison of the representative agent model against two alternative methods that allow for extensive behavioural

Table 5: Estimates using the Bouchouicha et al. (2019) data.

	MLE	MSLE	HB
λ	1.596***	1.672***	1.615***
s.e.	0.099	0.112	0.113
λ_{FEMALE}	0.321	0.430***	0.398***
s.e.	0.211	0.141	0.136
σ_λ	-	0.467***	-
s.e.	-	0.044	-
r	-0.133	0.146**	-0.026
s.e.	0.07	0.061	0.019
r_{FEMALE}	-0.011	-0.061	-0.082^{*}
s.e.	0.115	0.083	0.041
σ_r	-	0.000	-
s.e.	-	0.053	-
w_g	0.426***	0.433***	0.444***
s.e.	0.019	0.023	0.017
$w_{g_{FEMALE}}$	-0.009	-0.032	-0.037
s.e.	0.031	0.031	0.056
σ_{w_g}	-	0.38***	-
s.e.	-	0.051	-
w_l	0.478***	0.467***	0.501***
s.e.	0.019	0.022	0.010
$w_{lFEMALE}$	0.007	-0.010	-0.011
s.e.	0.031	0.031	0.037
σ_{w_l}	-	0.436***	-
s.e.	-	0.048	-
ξ	0.173***	0.178***	0.097***
s.e.	0.007	0.012	0.012
ξ_{FEMALE}	0.028***	0.000	0.051
s.e.	0.012	0.000	0.261
ξ_l	0.150***	0.112	0.063***
s.e.	0.006	0.008	0.007
$\xi_{lFEMALE}$	0.033***	0.000	0.055
s.e.	0.011	0.000	0.311

The Table reports estimates from all three inference methods: Maximum Likelihood Estimation (MLE), Maximum Simulated Likelihood Estimation (MSLE) and Hierarchical Bayesian (HB). For each parameter θ, the Table reports the point estimate for the MLE, the mean of the distribution μ_θ for the MSLE, and the mode of the posterior distribution for the HB. Standard errors are reported (standard deviation for the HB). *p<0.1; **p<0.05; ***p<0.01

heterogeneity, even when the available sample size is small. Given that all three methods result is quantitatively different estimates, it raises the question of which method should one adopt. In this section we aim to provide an answer to this question, by means of an extensive Monte Carlo simulation exercise. Several studies have focused on the comparison between classical and Bayesian estimates, providing support on the latter (see for example Nilsson et al., 2011 or Gao et al., 2020). Here we repeat a similar exercise, suitably adapted to our objective of identifying gender differences in the elicited behaviour.

The main goal of this simulation study is two-fold. First, we want to confirm whether all estimation procedures are able to accurately recover the true parameter values from simulated data. Secondly, we test whether the inference methods under consideration, are equally efficient in detecting gender effects. To make the simulation as general as possible, we focus on the Bouchouicha et al. (2019) design and the CPT model, which satisfies the conditions for which researchers usually resort to pool their data (a relatively large number of parameters to estimate using a relatively low number of data points per subject). For our exercise, we simulate data of 100 subjects which we then estimate using each of the three inference methods: MLE, MSLE and HB.

We assume that gender differences exist only in two of the model's parameters, the coefficient of loss aversion and the risk coefficient[6]. The parameters used in the simulation, are normally distributed across the experimental population with mean θ_n and standard deviation σ_θ. In the simulation we set the gender difference in the risk coefficient to be small but significant (mean of 0.500 for males and 0.600 for females) with a standard deviation equal to 0.05[7]. The loss aversion is set to 1.648 for males and 2.013 for females[8] with a standard deviation of 0.100. The probability weighting coefficient for gains w_g is set equal to 0.540 while the

[6] We make this assumption in order to keep the simulation as simple as possible. Of course this analysis can be extended to any of the parameters of the model (i.e. probability weighting function, noise coefficient) since empirically, gender effects are observed in all components of preferences.

[7] We confirmed that the statistical significance of the two distributions is indeed significant based on a two-sided t-test (p<0.000).

