The nature of heterogeneity in risk attitudes: evidence from a high-stake experiment *

Thibault RICHARD

Ph.D. Student at ENS Paris-Saclay

June 2022

*I am thankful to Aurélien Baillon, who gave me the permission to use the rich dataset drawn from his paper, written with Han Bleichrodt and Vitalia Spinu, "Searching for the reference point," and to Daniel Zizzo, Julien Monardo, Thomas Vendryes, Francois Pannequin, Anne Corcos, and the participants at the EEA-ESEM Virtual Conference 2021 for their helpful comments. I am also thankful to Glenn Harrison, Lionel Page, Nikolaos Georgantzis, Serge Blondel, and Louis Levy-Garboua for their feedback on an early and incomplete version of this project.

Abstract

While it is commonly known that individuals differ widely in their risk attitudes, there is little understanding about the nature and origin of this heterogeneity. Following recent advancements in Bayesian statistics in the field of decision under risk, we provide evidence that all heterogeneity in risk attitudes is driven by the heterogeneity in probability weighting. Our results indicate that participants in a high-stake experiment exhibit a homogeneous cognitive treatment of wealth but a heterogeneous treatment of probabilities. More specifically, most of the individual heterogeneity in risk attitudes seems to result from differences in the elevation of the probability weighting function. As an alternative to existing risk-preference elicitation procedures, we propose a *targeted heterogeneity approach* wherein only some elements in the models are estimated at an individual level. We argue that measurement procedures applied entirely at the subject level could mistake pure noise for heterogeneity in the distribution of some model parameters. This paper also proposes several methodological innovations to demonstrate its results, such as the use of both parametric and non-parametric methods for estimating risk preferences.

Keywords: Risk Preferences, High-Stake Experiment, Probability Weighting, Bayesian Statistics.

1 Introduction

Since its origin, the elicitation of risk preferences has been a key component in behavioral economics. Up until the mid-1990s, most empirical studies focused on the ability of dominant models to predict some paradoxes or anomalies (e.g., Kahneman and Tversky, 1979). However, around the early 2000s, the focus changed to the elicitation and the test of the models on their structural or functional forms.[1] One of the main findings of this literature, and probably also one of the most robust, is that individuals considerably differ in their risk attitudes. The measurement of heterogeneity is straightforward, and it is typically assumed that all the decision model parameters of each subject are simply different. However, as pointed out for instance by Wilcox (2008), decisions under risk have a strong stochastic component, and the set of individual observations in experimental data is generally too small to make the estimation of risk preferences precise or reliable. Thus, models are prone to overfitting and generate what we could call *spurious heterogeneity* or heterogeneity in the estimates caused only by noise.[2]

So far, there has been little discussion about the origin and nature of heterogeneity in risk attitudes. This paper argues that a systematic understanding of this issue could help to separate pure noise from genuine differences in the decision models of subjects. Using a traditional Rank-Dependent Utility (RDU) model and following a Hierarchical Bayesian Estimation (HBE) approach, we provide evidence that all heterogeneity in risk behavior can be attributed to the heterogeneity in the subjects' Probability Weighting Function (PWF). When we assume a unique utility function among individuals, the predictive power of the RDU model remains identical, even for stakes that are relatively high according to experimental economics standards. Conversely, assuming a universal probability function sharply reduces the predictive accuracy of the same model. Thus, we reconsider the standard elicitation of risk preferences made at an individual level, and we propose the notion of *targeted heterogeneity*, where heterogeneity is driven only by some elements of the models.

[1] Prominent studies include Hey and Orme (1994), Camerer and Ho (1994), Stott, (2006), Harrison and Rutström (2007), and Post et al., (2008). At least two reasons explain this evolution in the literature. First, decision models have become numerous and flexible enough to rationalize many regularities, and this criterion has tended to be less and less discriminant. For instance, most non-expected utility theories proposed since Kahneman and Tversky (1979) explain the Allais paradox. Meanwhile, improvement in the calculation capacity has made model estimation and evaluation easier during the last two decades.

[2] See Nilsson et al. (2011), who demonstrate that regularization methods, by penalizing parameters far from the median estimates of the population, generally give better predictive performances than traditional approaches.

According to our results, decision-makers (DM) could be considered homogeneous in their cognitive treatment of wealth but heterogeneous in their treatment of probabilities. The possible existence of a unique utility function among individuals has many implications. Theoretically, such a hypothesis could make non-Expected Utility (non-EU) models more parsimonious, and, thus, more tractable. Concerning empirical applications, the idea of targeting heterogeneity may considerably facilitate non-EU estimates when the number of individual decisions is low, as is generally the case when it comes to field or large-scale survey data.

To demonstrate our results, we introduce two methodological contributions. First, we investigate this question through both a classical parametric approach and a new non-parametric approach. Some well-known non-parametric methods have been proposed, such as in Hey and Orme (1994) and Wakker and Deneffe (1996). However these two approaches can be applied only to specific datasets. In contrast, we show that our method is flexible and can be applied to more general contexts with reliable results. Second, an additional contribution of this paper is to introduce a relatively new way of testing the predictive power of models. This procedure, inspired by the standard practices of machine learning and data sciences, enables us to discuss models' performance more qualitatively and not only through the statistical significance of a few tests. As we will see, this design is especially helpful in interpreting the results of a model competition study.

This paper proceeds as follows. First, we describe the different streams of literature that this paper brings together: the characterization of heterogeneity in risk attitudes, non-parametric measurement of risk preferences, and the use of Bayesian statistics in decision sciences (Section 2). Then, we detail the methodology used in this study, more specifically, the elicitation of the risk preferences (Section 3) and the measurement of the predictive power (Section 4). We finally describe our results (Section 5) and, subsequently, the possible implications of this paper for further research (Section 6).

2 Background literature

2.1 Heterogeneity in risk preferences

The existence of substantial heterogeneity in risk attitudes is a well-established result in decision sciences. This point has already been underlined by early studies on the elicitation of risk

preferences (e.g., Hey and Orme, 1994, Holt and Laury, 2002). Harrison and Rutström (2008) conclude their seminal survey on risk preference elicitation by noticing the "considerable individual heterogeneity in risk attitudes [observed] in the laboratory." Moreover, the results of experiments implemented in large-scale surveys representative of the population generally confirm the substantial heterogeneity found in the laboratory (e.g., Harrison et al., 2006, von Gaudecker et al., 2011).

However, it is worth noting that the very notion of heterogeneity remains quite polysemic in this field. Moffatt (2015), for instance, differentiates the concepts of *continuous heterogeneity* and *discrete heterogeneity*, a typology also presented by Harrison and Rutström (2008) under the terms *heterogeneity in theories* and *heterogeneous theories*. Continuous heterogeneity corresponds to a situation where all the subjects share the same model but with different parameters. Discrete heterogeneity is, in some sense, more radical. In the case of discrete heterogeneity, subjects differ not only in their individual parameters, but also in the model they use. For instance, Harrison and Rutström (2009) famously elicit a mixture model where some individuals are supposed to follow an EU model and the others, the Prospect Theory.

Despite the fact that individuals differ in their risk attitudes, we have reasons to believe that heterogeneity in risk preferences has been overstated in previous studies. As we show in the next section, shrinking individual parameters toward their median values through a hierarchical approach tends to significantly improve the predictive power of decision models and the reliability of their estimates (e.g., Nilsson et al., 2011). In this paper, we push the analysis further, and we discuss the possibility that only one aspect of decision models could be driving the entire heterogeneity of the whole population. Therefore, we introduce the notion of *targeted heterogeneity*, a special case of *continuous heterogeneity* where only some aspects of the decision models are estimated at the individual level. In terms of complexity, this approach constitutes an intermediate possibility between the restrictive representative individual hypothesis and the assumption that each agent has her own model.

2.2 Non-parametric measurement of risk preferences

In this study, we used both a non-parametric and a parametric approach. In the elicitation of risk preferences, each approach has limitations that, in fact, complement the other. On the one hand, the main limitation of the parametric elicitation method is its lack of flexibility, which could lead to deceptive results if the functional forms are misspecified. On the other hand, the

non-parametric approach may be sometimes prone to overfitting. Therefore, we suggest that a two-fold approach is especially relevant, one method constituting the most natural robustness check of the other.

To the best of our best knowledge, there currently exist only two methods to elicit risk preferences without parametric assumptions: Hey and Orme's (1994) approach (HO) and Wakker and Deneffe's (1996) approach (WD).

The WD procedure, or Trade-off method, is based on successive adaptive decisions that determine equally spaced outcomes in terms of utility units, or *utils*, and consequently, the shape of the utility function. This technique has generated a considerable literature and has been successfully extended to such functions as the Probability Weighting Function (Abdellaoui, 2000; Bleichordt and Pinto, 2000), the Prospect Theory's value function (Abdellaoui et al., 2007) and the Regret function (Bleichrodt et al., 2010). Nevertheless, since this method relies on an adaptive experimental design, the choices proposed to the subjects depend on their previous answers, and this method cannot be applied to most datasets with risky decisions. This technique also suffers from several drawbacks; the procedure is generally not incentive-compatible (Harrison and Rutström, 2008), and its adaptive component can lead to error propagation issues (WD, p.1148; Blavatskyy, 2006; Richard and Baudin, 2020).

A possible alternative to the Trade-off method is the HO approach. The HO strategy consists in treating different levels in the functions to elicit, such as the utility function, as parameters to estimate. For example, in the EU model, one can maximize the likelihood function with respect to the utility levels associated with each outcome in the database (e.g., the utility of 0, 10, or 100 dollars, euros, or experimental units). An apparent limitation of this technique is that it becomes unrealistic when the number of levels in the elicited functions is large. This is why the authors themselves apply a non-parametric analysis to the EU model but rely on parametric forms to estimate an RDU model (see below), that gives less degrees of freedom. Otherwise, without parametric specification, the estimates given by the maximum likelihood could be misleading: in that case, the likelihood function is generally complex, and the solutions found by optimization algorithms are sensitive to initial values. Thus, like the Trade-off method, the HO approach can only be applied to databases that come from experiments especially designed to use this method, which generally have a reduced number of possible outcomes.[3] Because of this, and while this technique is detailed in the seminal

[3]For instance, HO consider only four outcomes: £0, £10, £20, and £30.

literature reviews on risk preferences (Harrison and Rutström, 2008; Moffatt, 2015), it has been little used since its introduction. Some exceptions include Gonzalez and Wu (1999), Hey et al. (2010), Kothiyal et al. (2014), and the multiple reconsiderations of the original HO dataset made, for instance, by Wilcox (2011) or Blavatskyy (2011).