[8] Since we transform the parameters to be drawn from a log-normal distribution, the values of loss aversion correspond to $\exp(0.500)$ for men, and $\exp(0.700)$ for women.

probability coefficient for losses w_l is set equal to 0.510. We assume no heterogeneity by setting the standard deviation equal to 0 for the weighting parameters and we also assume a common Fechnerian error for gains and losses. We conducted the simulation for three different levels of noise by setting the value of the error term equal to 0.130 (low noise), 0.150 (medium noise) and 0.200 (high noise). We report the results of the medium noise specification as they are the most representative[9]. For each simulation, we generate the data of the 100 artificial subjects by drawing parameters from the relevant distributions that were described above. This dataset was then estimated using each of the methods. Table 6 reports the results of 100 simulations. In particular, we report the mean and the standard deviation of the point estimates, in the case of MLE, the mean of the distribution means in the case of MSLE, and the mean of the posterior means of the distributions in the case of HB.

[9] Bouchouicha et al., 2019 using this dataset, estimate the noise parameter to be equal to 0.170. For our simulations, we are using a noise parameter of 0.150, which is in the middle of the interval between the low noise parameter (0.130) and the empirically observed parameter (0.170). We delegate the estimates from the low and high noise simulations to the online Appendix (see Tables A1 and A2).

Table 6: Mean and standard deviations of the parameters.

Parameter	True value	MLE	MSLE	HB
λ	1.648	1.575	1.637	1.657
s.e.	-	0.056	0.062	0.067
λ_{FEMALE}	0.365	0.698	0.374	0.389
s.e.	-	0.132	0.118	0.129
σ_λ	0.100	-	0.081	-
s.e.	-	-	0.046	-
r	0.500	0.538	0.500	0.493
s.e.	-	0.019	0.031	0.025
r_{FEMALE}	0.100	0.064	0.105	0.108
s.e.	-	0.034	0.031	0.032
σ_r	0.050	-	0.046	-
s.e.	-	-	0.014	-
w_g	0.540	0.559	0.543	0.536
s.e.	-	0.013	0.014	0.014
w_l	0.510	0.528	0.510	0.510
s.e.	-	0.014	0.013	0.014
ξ	0.150	0.153	0.150	0.148
s.e.		0.007	0.005	0.006

The Table reports estimates from the simulation exercise on the three inference methods : Maximum Likelihood Estimation (MLE), Maximum Simulated Likelihood Estimation (MSLE) and Hierarchical Bayesian (HB), for the medium level of noise. For each parameter θ, the Table reports the mean of the point estimates, in the case of MLE, the mean of the distributions in the case of MSLE, and of the posterior mean of the distributions in the case of HB. Standard deviations in parentheses.

We first focus on the parameter recovery performance of each of the methods. The first column of the Table reports the true values of the coefficients that were used in the simulation. Compared to the true value, it is apparent that the MLE estimates have the worst performance in terms of precision. First, most of the parameters deviate significantly from the true value, compared to the other two methods. Then, in terms of gender effects, there is significant overestimation of the difference in loss aversion where the parameter is estimated to be almost twice the true value (0.698 compared to the true value of 0.365) while there is underestimation of the difference in the risk coefficient (0.064 compared to the true value of 0.100). As far as the MLE and MSLE estimates are concerned, both are remarkably close to each other and both have recovered the true parameters with quite high precision. The first conclusion from this simulation exercise is that if one is interested in the mean values of the parameters of different groups, then both MSLE and HB are equally good in recovering unbiased parameter values compared to the MLE.

We now turn to the identification of gender effects. For each of the simulations, we generate the 95% confidence interval (credible interval in the case of HB) to test the statistical significance of the estimate. When we focus on the gender effect for the risk coefficient, the MLE estimate is statistically significant for 55% of the simulations, the MSLE for 66% while the HB for 96%. Similarly, when we focus on the loss aversion parameter the MLE estimate is statistically significant for 53% of the simulations, the MSLE for 67% while the HB for 89%. Table 7reports the frequency with which statistically significant gender effects were detected, for each of the three inference methods, and for each of the three levels of noise (low, medium and high). The Table confirms the pattern that higher levels of noise lead to lower detection levels of gender effects, with MLE having the worst performance, HB the best, and MSLE in the between.

Table 7: Identification of gender effects.

	r_{FEMALE}		
	$\xi = 0.130$	$\xi = 0.150$	$\xi = 0.200$
MLE	62%	55%	36%
MSLE	82%	66%	59%
HB	98%	96%	76%
	λ_{FEMALE}		
	$\xi = 0.130$	$\xi = 0.150$	$\xi = 0.200$
MLE	63%	53%	37%
MSLE	81%	67%	50%
HB	97%	89%	66%

The Table reports the rate of success of each inference method to identify gender effects for each of the three levels of noise, for the gender specific parameter for risk attitude (r_{FEMALE}) and loss aversion (λ_{FEMALE}) are statistically significant, at the 5% level.

Our results mirror the conclusions of Huber and Train (2001). In this study the authors compare classical and Bayesian estimates by providing a comparison between MSLE and HB. They show that both methods result in virtually equivalent conditional estimates of the parameters. Then, they provide a list of differences between the two methods including (1) the difficulty of MSLE to locate the maximum of the likelihood function; (2) the computational burden that the variance-covariance matrix poses to the estimation of the MSLE parameters, and; (3) the identification issues that the classical approach faces compared to the Bayesian estimation. Our simulation shows that when the identification of differences between different populations is the objective, then HB is the clear winner as the most appropriate inference method. This result can be attributed to the way each of the methods handles uncertainty in the estimates and the fact that the estimate of the unobserved heterogeneity in the MSLE estimates is much noisier (larger standard errors) compared to the HB ones.

To investigate the role of the sample size in the detection of gender effects, we ran some additional simulations for the MSLE methods, varying the sample size. Assuming a fixed level of noise ($\xi = 0.150$), we repeated the simulation exercise for $N = 200$ and $N = 500$ and again we report the rate of success to identify gender effects for the risk attitude (r_{FEMALE}) and the loss aversion (λ_{FEMALE}) gender specific parameters[10]. When the sample size is equal to 200, the risk (loss aversion) coefficient is significant for 89% (90%) of the simulations, while when the sample size increases to 500, the risk (loss aversion) coefficient is significant for 93% (94%) of the simulations. This analysis further highlights the advantages of the HB modelling since this inference method needs only half of the sample that MSLE needs in order to achieve the same detection rate of success in the case of loss aversion, while it needs only one fifth of the sample that MSLE needs, to reach the same success rate, in the case of the risk coefficient.

7 Concluding remarks

In this study, we focus on gender differences and compare the inference made by three econometric methods, Maximum Likelihood Estimation, Maximum Simulated Likelihood Estimation and Hierarchical Bayesian modelling, on three representative domains of risk preferences. We show that when all the data are assumed to come from a representative agent, and assume heterogeneity (gender differences or any other demographic differences) at a very basic level (e.g. all black females have the same level of loss aversion), valuable information might be ignored, and therefore, distorted conclusions may be drawn. Nevertheless, opting for a more flexible approach, and taking into consideration both the individual variation and the population-level characteristics, the inference about individual risk preferences is massively improved, and significant differences are captured.

[10] The estimates are delegated to the online Appendix (see Table B1). There, it can be seen that as the sample size increases, the standard errors decrease, which allows for better identification of the effects.

In particular, we compare the representative agent modelling approach, to two more flexible and informative methods of parameter estimation that allow one to simultaneously make inferences at both the individual subject and the experimental population level. We compare the frequentist and the Bayesian methods, by analysing the data using Maximum Simulated Likelihood Estimation techniques (MSLE), as well as Hierarchical Bayesian (HB) econometric modelling. We use data from three representative studies on decision making under risk and we study Expected Utility preferences, for a simple analysis of risk attitudes, Rank Dependent Utility preferences, to incorporate probability weighting, and Cumulative Prospect Theory, to investigate loss averse behaviour. We show that by ignoring heterogeneity at the subject level, it may lead to incorrect inference regarding the difference between distinct demographic groups.