In this paper, we return to the HO idea of treating different levels of the utility function as parameters, and we demonstrate that this method can actually be suitable to a large number of different outcomes if some elements from Bayesian statistics are added.

2.3 Bayesian statistics and decision sciences

The key aspect of Bayesian statistics is to treat the elicited parameters as random variables. As a result, the estimation of the parameters does not correspond to a precise value but to a credible distribution—or posterior distribution—given the observed data. By estimating a posterior distribution rather than precise values, some computational issues posed by the maximization algorithms in the non-parametric context can be avoided.

The posterior distribution of the parameters depends both on a prior distribution (generally vague) and on the choices made by the subjects. Then, the posterior distribution is provided by a Bayesian updating, such that the density of the posterior distribution in the case of a discrete predicted variable is

$$f_{post.}(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta) f_{prior}(\theta), \tag{1}$$

where θ is the vector of parameters, $\mathbf{x}$ is the set of observations, $f_{prior}(\theta)$ is the density of the prior distribution, and $f_{post.}(\theta|\mathbf{x})$ is the density of the posterior distribution. $P(\mathbf{x}|\theta)$ corresponds to the likelihood, namely, the probability to observe $\mathbf{x}$ given the parameters θ.[4]

The introduction of Bayesian statistics in the field of decision making under risk can be attributed to Nilsson et al. (2011), who developed a Hierarchical Bayesian measurement of Prospect Theory parameters.[5] In a hierarchical approach, *individual parameters* are supposed

[4] In Equation (1), $f_{post.}(\theta|\mathbf{x})$ is proportional and not necessarily equal to $P(\mathbf{x}|\theta) f_{prior}(\theta)$, since the integration of $P(\mathbf{x}|\theta) f_{prior}(\theta)$ on the parameters space Θ is also not necessarily equal to 1.

[5] In this paper, the terms "Prospect Theory" refer to the model initially called Cumulative Prospect Theory (Tverky and Kahneman, 1992), following the terminology from Wakker (2010). Let us also note here that Jarnebrant et al. (2009) are technically the first authors to introduce a Bayesian framework in decision theory. However, their work significantly departs from the rest of the literature, since they apply Bayesian statistics to the elicitation of a probit model, and not directly to a structural model of decision under risk.

to be drawn in the same prior distribution of parameters—called *hyper-parameters*—that have to be determined. This procedure allows collective inference since each subject contributes to the elicitation of hyper-parameters and thus, indirectly to the elicitation of the other subjects' parameters. Nilsson et al. (2011) demonstrate that this method outperforms classical maximum likelihood approaches, obtaining estimates that are more stable and reliable. This Hierarchical Bayesian approach is also supported by Scheibehenne and Pachur (2015), who obtain relatively similar results using another dataset.

Hierarchical Bayesian models have been extended to various research questions. Toubia et al. (2013) use this framework to evaluate the predictive power of their preference elicitation procedure. Balcombe and Fraeser (2015) pursue further research in this direction and compare the statistical performance of different functional forms of the Prospect Theory in the gain domain as well as those of a few of alternative models. This methodology has been applied by Ferecatu and Onçuler (2016) to the study of both risk and time decisions. Finally, Baillon et al. (2020) complete this literature by adding the possibility to elicit a mixture model within an HBE framework. More precisely, they assume that subjects are heterogeneous in the fixation of their reference point and measure the probability for each individual of following different reference-dependent decision models. Note, however, that Murphy and ten Brincke (2018) have shown that hierarchical models can also be implemented without Bayesian statistics and provide similar gains in terms of stability and predictive performance of the decision models.

3 Methodology

3.1 Data description

The data we use are drawn from an experimental study by Baillon et al. (2020). In this dataset, subjects, who were located in Moldova, had to select the lotteries they prefer in a series of different binary decisions. Figure 2 presents an example of such decisions, with outcomes expressed in the local currency (Lei). A common experimental protocol in the field (e.g., Hey and Orme, 1994; Harrison and Rutström, 2009), this experiment provides nevertheless several specific features particularly suitable for our study.

Although the subjects had only one chance out of three to see one of the selected options eventually played, the stakes involved were large according to the standards of experimental economics. Baillon et al. (2020) argue that "the subjects who played out their choices for real earned 330 Lei on average, which was more than half the average weekly salary [at the time

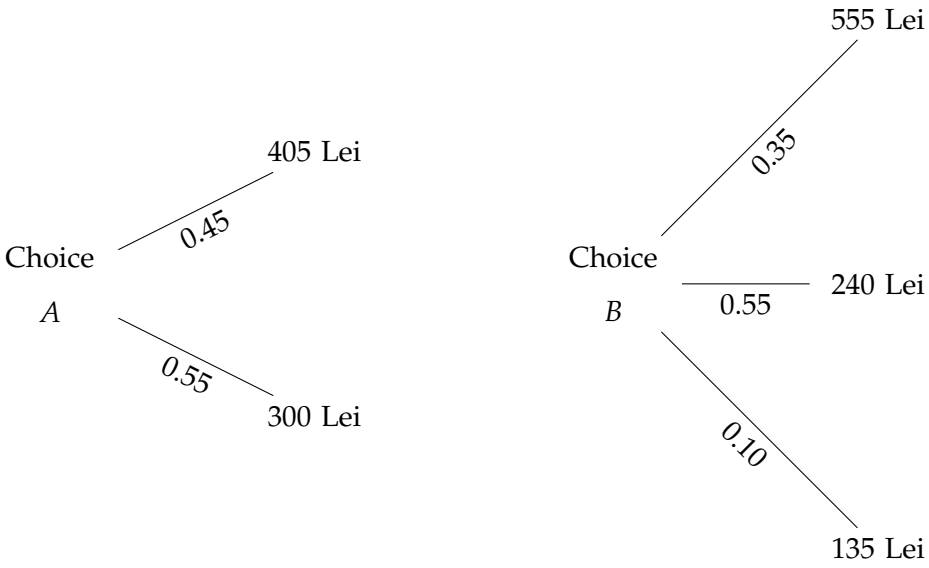


Figure (1) An example of binary lottery choice

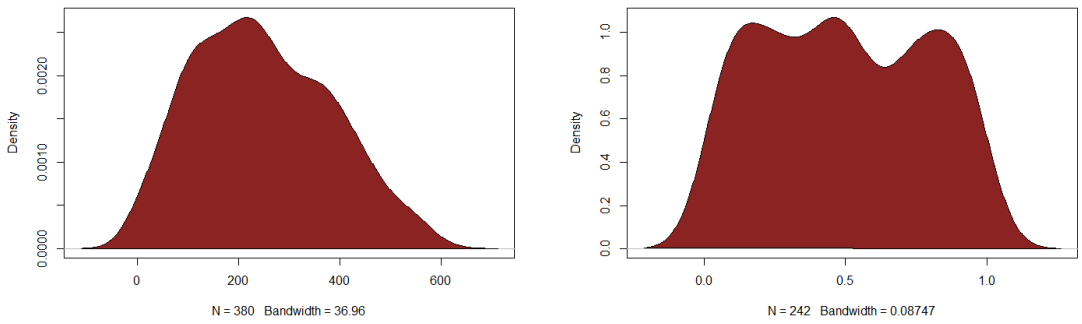
of the experiment]". Two subjects won the highest price possible of 601 Lei, the equivalent of an average week's salary. In addition to their potential gains from playing the lottery, all the participants received a fixed fee of 50 Lei.

To the best of our knowledge, the data from Baillon et al. (2020) are the only ones dealing with such monetary incentives within a Random Lottery Pair design (Harrison and Rutström, 2008). The few other databases that include monetary rewards of this magnitude generally offer a limited number of questions to the participants (e.g., Tanaka et al. 2010). Even if having a high-stake experiment is almost always an advantage in itself, this feature is especially important for our purpose. Determining the origin of heterogeneity in risk attitudes using low-stake lotteries necessarily downplays the importance of the utility function, since the latter is often assumed to be linear in small intervals. It should also be noted that since only one decision could potentially be played, we avoid hedging behaviors from the participants, thus facilitating the elicitation of risk preferences.

Subjects were students or employees at the Technical University of Moldova and had sufficient numeracy skills to understand the questions. Overall, 139 subjects participated in the experiment, and each of them made 70 different binary choices. The participants were aged between 17 to 47 years, with an average of 22, and approximately 60 % of them were male. Three observations out of 9730 were discarded for apparent mistakes in the data recording.[6]

[6]For those three observations, the values taken by the decision variable were inconsistent with the codebook.

This experiment was designed such that a large variety of binary choices would be presented to the subjects. The experiment thus comprised both simple choices with certainty amounts (8 choices) and complex choices with at least three outcomes for each option (19 choices). More importantly, the algorithm employed by Baillon et al. (2020) to construct their choices ensures the complete coverage of the outcome and of the probability spaces (from 0 to 601 Lei and from 0 to 1, respectively). The main goal of their algorithm was to have, as they state, "minimally correlated choices" that should "lead to more efficient and more robust estimates." Figure 2 introduces the frequencies at which the different outcomes and cumulative probabilities are present in the choices offered to the subjects, confirming full coverage of both the outcome and of the probability spaces. This feature of the database is particularly suitable for non-parametric methods of risk preference elicitation. Otherwise, the estimation of some of the values in the utility function or in the PWF would have to rely on a small number of answers from the participants.



(a) The kernel density of the experimental outcomes in the experiment (in Lei)

(b) The kernel density of the cumulative probabilities in the experiment

Figure (2) The distributions of the probabilities and outcomes presented to the subjects

The last attractive feature of this dataset is that five questions were proposed twice to the participants, enabling us to compute a consistency rate. This rate is approximately 70 %, which corresponds to the consistency rate generally reported for non-trivial decisions under risk (Stott, 2006). Thus, we obtain the maximal accuracy rate achievable, which constitutes a useful benchmark of predictive performance.[7]

This consistency rate can be employed to measure the "completeness" of decision models. In a recent study, Fudenberg et al. (2019) consider a theory as "complete" when its prediction

[7]Baillon et al. (2020) also interpret the consistency rate in a similar way.

errors correspond to irreducible errors, given the existence of a stochastic component in decision-making. A low predictive power is uninformative when it comes to the completeness of a theory, since it could correspond to the best performance that one can achieve with noisy data. Here, the consistency rate can be interpreted as determinant information for measuring the completeness of a theory, because any decision model assuming stable preferences during the experiment cannot provide an accuracy rate higher than the consistency rate.

For the more technical aspects, the experiment was run on computers, and the order of the questions was randomized. The order of the choices presented to the subject— i.e., which option is considered as lottery A or lottery B — was also randomly selected. The overall experiment, including the reading of the instruction, was completed, on average, in 30 minutes by the participants.

3.2 Eliciting risk preferences

3.2.1 Rank-dependent utility theory

We apply our targeting approach of dealing with heterogeneity, which is developed in the next section, to one of the most prevailing theories of decision under risk: the Rank-Dependent Utility model. The Rank-Dependent Utility model was developed by Quiggin (1982), and its primary objective is to capture the propensity for an individual to overweight or underweight probabilities in her decision process, without violating the stochastic dominance principle.[8]

In this model, the decision-maker evaluates different lotteries and then chooses her best option according to her preferences. Each lottery corresponds to a probability distribution over money. Here, x_i refers to an amount of money, and p_i refers to the probability that is assigned to the outcome x_i. The lottery the subject has to evaluate is commonly denoted by the vector $(p_1 : x_1, ..., p_n : x_n)$. This vector is presented here in decreasing order, such that $x_i > x_j$ for any $i < j$.

Compared with the EU model, in an RDU model, subjects do not only transform money into utility but also probabilities into decision weights. According to the RDU model, the decision-maker chooses the option that maximizes the following decision criterion:

$$U = \sum_{i=1}^{n} \pi_i u(x_i) \tag{2}$$

[8]Having only outcomes framed as gains here, we can also argue, such as in Balcombe and Fraeser (2015), that we equivalently study the Cumulative Prospect Theory in the gain domain.

with

$$\pi_i = w(\sum_{j=1}^{i} p_j) - w(\sum_{j=1}^{i-1} p_j), \tag{3}$$

where $u(\cdot)$ designates the utility function and $w(\cdot)$ the probability weighting function.

The utility function is normalized, such that the highest outcome in the database corresponds to a utility of 1 and the lowest to a utility of 0. Similarly, we normalize the outcomes such that the highest outcome in the whole experiment (601 Lei) is now equal to 1 and the lowest (0 Lei) to 0.

In the parametric part of this study, we assume a traditional power specification of the utility function,

$$u(x) = x^r, \tag{4}$$

with $r > 0$.[9] For the PWF, we assume another classical functional form proposed by Prelec (1998), such that

$$w(p) = exp(-\beta(-\log(p))^\gamma). \tag{5}$$

In Equation (5), β determines the elevation of the function, while γ captures its curvature. Most especially, a γ inferior to one gives to the function the inverse-S shaped aspect generally found at an aggregate level. This decomposition is useful because, as explained in the Introduction, our results indicate that all heterogeneity in risk behavior can be attributed to the heterogeneity of the PWF. Thus, this distinction enables us to determine which specific element in the PWF, its elevation, or its curvature is essential for describing individual differences in risk attitudes.

For the non-parametric part of this paper, we estimate n different equally-spaced levels of the utility function and the PWF. For $n = 3$, for instance, we would determine $u(0.25)$, $u(0.5)$, $u(0.75)$, $w(0.25)$, $w(0.5)$, and $w(0.75)$. The intermediate values of the utility function and the PWF are inferred using a linear interpolation. If we write $N = n + 1$, then for an outcome x between $(k-1)/N$ and k/N, its corresponding utility is given by the equation

[9] The more general form $u(x) = \frac{x^{1-\rho}}{1-\rho}$ is not relevant here because one of the possible outcomes in this experiment was 0, which could give an infinite negative utility in the case where $\rho > 1$.

$$u(x) = \frac{x - \frac{k-1}{N}}{\frac{k}{N} - \frac{k-1}{N}}(u_k - u_{k-1}) + u_{k-1}. \tag{6}$$

Similarly, for a probability p between $(k-1)/N$ and k/N, its corresponding decision weight is given by

$$w(p) = \frac{p - \frac{k-1}{N}}{\frac{k}{N} - \frac{k-1}{N}}(w_k - w_{k-1}) + w_{k-1}. \tag{7}$$

Therefore, the shape of the two functions depends on two vectors, $\theta_u = (u_0, u_1, ..., u_N)$ and $\theta_w = (w_0, w_1, ..., w_N)$. Thus, u_0 designates the utility of the worst possible outcome in the experiment (normalized to 0), and u_N refers to the utility of the best possible outcome in the experiment (normalized to 1). Likewise, w_0 and w_N are defined such that $w_0 = w(0) = 0$ and $w_N = w(1) = 1$. The utility and the PWF can be then rewritten as

$$u(x) = \sum_{k=1}^{N} \min\{\max\{Nx - (k-1), 0\}, 1\}(u_k - u_{k-1}), \tag{8}$$

and

$$w(p) = \sum_{k=1}^{N} \min\{\max\{Np - (k-1), 0\}, 1\}(w_k - w_{k-1}). \tag{9}$$

3.2.2 Defining appropriate priors

As stated in the previous sections, using Bayesian statistics requires assuming a prior belief on the parameters' distribution. These prior distributions are relatively straightforward to define for the parametric part of this study. As there exists a positivity constraint on all the individual parameters (r, γ, and β), we assume that they are drawn from a log-normal prior distribution with $\log r \sim \mathcal{N}(\mu_r, \sigma_r)$, $\log \gamma \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma)$, and $\log \beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$. The position and dispersion hyper-parameters, $\mu_\star$ and $\sigma_\star$, have a normal and an inverse-gamma distribution, respectively, as hyper-priors. These hyper-priors are diffuse and uninformative.

Defining appropriate priors is more challenging in the non-parametric approach and requires more details. In this case, there exists two different but interdependent problems. Supposing a hierarchical model, the first issue is to define a relevant hyper-prior on the hyper-parameters; the vectors $\theta_u^* = (u_0^*, u_1^*, ..., u_N^*)$ and $\theta_w^* = (w_0^*, w_1^*, ..., w_N^*)$. θ_u^* or θ_w^* correspond to the levels of the utility and the PWF of the population on average, such that we penalize the elicited parameters far from those values. Therefore, θ_u^* and θ_w^* mostly correspond to the non-parametric

counterpart of the position parameters $\mu_\star$ for the parametric study. The second problem is to determine a prior that penalizes individual estimates far from the hyper-parameters.

Let us start by solving the first issue, namely, the definition of an appropriate hyper-prior for the hyperparameters θ_u^* and θ_w^*. The first proper characteristic required for our hyper-prior distribution is to have a null joint-density for any non-increasing sequence of θ_u^* or θ_w^*. Otherwise, the hyper-parameters would implicitly correspond to a non-increasing utility function or PWF, and thus to a DM that could prefer stochastically dominated options.

The second important property is to assume a prior distribution that does not favor either convex or concave forms. Thus, another criterion is that the prior should be centered on linear functions, such that the utility and the probability weighting levels have the following expected values:

$$E(u_k^*) = k/N \text{ and } E(w_k^*) = k/N,$$

for any $k \in \{1, ..., N-1\}$. The last property is to have a prior distribution vague enough such that it is possible to recover any shape of the utility or of the PWF for a large number of observations.

The solution we propose in this paper is to define this hyper-prior sequentially. First, we fix n levels for each function, with $n = 2^K - 1$ (or $N = 2^K$) and K any natural number. Then, we assume that the "middle levels", $u_{N/2}^*$ and $w_{N/2}^*$, follow a uniform distribution of parameters 0 and 1. This prior reflects our general ignorance about the overall shape of the functions. From the first distribution on the hyper-priors, we define the prior distribution of two additional intermediate utility levels as

$$\begin{cases} u_{N/4}^* | u_{N/2}^* \sim \mathcal{U}(0, u_{N/2}^*) \\ u_{3N/4}^* | u_{N/2}^* \sim \mathcal{U}(u_{N/2}^*, 1), \end{cases} \tag{10}$$

and we define $w_{N/4}^*$ and $w_{3N/4}^*$ similarly. With a second iteration of this process, we can also define four additional intermediate levels :

$$\begin{cases} u_{N/8}^* | u_{N/4}^* & \sim & \mathcal{U}(0, u_{N/4}^*) \\ u_{3N/8}^* | (u_{N/4}^*, u_{n/2}^*) & \sim & \mathcal{U}(u_{N/4}^*, u_{N/2}^*) \\ u_{5N/8}^* | (u_{N/2}^*, u_{3n/4}^*) & \sim & \mathcal{U}(u_{N/4}^*, u_{3N/4}^*) \\ u_{7N/8}^* | u_{3N/4}^* & \sim & \mathcal{U}(u_{3N/4}^*, 1). \end{cases} \tag{11}$$

Then, we define the additional $w_{N/8}^*$, $w_{3N/8}^*$, $w_{5N/8}^*$, and $w_{7N/8}^*$ levels in a similar fashion. The prior distribution eventually obtained is presented in Figure 3, with the 2.5%, 25%, 50%,

75%, and 97.5% quantiles of the hyper-prior distribution.

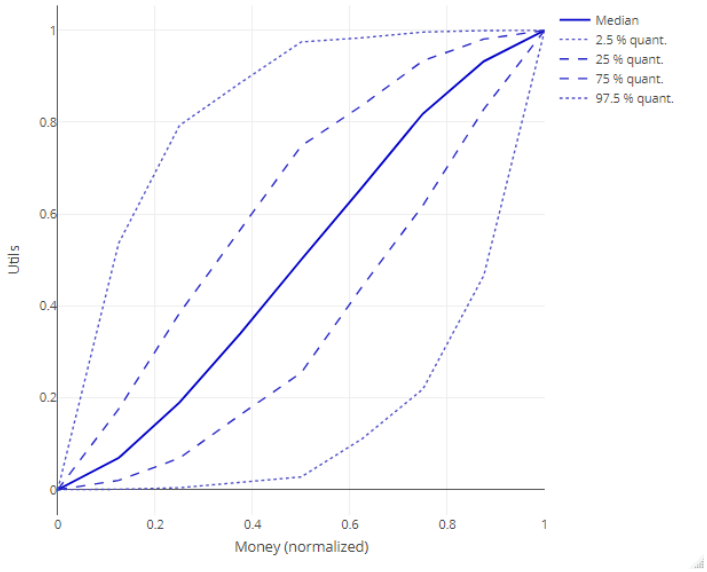


Figure (3) Prior distribution of the utility function.