Recent research on Hierarchical Bayesian modelling has shown that MLE estimates are both susceptible to overfitting and dominated by outliers (Nilsson et al., 2011, Murphy and ten Brincke, 2018), while Bayesian modelling improves the robustness of the estimation, by shrinking the parameters towards the group's mean. This method allows the robust estimation of preferences, and it is particularly useful, especially when one has a limited number of data points from each subject, as is often the case with field studies, or when additional tasks are used, along with the main experimental design, to control for particular preferences. With the aid of an extensive simulation exercise, we show that Bayesian methods are better placed to capture differences between groups, and this result can be attributed to the way that each of the methods handles uncertainty in the estimates.

In this study, we do not argue in favour of any particular preference functional or model, nor we claim that there is a uniform pattern of gender differences. In our analysis, we opted for the models and the preference functionals that are often assumed in this literature. These models acted as "vehicles" to illustrate the machinery behind both estimation techniques, and this approach could be extended to any alternative model. Our main objective is to warn researchers on the dangers of small sample datasets and ignoring heterogeneity of the subjects. Of course this method could be extended to other important fields of decision making such as ambiguity preferences, time preferences or social preferences. Even more, as Gao et al. (2020) highlight, HB methods are particularly useful when one is interested in joint estimation of perhaps non-correlated preferences (e.g. joint estimation of risk and time preferences) where the need of robust estimates is important at the individual level.

References

Baillon, A., Bleichrodt, H., and Spinu, V. (2020). Searching for the Reference Point. *Management Science*, 66(1):93–112.

Balcombe, K. and Fraser, I. (2015). Parametric Preference Functionals under Risk in the Gain Domain: A Bayesian Analysis. *Journal of Risk and Uncertainty*, 50(2):161–187.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

Booij, A., Van praag, B., and Kuillen, G. (2009). A Parametric Analysis of Prospect Theory's Functionals for the General Population. *Theory and Decision*, 68:115–148.

Bouchouicha, R., Deer, L., Eid, A., McGee, P., Schoch, D., Stojic, H., Ygosse-Battisti, J., and Vieider, F. (2019). Gender effects for loss aversion: Yes, no, maybe? *Journal of Risk and Uncertainty*, 59:171–184.

Brooks, P. and Zank, H. (2005). Loss Averse Behavior. *Journal of Risk and Uncertainty*, 31:301–325.

Camerer, C. (1995). Individual Decision Making. In Kagel, J. and Roth, A., editors, *The Handbook of Experimental Economics*. Princeton University Press.

Charness, G. and Gneezy, Y. (2012). Strong Evidence for Gender Difference in Risk Taking. *Journal of Economic Behavior & Organization*, 83(1):50–58.

Conte, A., Hey, J. D., and Moffatt, P. G. (2011). Mixture models of choice under risk. *Journal of*

Econometrics, 162(1):79 – 88.

Croson, R. and Gneezy, Y. (2009). Gender Differences in Preferences. *Journal of Economic Literature*, 47(2):448–474.

Eckel, C. C. and Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. volume 1 of *Handbook of Experimental Economics Results*, pages 1061 – 1073. Elsevier.

Ferecatu, A. and Önçüler, A. (2016). Heterogeneous Risk and Time Preferences. *Journal of Risk and Uncertainty*, 53(1):1–28.

Filippin, A. and Crosetto, P. (2016). A Reconsideration of Gender Differences in Risk Attitudes. *Management Science*, 62(11):3138–3160.

Fox, C. R. and Poldrack, R. A. (2009). Chapter 11 - Prospect Theory and the Brain. In Glimcher, P. W., Camerer, C. F., Fehr, E., and Poldrack, R. A., editors, *Neuroeconomics*, pages 145 – 173. Academic Press, London.

Gao, X., Harrison, G., and Tchernis, R. (2020). Estimating Risk Preferences for Individuals: A Bayesian Approach. CEAR Working Paper 2020-15.

Gelman, A. and Rubin, D. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472.

Harrison, G. and Rutström, E. (2008). Expected Utility Theory and Prospect Theory: one Wedding and a Decent Funeral. *Experimental Economics*, 12(2):133.

Hey, J. (2014). Choice Under Uncertainty: Empirical Methods and Experimental Results. In Machina, M. and Viscusi, W., K., editors, *Hanbook of the Economics of Risk and Uncertainty*, pages 809–850. Eslevier B.V.

Hey, J. and Orme, C. (1994). Investigating Generalizations of Expected Utility Theory Using Experimental Data. *Econometrica*, 62(6):1291–1326.