We stop this process at three iterations, and we keep 7 points for each function (which also corresponds to $N = 8$). According to our estimates, limiting ourselves to $n = 7$ seems sufficient to obtain a reliable measurement of the utility and of the PWF. We can verify without demonstration that the expected values of the utility and probability weighting levels do not bias the estimates toward concave or convex forms, and that the density of any non-increasing sequence of utility or probability weighting levels is null. Also, as indicated in Figure 3, the distribution eventually obtained can be qualified as uninformative.

Now, we propose a strategy to penalize individual estimates that are distant from the functions denoted by θ_u^* and θ_w^*. As above, we reason sequentially, and we assume, as a prior for $u_{N/2,j}$, a normal distribution of parameters $u_{N/2}^*$ and σ_u, truncated at 0 and 1, such that we penalize the individual estimates of $u_{N/2,j}$ that are far from $u_{N/2}^*$. Then, we assume that $u_{N/4,j}$ follows a normal distribution of parameter $u_{N/4}^*$ and σ_u, but is truncated both on the left at 0 and on the right at $u_{N/2,j}$ to guarantee that the final estimates correspond to an increasing function. Similarly, $u_{3N/4,j}$ follows a normal distribution of parameter $u_{3N/4}^*$ and σ_u, but is truncated on the left at $u_{N/2,j}$ and on the right at 1. Thus, we have

$$\begin{cases} f_{N/4,j}(u_{N/4,j}|u_{N/2,j}) & = & \begin{cases} \frac{f(u_{N/4,j}|u^*_{N/4},\sigma_u)}{F(u_{N/2,j}|u^*_{N/4},\sigma_u)-F(0|u^*_{N/4},\sigma_u)} & \text{if } u_{3N/4,j} \in [0; u_{N/2,j}], \\ 0 & \text{otherwise;} \end{cases} \\ f_{3N/4,j}(u_{3N/4,j}|u_{N/2,j}) & = & \begin{cases} \frac{f(u_{3N/4,j}|u^*_{3N/4},\sigma_u)}{F(1|u^*_{3N/4},\sigma_u)-F(u_{N/2,j}|u^*_{3N/4},\sigma_u)} & \text{if } u_{N/4,j} \in [u_{N/2,j}; 1], \\ 0 & \text{otherwise,} \end{cases} \end{cases} \tag{12}$$

where $f(x|\mu,\sigma)$ and $F(x|\mu,\sigma)$ are the density and the cumulative distribution function of a normal distribution of parameters μ and σ. Thus, $f_{N/4,j}$ and $f_{3N/4,j}$ are the densities of our priors. We also define four additional intermediate levels in a similar fashion.

As a consequence, we have as a prior a distribution similar in its construction to the hyper-prior, since when σ_u or σ_w tend toward infinity, each prior distribution tends toward a uniform distribution. The hyper-prior distribution is a limit case of the prior distribution, which implies that it is only possible to improve the reliability of our individual estimates with that prior. If no collective inference can be made concerning the shape of the utility function or of the PWF—in other words, if it is impossible to improve the quality of the estimates by using the decisions of others subjects—σ_u and σ_w become high, and the prior does not significantly differ from the hyper-prior.

3.3 The error model

To estimate the parameters of an RDU model using Bayesian statistics, we have to assume an error model that specifies the probability of choosing a particular option, given the utility and the probability weighting functions of a subject. Throughout this paper, we suppose a standard logit specification with an additional tremble parameter, where the probability of choosing an option B over an option A is given through

$$P(X = B) = \omega/2 + (1-\omega)\frac{e^{U_B/\xi}}{e^{U_A/\xi} + e^{U_B/\xi}}, \tag{13}$$

where U_A and U_B are the deterministic utilities of the options A and B (given by the RDU model, see Equation (2)), and ξ and ω are two noise parameters.[10]

[10]The tremble parameter ω possesses two different interpretations. First, its psychological interpretation could be the following: in addition to their ability to discriminate the best choice, which is captured by ξ, the subjects can also make another kind of mistake when they sometimes randomly click on an option. The second interpretation is statistical: the addition of a tremble parameter is generally recommended in Bayesian analysis to make the estimates more robust when the sample corresponds to a small number of dichotomous variables (see, for instance, the discussionabout on the robust logistic regression model, in the reference textbook by Kruschke, 2015, p. 621).

Individual ξ are assumed to be drawn in the same log-normal distribution. For simplicity's sake, and to avoid possible identification issues with ξ, we also assume that the tremble parameter is universal. As ω is likely to be relatively small, we define its prior through a beta distribution of parameters 1 and 9, following recommendations from Kruschke (2015, p. 621). Note that additional noise specifications will also be considered in the robustness checks, with similar results. The model was estimated using JAGS, and the details of the Markov Chains Monte Carlo (MCMC) are introduced in the online appendix.

4 Testing the predictive power

4.1 The importance of the predictive power in decision sciences

Before introducing the evidence that all heterogeneity in risk attitude can be attributed to the heterogeneity in probability weighting, we should explain another relative innovation of this study with respect to the rest of the research on decisions under risk.

As stated in Introduction, comparing the statistical performance of structural models of decision-making, or model competition, is a widespread approach in behavioral economics for evaluating theories. The earliest theory competitions using the functional forms of the models of decisions under risk were Camerer and Ho (1994), Hey and Orme (1994), and Loomes and Sugden (1995). This approach has been especially employed to study models of error[11] but has also been extended to various issues such as decisions under ambiguity[12] or time decisions[13]. However, as Hey et al. (2010) state, in those competitions, "statistical significance tells us nothing about economic significance. Nor does it tell us whether the increase in statistical predictability is worth the reduction in theoretical parsimony." (p.83). Given the importance of this approach, it is surprising that, until now, the characterization in absolute terms of the predictive power of the models has not been a primary issue, as it would be interesting to know whether a difference in terms of predictive accuracy or goodness-of-fit between two models corresponds to, for instance, a difference between an excellent model and a poor model, or a difference between a fair model and a poor model.

As demonstrated in the robustness section and in the appendices, the addition of a tremble parameter tends to improve the predictions of the models.

[11] e.g. Loomes et al. (2002), Blavatskyy (2011), and Wilcox (2011).

[12] e.g. Hey et al. (2010) and Kothiyal et al. (2014).

[13] e.g. Arfer and Luhmann (2015) and Blavatskyy and Maafi (2018).

This eventually constitutes a paradoxical situation: while the insufficient predictive power of the standard theory is often given as a primary reason for the development of behavioral economics, far too little attention has been paid on the *qualitative* and *absolute* assessment of the models' predictive power. Glöckner and Pachur (2012) and Peysakhovich and Naecker (2017) are two exceptions, but it should be noted that their manner of measuring the quality of predictions significantly differs from ours.

To deal with the predictive power of the behavioral models in-depth, we adopt the standard approach from machine learning, an area in which the issue of prediction is crucial. Nowadays, the idea that machine learning will transform the practice of econometrics and applied economics is widespread.[14] In this section, we argue that the methodology usually recommended in machine learning to assess the predictive performance of models can be particularly suitable for model competitions in decision sciences.[15] In comparison with most of the literature on risk-preferences and related topics, the key differences in this approach are (1) the choice of the indicators of predictive performance that possess a qualitative interpretation and (2) the existence of a train set/test set distinction (the second point is already present in several papers such as Wilcox, 2011).

4.2 Indicators of predictive performance

The simplest of the predictive indicators used in this study is probably the accuracy rate, or hit rate, but it can also be very misleading. It is well-known that for imbalanced data, it is easy to reach a high accuracy rate with a simple random guess (this result is sometimes called the "accuracy paradox.").[16] However, it is worth noting that the accuracy rate can be even more

[14] In an already seminal paper on this issue, Varian (2014) writes, "My standard advice to graduate students these days is go to the computer science department and take a class in machine learning" (p. 3). More recently, Camerer (2018) also highlights that the impact of machine learning should be more potent in behavioral economics than anywhere else in economics. Machine learning allows economists to handle high-dimensional data more easily, and as Camerer (2018) says, behavioral economics precisely tends to deal with a more significant number of relevant factors than the other sub-fields of economics.

[15] Let us clarify the relationship of our approach with machine learning. We do not pretend to have a machine learning approach for this issue. We have simply chosen to import its practices to treat a question for which this field is especially advanced. As Varian (2014) states, "Machine learning is concerned primarily with prediction" (p.4).

[16] By imbalanced data, we mean, in the context of binary choices, a situation where an option is chosen with a higher frequency than the other.

problematic when using lottery choices data. While the data might appear to be balanced at the aggregate level with an almost equal distribution of risky choices and safe choices for the whole database, the dataset of each subject is generally imbalanced, as the population is heterogeneous in terms of risk preferences. As models are estimated at an individual level, computing the aggregate accuracy rate on the whole dataset is no different from computing the mean of the individual accuracy rates. Consequently, even though the aggregate accuracy rate can be seen as relevant at first glance, it only corresponds to the mean of accuracy rates obtained using imbalanced data, and, thus, to the mean of irrelevant indicators of predictive performance. Therefore, the accuracy rate will only be given as a complementary piece of information, and to compare it to the consistency rate (see above). There exists a vast literature on predictive performance measures, and multiple indicators more relevant than the hit rate have been developed.[17] The ones that we introduce in this paper avoid the biases linked to imbalanced data issues.

The second indicator that we use is Cohen's kappa (Cohen, 1960), a statistic initially designed to measure the degree of agreement between two judges or experts, that has been generalized to the measurement of the adequacy between predictions and observations. Its formula is

$$\kappa = \frac{\text{accuracy rate} - p_e}{1 - p_e}, \tag{14}$$

where p_e is the probability of being correct only by chance,

$$p_e = P(\text{prediction} = 1)P(\text{observation} = 1) + P(\text{prediction} = 0)P(\text{observation} = 0).$$

Thus, Cohen's kappa is found between $-p_e/(1 - p_e)$ and 1. The following scale is often given in machine learning textbooks for the interpretation of this indicator (see for instance Lantz, 2015, p. 323):

- $\kappa < 0$: Disagreement
- $\kappa \in [0.00; 0.20]$: Poor agreement
- $\kappa \in]0.20; 0.40]$: Fair agreement
- $\kappa \in]0.40; 0.60]$: Moderate agreement

[17] See Sokolova et al. (2006) for an overview of the most popular indicators in machine learning.

- $\kappa \in]0.60; 0.80]$: Good agreement
- $\kappa \in]0.80; 1.00]$: Very good agreement.

The adequacy between predictions and observations can also be measured through a simple correlation coefficient. The obtained indicator is often called the Matthews' correlation coefficient or MCC. The value of this indicator is given by

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{15}$$

where TP, TN, FP and FN are the True Positive, True Negative, False Positive and False Negative rates, respectively. The MCC is defined between -1 and 1 and can be interpreted on the same scale as a standard correlation coefficient.

4.3 Estimation of predictive power

To test the predictive accuracy of a model, it is preferable to measure the adequacy between predictions and observations on a dataset that has not been used for training the model. Otherwise, it is not the predictive performance of a model that is measured, but its capacity to rationalize a given dataset, thereby favoring complex models prone to overfitting. Thus, one-fifth of the data were reserved to measure the predictive accuracy (1933 observations in total, between 12 and 16 observations per subject). Except for the accuracy rate, predictive performance measures can differ depending on the manner in which we define the variable to predict. In this study, we measure our indicators using the dummy variable that takes 1 when the DM chooses the "risky option" over the "safe option" and 0 otherwise.[18]

To make predictions from the estimates, we use the *posterior predictive distribution*, such that the probability of taking the risky option is given by

$$P(X = 1) = \int_{\theta \in \Theta} P(X = 1|\theta) f_{post.}(\theta) d\theta. \tag{16}$$

In our setup, we can use the indicators of predictive performance introduced in the previous section in two ways:

[18] We define the risky option as the option with the highest variance.

- First, by considering the data as a whole and then measuring the predictive power on the entire test set.
- Second, by considering each subject independently and studying the distribution of the indicators over the individual test sets.

For this reason, the Tables below present the predictive performance measures at the aggregate level, as well as the medians of individual Cohen's kappas.[19]

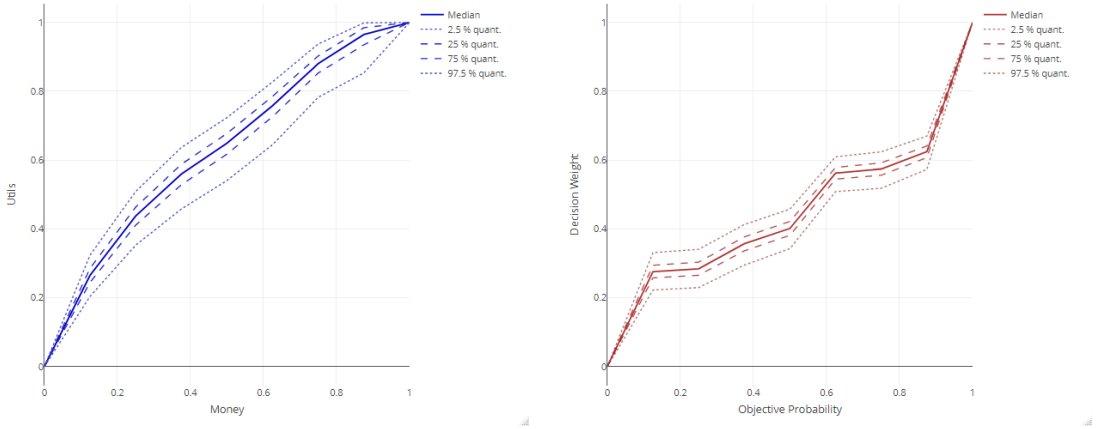
5 Results

5.1 Results from the baseline RDU model

Regarding our non-parametric approach, the first way to describe the participants' risk preferences is to introduce the estimates of a hypothetical representative agent who would have taken all the decisions from the database. Despite its limitations, this popular approach enables us to verify that the preliminary results given by our non-parametric approach are consistent with the rest of the literature. The estimates of this fictitious individual are described in Figure 4, which shows the distribution of the median utility (left graph) or probability weighting (right graph) levels of the posterior distribution, as well as two lower (2.5 % and 25%) and two upper (75% and 97.5 %) quantiles. As shown in Figure 4, a classical curvature can be recovered for both for the utility function and the PWF. The utility is thus mildly concave, and the probability weighting function is inverse-S shaped. Thus, the representative agent tends to overweight small probabilities (below 0.3) and to underweight larger ones. However, individual estimates possess, on average, more unstable or atypical shapes, such as exclusively concave or convex PWF or irregular utility functions (see the online appendix).

Concerning the parametric part of this study, Table 1 summarizes the results given by the estimation of the baseline RDU model, with no constraint on the PWF or on the utility function. For each parameter, we took the median values of the posterior distribution as point estimates. As the parameters are strictly positive and assumed to be drawn from the same log-normal distribution, Table 1 introduces the log transformation of these values for more readability. Predictably, the utility function appears to be concave (with $\log r < 0$) for all the subjects, while most of the subjects display an inverse-S shape of the PWF (with $\log \gamma < 0$). The estimation of

[19] Individual MCC are not presented here because this indicator could not be computed for a significant part of the subjects.



(a) Utility Function (b) Probability Weighting Function

Figure (4) The representative agent's estimates

Table (1) Distribution of the point estimates, Baseline RDU model

Parameters	2.5 %	25 %	Median	75 %	97.5 %
$\log r$	-1.8	-1.39	-1.16	-0.97	-0.14
$\log \beta$	-1.38	-0.40	0.20	0.61	1.41
$\log \gamma$	-1.75	-1.07	-0.62	-0.05	0.71
$\log \xi$	-4.31	-3.86	-3.50	-3.29	-2.87

Note: This table describes the distribution of the point estimates of each parameter at an individual level of a classical RDU model. Each point estimate corresponds to the median of the posterior distribution of each individual parameter.

the universal tremble parameter ω is approximately 0.08.

From this baseline model, the dispersion of the individual estimates seems already more pronounced for the PWF parameters $\log \gamma$ and $\log \beta$ compared with the utility parameter, $\log r$. However, this result is not sufficient to conclude that the heterogeneity in risk behavior is driven by the heterogeneity of the PWF, a conclusion we will more formally demonstrate below.

5.2 Model competition and main results

We now introduce the main result of this paper, namely that the probability weighting function appears to drive all the heterogeneity in risk attitudes. To do so, we compare the performance of the baseline RDU model measured above with that of two constrained RDU models in predicting the experimental choices made by the subjects. The first of those constraints is the

Table (2) Out-of-sample predictive power, non-parametric approach

Model	Constraint	Kappa	MCC	Acc.	Median kappa	Log-likelihood	Vuong test
RDU	–	0.32	0.32	0.69	0.18	−1118.21	
	Unique Utility	0.34	0.35	0.70	0.24	−1098.95	0.03
	Unique PWF	0.20	0.22	0.66	0.00	−1139.39	0.002
EU	–	0.20	0.20	0.63	0.14	−1227.18	< 0.001
DT	–	0.33	0.33	0.69	0.19	−1122.44	0.001

Note: The kappa, MCC, Acc. columns correspond to the Cohen's Kappa, MCC and Accuracy rate on the complete test set, respectively. The median kappa corresponds to the median Cohen's kappa calculated for each individual test sets (with a test set by subject). The log-likelihood corresponds to the log-likelihood of the model on the complete test set. The "Vuong test" column corresponds to the p-value of a Vuong test that tests each model against the baseline rank-dependent utility model (or the baseline rank-dependent utility model against the model under consideration if this one has a higher log-likelihood). In the column constraint, "-" indicates an absence of constraint in the elicitation of the model, "Unique Utility" indicates that all the subjects share the same utility function and "Unique PWF" indicates that all the subjects share the same PWF.

existence of a unique utility function, and the second, the existence of a unique PWF. In other words, we assume in the first case that all individual heterogeneity in terms of risk attitudes is driven by the PWF, while in the second case, all individual heterogeneity can be attributed to the utility function.

Following Wilcox (2011), we add the out-of-sample log-likelihood and the p-value of a Vuong test to the different predictive performance indicators introduced in the previous section. Each Vuong test in Tables 2 and 3 is performed in comparison with the baseline RDU model. We also add to these results the performances of an EU model and a Dual Theory model (Yaari, 1987). The Dual Theory (DT) describes a special case of RDU models where the utility function is linear. Thus, the DT corresponds to the exact opposite of the other noticeable special case of the RDU models, the EU model, that implicitly adopts a linear probability weighting function.

Tables 2 and 3 reveal that the constraint of a unique PWF significantly decreases the predictive performance of the RDU model, while the same restriction applied to the utility function does not have a similar effect. Conversely, when a universal utility function is assumed, the predictive power tends to be higher, although this increase is not statistically significant. Moreover, both the unconstrained RDU model and the RDU model with a unique utility

Table (3) Out-of-sample predictive power, parametric approach

Model	Constraint	Kappa	MCC	Acc.	Median kappa	Log-likelihood	Vuong test
RDU	–	0.36	0.37	0.71	0.29	−1065.60	
	Unique Utility	0.36	0.37	0.71	0.29	−1064.04	0.24
	Unique Utility + Unique Sensitivity	0.33	0.34	0.70	0.29	−1082.03	< 0.001
	Unique Utility + Unique Elevation	0.22	0.27	0.67	0.19	−1115.67	< 0.001
	Unique PWF	0.22	0.25	0.66	0.14	−1121.95	< 0.001
EU	–	0.19	0.19	0.63	0.11	−1279.40	< 0.001
DT	–	0.34	0.35	0.70	0.27	−1077.90	0.02

Note: The signification of the columns "Kappa," "MCC," "Acc.," "Median kappa," "Log-likelihood," and "Vuong test" remain the same as in Table 2. In the constraint column, "Unique utility + Unique Sensitivity" means that the individuals share the same r and γ parameters while "Unique utility + Unique Elevation" means that the individuals share the same r and β parameters.

function have a predictive performance that can be described as fair or moderate. In contrast, the predictive performance of the RDU model with a unique probability function may only be characterized as poor.[20] This result is confirmed both in the parametric and non-parametric studies. Consequently, heterogeneous probability weighting seems crucial to capture systematic differences in risk attitudes, while the utility function does not seem to possess similar importance.

Thus, these estimates argue in favor of a unique utility function. While supposing a unique utility function does not significantly improve the model's predictive power, it makes the model simpler; if we apply Ockham's razor principle, the particular case of a unique utility function among subjects must be considered as true. The unique utility functions we obtain in the parametric and non-parametric studies are described in Figure 5. As expected, the utility function is, in both cases, increasing and mildly concave.

In the case of the parametric study, we can now deepen our analysis and test additional constraints on the parameters by measuring the performance of two models, by assuming (1) a unique utility function and a unique curvature parameter γ and (2) a unique utility function and a unique elevation parameter β. When a unique curvature parameter γ is assumed, the

[20]The confidence interval at 95% of each predictive performance measure is introduced in the online appendix for more readability. The confidence interval of each of these indicators is narrow, with a precision around 0.04.