Holt, C. and Laury, S. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, 92(5):1644–1655.

Holt, C. A. and Laury, S. K. (2014). Chapter 4 - assessment and estimation of risk preferences.

In Machina, M. and Viscusi, K., editors, *Handbook of the Economics of Risk and Uncertainty*, volume 1 of *Handbook of the Economics of Risk and Uncertainty*, pages 135 – 201. North-Holland.

Huber, J. and Train, K. (2001). On the Similarity if Classical and Bayesian Estimates of Individual Mean Partworths. *Marketing Letters*, 12(3):259–269.

Jacquement, N. and L'Haridon, O. (2018). *Experimental Economics: Methods and Applications*. Cambridge University Press.

Moffatt, P. (2016). *Experimetrics*. Macmillan Palgrave.

Murphy, R. and ten Brincke, R. (2018). Hierarchical Maximum Likelihood Parameter Estimation for Cumulative Prospect Theory: Improving the Reliability of Individual Risk Parameter Estimates. *Management Science*, 64:308–326.

Nilsson, H., Rieskamp, J., and Wagenmakers, E.-J. (2011). Hierarchical Bayesian Parameter Estimation for Cumulative Prospect Theory. *Journal of Mathematical Psychology*, 55:84–93.

Plummer, M. (2017). JAGS Version 4.3.0 User Manual. Technical report.

Qui, J. and Steiger, E. (2011). Understanding the two Components of Risk Attitudes: an Experimental Analysis. *Management Science*, 57:193–199.

Quiggin, J. (1982). A Theory of Anticipated Utility. *Journal of Economic Behavior and Organization*, 3(4):323–343.

Rouder, J. and Lu, J. (2005). An Introduction to Bayesian Hierarchical Models with an Application in the Theory of Signal Detection. *Psychonomic Bulletin & Review*, 55:84–93.

Schmidt, U. and Traub, S. (2002). An Experimental Test of Loss Aversion. *Journal of Risk and Uncertainty*, 25(3):233–249.

Schmidt, U. and Zank, H. (2005). What is Loss Aversion. *Journal of Risk &Uncertainty*, 30(2):157–167.

Stahl, D. (2014). Heterogeneity of Ambiguity Preferences. *The Review of Economics and Statistics*, 96(5):609–617.

Starmer, C. (2000). Developments in Non-expected Utility Theory: The Hunt for a Descriptive

Theory of Choice under Risk. *Journal of Economic Literature*, 38(2):332–382.

Stott, H. (2006). Cumulative Prospect Theory's Functional Menagerie. *Journal of Risk and Uncertainty*, 32(2):101–130.

Toubia, O., Johnson, E., Evgeniou, T., and Delquie, P. (2013). Dynamic Experiments for Estimating Preferences: an Adaptive Method of Eliciting Time and Risk Parameters. *Management Science*, 59:613–640.

Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition.

Tversky, A. and Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Vieider, F. M., Lefebvre, M., Bouchouicha, R., Chmura, T., Hakimov, R., Krawczyk, M., and Martinsson, P. (2015). Common Componetns of Risk and Uncertainty Attitudes Across Contexts and Domains: Evidence from 30 Countries. *Journal of the European Economic Association*, 13(3):421–452.

von Gaudecker, H., Van Soest, A., and Wengström, E. (2011). Heterogeneity in Risky Choice Behaviour in a Broad Population. *American Economic Review*, 101(2):664–694.

Wakker, P. (2010). *Prospect Theory*. Cambridge University Press.

Wilcox, N. (2011). Stochastically more Risk Averse: A Contextual Theory of Stochastic Discrete Choice under Risk. *Journal of Econometrics*, 162(1):89 – 104.

Xie, Z., Page, L., and Hardy, B. (2017). Investigating Gender Differences under Time Pressure in Financial Risk Taking. *Frontiers in Behavioural Neuroscience*, 11(246).

Appendix A Monte Carlo Simulation

This Appendix presents further results of the simulation in Section 6 when the noise is low ($\xi = 0.130$) in Table A1, and when it is high ($\xi = 0.200$) in Table A2.

Table A1: Mean and standard deviations of the parameters.