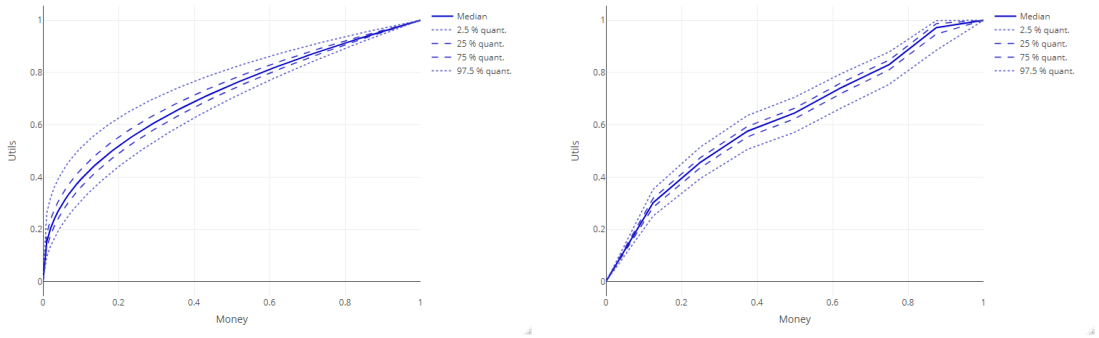
quality of the prediction decreases only moderately. Conversely, assuming a unique elevation parameter β makes the predictive performance of the RDU model almost as low as that of the EU model. The heterogeneity of risk attitudes is thus driven mainly by the heterogeneity in the elevation of the PWF, classically interpreted as optimism and pessimism (e.g., Wakker, 2010), and not by the heterogeneity in the shape or curvature of that same function. As most heterogeneity in risk attitudes is due to the probability weighting function, it is not surprising to find that the Dual Theory performs well on this dataset, while the EU model displays poor predictive performances. Note, however, that the performance of the Dual Theory does not imply that the utility function is linear, a hypothesis that is rejected both in our estimates (see Figure 5) and by a Vuong test (see Tables 2 and 3).

The most predictive models studied here reach a level of predictive accuracy close to the "completeness" level, as defined above. The unconstrained RDU and the "unique-utility RDU" models, parametric or non-parametric, provide an accuracy rate around 70%, slightly above the observed consistency rate.[21] From these results, we can also confirm that the non-parametric methodology gives reliable estimates despite its high degree of flexibility. Additionally, the results confirm the relevance of our methodology based on the qualitative assessment of predictive power, which extends beyond the issue of statistical significance. A Vuong test alone, for instance, is not sufficient to characterize the importance of the disparity in terms of predictive accuracy between two models. The qualitative differences between the performances of our models appear here not only as statistically significant but also economically significant. By analyzing the predictive performance qualitatively, we can also underline the considerable noise in the decision process of subjects that face risky choices since the best predictive performance achievable seems, at best, moderate.

5.3 Additional specifications and robustness checks

This section aims to check the robustness of our results to alternative tests or hypotheses. More specifically, we discuss the robustness of our results using a model competition based on the in-sample goodness-of-fit of the models rather than their predictive power. We also check whether the main results of our parametric study can be attributed to a misspecification of

[21] If five choices were repeated in the experimental protocol, these choices were not randomized. This explains the possibility for the accuracy rate to be somewhat above the consistency rate. However, these five choices were not especially easier nor more difficult than the others, so the consistency rate can still be considered as a good proxy for the maximal accuracy rate one can achieve on this database.



(a) Power specification (b) Non-parametric specification

Figure (5) The unique utility function

(1) the noise model, (2) the utility function, and (3) the prior chosen for the utility function parameter. We finally discuss the relevance of the RDU model in comparison with reference-dependent models. For all of these robustness checks, we limited ourselves to the parametric approach, as the non-parametric approach is computationally more costly.

5.3.1 Studying the internal goodness-of-fit

We first check the validity of our results by comparing the in-sample goodness-of-fit of our models rather than their predictive power. We then estimate the models on the complete dataset (with no train set/test set distinction), and we compute the Deviance Information Criterion (DIC) of each model.[22] As in the previous sections, the DIC favors the unique utility function model against all the other models, including the baseline RDU model (see the online appendix).

To check whether the difference in the goodness-of-fit between the two models is statistically significant, we run a mixture model that includes a parameter I, which equals 1 if the true model is the baseline RDU model and 0 if the RDU model with a unique utility function is the true model. The average I in the MCMC sample gives an estimation of the probability that the subjects have different utility function. Thus, we now assume

$$P(X = 1) = (1 - I) \cdot P_{\text{Unique}}(X = 1|\theta_{\text{Unique}}) + I \cdot P_{\text{Baseline}}(X = 1|\theta_{\text{Baseline}}), \tag{17}$$

where θ_{Unique} are the parameters of the unique utility model and, $\theta_{Baseline}$ are the parameters of the baseline RDU model. The prior and hyper-prior distributions are the same as those in

[22]The DIC is a measure similar in its construction to the Akaike Information Criterion but more adapted to Bayesian statistics.

the previous sections. The probability for the baseline RDU model to be true is inferior to 0.001, and thus our estimates still strongly support that the utility function is stable among subjects.[23]

5.3.2 Robustness to additional specifications

We also check the validity of our results exploiting two alternative noise specifications: the classical logit specification without the tremble parameter (see Equation (13)) and the heteroscedastic specification proposed by Wilcox (2011). In the Wilcox's heteroscedastic specification, or Contextual Utility model, the noise parameter ξ in Equation (13) becomes a function of the lottery choice options, labeled A and B, such that

$$\xi(A, B) = \bar{\xi} \cdot (u_{max}(A, B) - u_{min}(A, B)), \tag{18}$$

where $u_{max}(A, B)$ is the utility of the best outcome possible in the lotteries A and B, $u_{min}(A, B)$ the utility of the worst outcome possible, and $\bar{\xi}$ is a new parameter to be estimated. Thus, with this specification, choices that have an important gap between the worst and the best outcomes, also generate more incertitude in decision-making. We demonstrate in the appendices that with those new noise structures, the predictive performance of the models remains stable across specifications, and our conclusions remain unchanged.[24]

Regarding the other robustness checks, we focus on the RDU model with a unique PWF, and we verify that its low predictive performance is not due to a misspefication. We test two additional prior distributions on r for the RDU model with a unique PWF: a normal distribution truncated at 0, and a gamma distribution. In each of these specifications, the predictive power of the RDU model with a unique PWF remains unchanged and is largely inferior to both the performance of the baseline RDU model and to an RDU model with only a unique utility function. We also test whether other functional forms of the utility function could provide a better predictive performance. More precisely, we use an expo-power utility function, as proposed by Saha (1993), and a utility function that has the same functional form as our PWF (see Equation (5)). The purpose of this latter specification is to take into account that the importance given to the PWF over the utility function in our results might only reflect the

[23] Bayesian hypothesis testing is different in its construction from the tests performed in frequentist statistics, and indirectly penalizes complex models. As a consequence, the probability of the constrained model being the true model against the unconstrained model can be greater than 0.5 or even close to 1 if that constraint is not associated with a decrease in the goodness-of-fit.

[24] Note thain this case, we adopt a probit model with a tremble parameter, instead of a logit model, following the specification used by Wilcox (2011).

flexibility of the functional form proposed by Prelec (1998).[25] As before, the results obtained with these two new specifications of the utility function do not affect our conclusions (see the online appendix).

5.3.3 Possibility of a reference-dependent behavior

As already stated, the database used here was initially designed to test different reference points in decision-making under risk—such as the status quo or the stochastic reference point proposed in the Kőszegi-Rabin model (Kőszegi and Rabin, 2006). Thus, there could exist a misspecification of our own statistical design if the participants have indeed reference-dependent preferences and not the basic RDU model assumed above. The first counter-argument against a misspecification of our model is that, as we have already seen, the RDU model appears as a "complete theory" in the sense of Fudenberg et al. (2019), since the accuracy rate of our highest-performing models is equivalent to the consistency rate. Thus, no significant gains in terms of predictive accuracy can be achieved through new and more complex decision models.

Nevertheless, this question is investigated in further detail in the online appendix where we revisit some of the results from Baillon et al. (2020). In particular, we show that the evidence for reference-dependent behaviors in this database can be reevaluated. We demonstrate more specifically that with the noise specification we proposed in this study, a mixture model of different reference-point rules gives a lower statistical performance than a simpler RDU model (as measured by the DIC). Moreover, the RDU model, or the status quo reference point, was already considered by Baillon et al. (2020) in their conclusions as the most widespread model of decision under risk among the subjects. Therefore, we may conclude that the decisions not explained by the RDU model most likely reflect pure noise.

6 Discussion and further research

The possibility of a utility function stable among individuals has numerous implications. From a theoretical point of view, the idea that heterogeneity in risk attitudes is a matter of optimism or pessimism contradicts the idea that the EU theory could constitute even a relevant approximation of true risk preferences. According to our results, the PWF is more than a simple auxiliary feature that improves the descriptive performance of the standard model. Since the PWF is, by

[25] Considering the function provided by Prelec (1998) to describe the PWF as a possible utility function is relevant here since both the outcomes and the utility were normalized between 0 and 1.

definition, missing from the EU theory, this model can only mischaracterize the very nature of the heterogeneity in risk attitudes. This mischaracterization can be particularly harmful when the EU model is used to provide welfare evaluation, which is in general very dependent on the shape of the utility function (e.g., Harrison and Ng, 2016). Moreover, assuming a unique utility function could make behavioral economics models' applications more tractable, especially in situations involving several interacting agents or individuals.

From an empirical point of view, our *targeted heterogeneity approach* constitutes a reliable alternative to what we could call the "one individual, one model" approach. By targeting heterogeneity only in some aspects of the model, estimates could also become more readable. In the context of decisions under risk, for instance, the optimism or elevation parameter can be used as a unique indicator of risk aversion, even when facing a non-EU model. Currently, the only alternative approach to obtain a unique indicator of risk attitude when there are multiple decisions to aggregate is to assume an EU model with one risk aversion parameter, even when it fits the data poorly — which is, as we have seen above, often the case. Our *targeting heterogeneity approach* could be especially useful in measuring risk preferences on field or survey data, where the number of questions is typically low and the risk of overfitting is especially high. However, the subjects of this study were essentially composed of students from the same university, and it is unknown if more heterogeneous utility functions could be found on a sample more representative of the general population.