Parameter	True value	MLE	MSLE	HB
λ	1.648	1.576	1.640	1.648
s.e.	-	0.050	0.035	0.035
λ_{FEMALE}	0.365	0.696	0.368	0.379
s.e.	-	0.117	0.057	0.055
σ_λ	0.100	-	0.088	-
s.e.	-	-	0.036	-
r	0.500	0.537	0.500	0.497
s.e.	-	0.017	0.023	0.021
r_{FEMALE}	0.100	0.065	0.105	0.107
s.e.	-	0.030	0.028	0.027
σ_r	0.050	-	0.047	-
s.e.	-	-	0.012	-
w_g	0.540	0.559	0.540	0.536
s.e.	-	0.012	0.050	0.022
w_l	0.510	0.527	0.509	0.506
s.e.	-	0.012	0.048	0.023
ξ	0.130	0.134	0.130	0.128
s.e.	-	0.005	0.005	0.043

The Table reports estimates from the simulation exercise on the three inference methods : Maximum Likelihood Estimation (MLE), Maximum Simulated Likelihood Estimation (MSLE) and Hierarchical Bayesian (HB), for the low level of noise (0.13). For each parameter θ, the Table reports the mean of the point estimates, in the case of MLE, the mean of the distributions in the case of MSLE, and of the posterior mean of the distributions in the case of HB. Standard deviations in parentheses.

Table A2: Mean and standard deviations of the parameters.

Parameter	True value	MLE	MSLE	HB
λ	1.648	1.572	1.640	1.670
s.e.	-	0.072	0.051	0.054
λ_{FEMALE}	0.365	0.705	0.371	0.398
s.e.	-	0.173	0.081	0.089
σ_{λ}	0.100	-	0.072	-
s.e.	-	-	0.057	-
r	0.500	0.539	0.500	0.489
s.e.	-	0.024	0.031	0.032
r_{FEMALE}	0.100	0.062	0.106	0.111
s.e.	-	0.045	0.040	0.041
σ_r	0.050	-	0.043	-
s.e.	-	-	0.022	-
w_g	0.540	0.560	0.540	0.534
s.e.	-	0.018	0.077	0.036
w_l	0.510	0.528	0.509	0.502
s.e.	-	0.018	0.075	0.038
ξ	0.200	0.202	0.199	0.198
s.e.		0.008	0.006	0.042

The Table reports estimates from the simulation exercise on the three inference methods : Maximum Likelihood Estimation (MLE), Maximum Simulated Likelihood Estimation (MSLE) and Hierarchical Bayesian (HB), for the high level of noise (0.20). For each parameter θ, the Table reports the mean of the point estimates, in the case of MLE, the mean of the distributions in the case of MSLE, and of the posterior mean of the distributions in the case of HB. Standard deviations in parentheses.

Appendix B Sample size

Table B1 reports the results of the simulation exercise when the size sample of 100 increases by a factor of 2 (N=200) and 5 (N=200). All the parameter values

Table B1: Mean and standard deviations of the parameters.

	True value	N=100	N=200	N=500
λ	1.648	1.637	1.640	1.648
s.e.	-	0.062	0.029	0.021
λ_{FEMALE}	0.365	0.374	0.370	0.368
s.e.	-	0.118	0.046	0.031
σ_{λ}	0.100	0.081	0.088	0.093
s.e.	-	0.046	0.029	0.016
r	0.500	0.500	0.503	0.504
s.e.	-	0.031	0.017	0.015
r_{FEMALE}	0.100	0.105	0.099	0.096
s.e.	-	0.031	0.022	0.017
σ_r	0.050	0.046	0.048	0.050
s.e.	-	0.014	0.008	0.005
w_g	0.540	0.543	0.542	0.542
s.e.	-	0.014	0.041	0.029
w_l	0.510	0.510	0.511	0.511
s.e.	-	0.013	0.043	0.029
ξ	0.150	0.150	0.150	0.150
s.e.	-	0.005	0.003	0.002

The Table reports estimates from the simulation exercise using Maximum Simulated Likelihood Estimation (MSLE) for three levels of sample size (N) namely 100, 200 and 500. Standard deviations in parentheses.