Given these results, it could be tempting to assume further implications with normative purposes. If all the subjects share the same utility function, and if probability weighting is considered as a rationality bias, then for each choice there should exist an ideal decision that every subject should adopt. However, this perspective was not considered in this paper for at least two reasons. The first is that probability weighting is not universally interpreted as an irrational feature; for instance, Harrison and Ng (2016, 2018) use an RDU model to measure the welfare gains from insurance decisions. Second, it is not certain that the utility function that describes the best risk attitudes is also the most relevant to describe welfare gains or well-being.

Regarding our methodological contributions, we proposed several innovations to demonstrate our results, and this paper could also constitute a blueprint for new research questions. This is particularly true concerning our non-parametric statistical design, which is adaptable to other error structures (e.g., Blavatskyy, 2011) and to other decision models (e.g., Bell, 1982, Loomes

and Sugden, 1982, Gul, 1991, or Kőszegi and Rabin, 2006). The framework proposed here could also be useful for topics concerning decision-making in general, beyond the issue of risk preferences. An extension of this methodology to ambiguity attitudes, for instance, would be relatively straightforward, by exploiting the similarities of the RDU model and the "Choquet-Expected Utility" model (see Wakker, 2010).

Our non-parametric approach could be helpful, even for studies that opt for parametric strategies, in two different ways. One is *ex ante*, in which our design could be employed for choosing a relevant functional form, which could be selected depending on the shape of the functions found non-parametrically at an aggregate level (e.g., from the representative agent estimates). The other is *ex post*; in this case, our non-parametric approach could serve as a robustness check, since the parametric forms that are clearly outperformed by the non-parametric method are likely to be misspecified.

7 Conclusion

Following the increasing interest being paid to Bayesian statistics in decision sciences during the last decade, we developed a new methodology to determine what drives heterogeneity in risk attitudes. This procedure is innovative in several ways. First, we apply both parametric and non-parametric approaches to test our hypotheses. Second, we measure the predictive performance of the models under consideration not only in relative, but also in absolute terms. Our results reveal that most of the heterogeneity in risk behaviors can be attributed to heterogeneity in probability weighting. More precisely, the variety of the risk behaviors generally observed in the laboratory is largely explained only by differences in the elevation of the probability weighting function among individuals.

Our results seem determinant in at least two aspects. First, it demonstrates all the dangers of overfitting in estimating risk preferences, since spurious results, such as the existence of heterogeneity in the utility function, may only correspond to pure noise. The issue of overfitting in the estimation of risk preferences has only been considered in a handful of papers in decision sciences and, yet, according to our findings, should no longer be ignored. Second, our findings contradict the idea that the expected utility model could be considered as a relevant approximation of true risk preferences. Given our results, when interpreting the RDU model, the PWF should not be regarded simply as an auxiliary feature that increases the goodness-of-fit of the standard theory, but can be considered as the key element making individuals more or less risk-averse.

References

Abdellaoui, M. 2000. Parameter-free elicitation of utility and probability weighting functions. *Management Science.* 46(11):1497–1512.

Abdellaoui, M., Bleichrodt, H., Paraschiv, C. 2007. Loss aversion under prospect theory: A parameter-free measurement. *Management Science.* 53(10):1659–1674.

Arfer, K. B., Luhmann, C. C. 2015. The predictive accuracy of intertemporal-choice models. *British Journal of Mathematical and Statistical Psychology.* 68(2):326–341.

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

Balcombe, K., Fraser, I. 2015. Parametric preference functionals under risk in the gain domain: A Bayesian analysis. *Journal of Risk and Uncertainty.* 50(2):161–187.

Bell, D. E. 1982. Regret in decision making under uncertainty. *Operations Research.* 30(5):961–981.

Blavatskyy, P. R. 2011. A model of probabilistic choice satisfying first-order stochastic dominance. *Management Science.* 57(3):542–548.

Blavatskyy, P. R., Maafi, H. 2018. Estimating representations of time preferences and models of probabilistic intertemporal choice on experimental data. *Journal of Risk and Uncertainty.* 56(3):259–287.

Bleichrodt, H., Cillo, A., Diecidue, E. 2010. A quantitative measurement of regret theory. *Management Science.* 56(1):161–175.

Bleichrodt, H., Pinto, J. L. 2000. A parameter-free elicitation of the probability weighting function in medical decision analysis. *Management science.* 46(11):1485–1496.

Camerer, C. F. 2018. Artificial intelligence and behavioral economics. *The economics of artificial intelligence: an agenda*. University of Chicago Press.

Camerer, C. F., Ho, T. H. 1994. Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty.* 8(2):167–196.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 20(1):37–46.

Ferecatu, A., Önçüler, A. 2016. Heterogeneous risk and time preferences. *Journal of Risk and Uncertainty.* 53(1):1-28.

Fudenberg, D., Kleinberg, J., Liang, A., Mullainathan, S. 2019. Measuring the completeness of theories. *arXiv preprint arXiv:1910.07022.*

Glöckner, A., Pachur, T. 2012. Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition.* 123(1):21–32.

Gonzalez, R., Wu, G. 1999. On the shape of the probability weighting function. *Cognitive Psychology.* 38(1):129–166.

Gul, F. 1991. A theory of disappointment aversion. *Econometrica.* 59(3):667–686.

Harrison, G. W., Ng, J. M. 2016. Evaluating the expected welfare gain from insurance. *Journal of Risk and Insurance.* 83(1):91–120.

Harrison, G. W., Ng, J. M. 2018. Welfare effects of insurance contract non-performance. *The Geneva Risk and Insurance Review.* 43(1):39–76.

Harrison, G. W., Rutström, E. E. 2008. Risk aversion in the laboratory. *Research in experimental economics.* 12(8):41–196.

Harrison, G. W., Rutström, E. E. 2009. Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental Economics.* 12(2):133–158.

Hey, J. D., Lotito, G., Maffioletti, A. 2010. The descriptive and predictive adequacy of theories of decision making under uncertainty/ambiguity. *Journal of Risk and Uncertainty.* 41(2):81–111.

Hey, J. D., Orme, C. 1994. Investigating generalizations of expected utility theory using experimental data. *Econometrica.* 62(6):1291–1326.

Holt, C. A., Laury, S. K. 2002. Risk aversion and incentive effects. *American Economic Review.* 92(5):1644–1655.

Jarnebrant, P., Toubia, O., Johnson, E. 2009. The silver lining effect: Formal analysis and experiments. *Management Science.* 55(11):1832–1841.

Kahneman, D., Tversky, A. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica.* 47(2):263–292.

Kőszegi, B., Rabin, M. 2006. A model of reference-dependent preferences. *Quarterly Journal of Economics*. 121(4):1133–1165.

Kruschke, J. 2015. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* Academic Press.

Lantz, B. 2015. *Machine Learning with R: Expert techniques for predictive modeling to solve all your data analysis problems.* Packt Publishing Ltd.

Loomes, G., Sugden, R. 1982. Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal.* 92(368):805–824.

Loomes, G., Moffatt, P. G., Sugden, R. 2002. A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty.* 24(2):103–130.

Moffatt, P. G. 2015. *Experimetrics: Econometrics for experimental economics*. Macmillan International Higher Education.

Moffatt, P. G., Peters, S. A. 2001. Testing for the presence of a tremble in economic experiments. *Experimental Economics.* 4(3):221–228.

Murphy, R. O., ten Brincke, R. H. 2018. Hierarchical maximum likelihood parameter estimation for cumulative prospect theory: Improving the reliability of individual risk parameter estimates. *Management Science.* 64(1):308–326.

Nilsson, H., Rieskamp, J., Wagenmakers, E. J. 2011. Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology.* 55(1):84–93.

Peysakhovich, A., Naecker, J. 2017. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization.* 133:373–384.

Prelec, D. 1998. The probability weighting function. *Econometrica.* 66(3):497–527.

Quiggin, J. 1982. A theory of anticipated utility. *Journal of Economic Behavior & Organization.* 3(4):323–343.

Richard, T., Baudin, V. 2020. Asymmetric noise and systematic biases: A new look at the Trade-Off method. *Economics Letters.* 191:109132.

Saha, A. 1993. Expo-power utility: a 'flexible' form for absolute and relative risk aversion. *American Journal of Agricultural Economics.* 75(4):905–913.

Scheibehenne, B., Pachur, T. 2015. Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review.* 22(2):391–407.

Sokolova, M., Japkowicz, N., Szpakowicz, S. 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*. Springer, Berlin, Heidelberg. pp. 1015–1021.

Stott, H. P. 2006. Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty.* 32(2):101–130.

Tanaka, T., Camerer, C. F., Nguyen, Q. 2010. Risk and time preferences: Linking experimental and household survey data from Vietnam. *American Economic Review.* 100(1):557–71.

Toubia, O., Johnson, E., Evgeniou, T., Delquié, P. 2013. Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters. *Management Science.* 59(3):613–640.

Tversky, A., Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty.* 5(4):297–323.

Varian, H. R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives.* 28(2):3–28.

Von Gaudecker, H. M., Van Soest, A., Wengstrom, E. 2011. Heterogeneity in risky choice behavior in a broad population. *American Economic Review.* 101(2):664–94.

Wakker, P. P. 2010. *Prospect theory: For risk and ambiguity.* Cambridge university press.

Wakker, P., Deneffe, D. 1996. Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management science.* 42(8):1131–1150.

Wilcox, N. T. 2008. Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. *Risk aversion in experiments.* 12:197–292.

Wilcox, N. T. 2011. 'Stochastically more risk averse:'A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics.* 162(1):89–104.

Yaari, M. E. 1987. The dual theory of choice under risk. *Econometrica.* 55(1):95–115.

Appendices

A Detailed results: predictive power

A.1 Parametric Study

A.1.1 Accuracy rate

Table (4) Accuracy rate, Parametric Study

Models		95 % Confidence interval
RDU	0.71	0.69 - 0.73
RDU Unique Utility	0.71	0.69 - 0.73
RDU Unique Utility and Sensi.	0.70	0.68 - 0.72
RDU Unique Utility and Eleva.	0.67	0.65 - 0.69
RDU Unique PWF	0.66	0.64 - 0.68
EU	0.63	0.61 - 0.65
Dual	0.70	0.68 - 0.72

A.1.2 M.C.C.

Table (5) M.C.C., Parametric Study

Models		95 % Confidence interval
RDU	0.37	0.33 - 0.41
RDU Unique Utility	0.37	0.33 - 0.40
RDU Unique Utility and Sensi.	0.34	0.30 - 0.38
RDU Unique Utility and Eleva.	0.27	0.22 - 0.31
RDU Unique PWF	0.25	0.20 - 0.29
EU	0.19	0.15 - 0.23
Dual	0.35	0.31 - 0.39

A.1.3 Kappa

Table (6) Cohen's Kappas, Parametric Study

Models		95 % Confidence interval
RDU	0.36	0.32 - 0.40
RDU Unique Utility	0.36	0.31 - 0.40
RDU Unique Utility and Sensi.	0.33	0.28 - 0.37
RDU Unique Utility and Eleva.	0.24	0.19 - 0.29
RDU Unique PWF	0.22	0.17 - 0.27
EU	0.19	0.14 - 0.24
Dual	0.34	0.30 - 0.39

A.1.4 Individual Kappas

Table (7) Individual Kappas Distribution, Parametric Study

Models	Min	25 %	50 %	75 %	Max
RDU	-0.32	0.07	0.29	0.47	1.00
RDU Unique Utility	-0.32	0.06	0.29	0.49	1.00
RDU Unique Utility and Sensi.	-0.35	0.00	0.29	0.47	1.00
RDU Unique Utility and Eleva.	-0.43	0.00	0.19	0.39	1.00
RDU Unique PWF	-0.47	0.00	0.14	0.36	1.00
EU	-0.40	-0.11	0.11	0.28	0.81
Dual	-0.24	0.00	0.28	0.47	1.00

Note: This table introduces the minimum, the maximum, as well as the 25%, 50%, and 75% quantiles of the distribution of individual kappas for each model.

A.2 Non-Parametric Study

A.2.1 Accuracy rate

Table (8) Accuracy rate, Non-Parametric Study

Models		95 % Confidence interval
RDU	0.69	0.67 - 0.71
RDU Unique Utility	0.70	0.68 - 0.72
RDU Unique PWF	0.66	0.64 - 0.68
EU	0.63	0.61 - 0.65
Dual	0.69	0.67 - 0.71

A.2.2 Kappa

Table (9) Cohen's Kappas, Non-Parametric Study

Models		95 % Confidence interval
RDU	0.32	0.28 - 0.36
RDU Unique Utility	0.34	0.31 - 0.38
RDU Unique PWF	0.20	0.16 - 0.24
EU	0.20	0.16 - 0.24
Dual	0.33	0.29 - 0.36

A.2.3 M.C.C.

Table (10) M.C.C., Non-Parametric Study

Models		95 % Confidence interval
RDU	0.32	0.28 - 0.36
RDU Unique Utility	0.35	0.31 - 0.38
RDU Unique PWF	0.22	0.17 - 0.26
EU	0.20	0.16 - 0.25
Dual	0.33	0.29 - 0.37

A.2.4 Individual Kappas

Table (11) Individual Kappas Distribution, Non-Parametric Study

Indicators	Min	25 %	50 %	75 %	Max
RDU	-0.41	-0.01	0.18	0.40	1.00
RDU Unique Utility	-0.32	0.00	0.24	0.44	1.00
RDU Unique PWF	-0.43	-0.09	0.00	0.30	1.00
Dual	-0.38	-0.05	0.14	0.33	0.87
EU	-0.36	0.00	0.19	0.42	1.00

Note: This table introduces the minimum, the maximum, as well as the 25%, 50%, and 75% quantiles of the distribution of individual kappas for each model.

B Detailed results : estimates

B.1 Parametric Study

B.1.1 Baseline RDU model

Table (12) Baseline RDU model, Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
$log(r)$ (Individual point estimates distri.)	-1.8	-1.39	-1.16	-0.97	-0.14
$log(\beta)$ (Individual point estimates distri.)	-1.38	-0.40	0.20	0.61	1.41
$log(\gamma)$ (Individual point estimates distri.)	-1.75	-1.07	-0.62	-0.05	0.71
$log(\xi)$ (Individual point estimates distri.)	-4.31	-3.86	-3.50	-3.29	-2.87
ω (Posterior distribution)	0.06	0.08	0.10	0.11	0.13
DIC	10773.26				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameter that is common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the complete dataset (with no train set/test set distinction).

B.1.2 Unique utility RDU model

Table (13) Unique utility RDU model, Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
$log(r)$ (Posterior distribution)	-1.22	-1.02	-0.94	-0.86	-0.73
$log(\beta)$ (Individual point estimates distri.)	-1.78	-0.38	0.19	0.58	1.44
$log(\gamma)$ (Individual point estimates distri.)	-1.72	-1.06	-0.61	-0.03	0.78
$log(\xi)$ (Individual point estimates distri.)	-4.28	-3.76	-3.31	-2.99	-2.46
ω (Posterior distribution)	0.06	0.08	0.09	0.1	0.13
DIC	10683.56				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameters that are common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the complete dataset (with no train set/test set distinction).

B.1.3 Unique PWF RDU model

Table (14) Unique PWF RDU model, Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
$log(r)$ (Individual point estimates distri.)	-3.65	-2.61	-1.98	-1.43	1.60
$log(\beta)$ (Posterior distribution)	0.05	0.1	0.13	0.16	0.22
$log(\gamma)$ (Posterior distribution)	-1.38	-1.29	-1.23	-1.18	-1.09
$log(\xi)$ (Individual point estimates distri.)	-5.79	-4.16	-3.34	-2.64	-1.14
ω (Posterior distribution)	0.09	0.11	0.13	0.14	0.18
DIC	11350.71				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameters that are common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the complete dataset (with no train set/test set distinction).

B.1.4 Unique utility and unique sensitivity RDU model

Table (15) Unique utility and unique sensitivity RDU model, Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
$log(r)$ (Posterior distribution)	-1.13	-0.94	-0.85	-0.77	-0.63
$log(\beta)$ (Individual point estimates distri.)	-2.76	0.03	1.50	3.46	5.28
$log(\gamma)$ (Posterior distribution)	0.03	0.09	0.13	0.17	0.24
$log(\xi)$ (Individual point estimates distri.)	-4.10	-3.36	-2.86	-2.20	-1.74
ω (Posterior distribution)	0.07	0.09	0.10	0.12	0.14
DIC	10876.92				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameters that are common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the complete dataset (with no train set/test set distinction).

B.1.5 Unique utility and unique elevation RDU model

Table (16) Unique utility and unique elevation RDU model, Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
$log(r)$ (Posterior distribution)	-1.77	-1.4	-1.25	-1.12	-0.92
$log(\beta)$ (Posterior distribution)	0.17	0.25	0.3	0.35	0.5
$log(\gamma)$ (Individual point estimates distri.)	-2.38	-1.34	-0.78	-0.04	1.12
$log(\xi)$ (Individual point estimates distri.)	-4.84	-3.70	-2.97	-2.17	-0.81
ω (Posterior distribution)	0.06	0.08	0.10	0.11	0.13
DIC	11195.27				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameters that are common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the complete dataset (with no train set/test set distinction).

B.1.6 Dual model

Table (17) Dual model, Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
$log(\beta)$ (Individual point estimates distri.)	-1.07	-0.05	0.40	0.86	1.63
$log(\gamma)$ (Individual point estimates distri.)	-1.69	-1.15	-0.65	-0.14	0.54
$log(\xi)$ (Individual point estimates distri.)	-4.16	-3.5	-3.12	-2.69	-2.21
ω (Posterior distribution)	0.07	0.09	0.11	0.12	0.15
DIC	10856.63				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameter that is common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the complete dataset (with no train set/test set distinction).

B.1.7 EU model

Table (18) EU model, Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
$log(r)$ (Individual point estimates distri.)	-8.5	-6.85	-5.26	-4.07	2.66
$log(\xi)$ (Individual point estimates distri.)	-10.2	-9.04	-8.59	-8.06	-7.7
ω (Posterior distribution)	0.34	0.48	0.51	0.54	0.57
DIC	12687.22				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameter that is common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the complete dataset (with no train set/test set distinction).

B.2 Non-parametric study

B.2.1 Baseline RDU model

Table (19) Baseline RDU model, Non-Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
u_1 (Individual point estimates distri.)	0.15	0.30	0.37	0.42	0.51
u_2 (Individual point estimates distri.)	0.26	0.44	0.51	0.57	0.66
u_3 (Individual point estimates distri.)	0.35	0.56	0.61	0.66	0.78
u_4 (Individual point estimates distri.)	0.41	0.63	0.69	0.75	0.84
u_5 (Individual point estimates distri.)	0.55	0.73	0.76	0.82	0.88
u_6 (Individual point estimates distri.)	0.69	0.83	0.86	0.89	0.93
u_7 (Individual point estimates distri.)	0.92	0.94	0.95	0.96	0.97
w_1 (Individual point estimates distri.)	0.03	0.16	0.28	0.48	0.67
w_2 (Individual point estimates distri.)	0.06	0.29	0.46	0.59	0.75
w_3 (Individual point estimates distri.)	0.12	0.40	0.55	0.68	0.80
w_4 (Individual point estimates distri.)	0.18	0.56	0.68	0.79	0.88
w_5 (Individual point estimates distri.)	0.30	0.65	0.77	0.85	0.92
w_6 (Individual point estimates distri.)	0.40	0.74	0.87	0.92	0.96
w_7 (Individual point estimates distri.)	0.47	0.87	0.94	0.97	0.98
$log(\xi)$ (Individual point estimates distri.)	-4.67	-3.83	-3.31	-2.82	-2.17
ω (Posterior distribution)	0.10	0.13	0.15	0.16	0.19
DIC	9971.45				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameter that is common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the test set only.

B.2.2 Unique utility RDU model

Table (20) Unique utility RDU model, Non-Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
u_1 (Posterior distribution)	0.26	0.30	0.32	0.34	0.38
u_2 (Posterior distribution)	0.40	0.44	0.47	0.49	0.54
u_3 (Posterior distribution)	0.50	0.55	0.59	0.61	0.66
u_4 (Posterior distribution)	0.56	0.62	0.65	0.68	0.72
u_5 (Posterior distribution)	0.65	0.71	0.74	0.77	0.81
u_6 (Posterior distribution)	0.75	0.80	0.83	0.86	0.89
u_7 (Posterior distribution)	0.84	0.90	0.94	0.97	1.00
w_1 (Individual point estimates distri.)	0.03	0.14	0.27	0.44	0.64
w_2 (Individual point estimates distri.)	0.06	0.24	0.39	0.53	0.71
w_3 (Individual point estimates distri.)	0.11	0.34	0.48	0.60	0.77
w_4 (Individual point estimates distri.)	0.15	0.46	0.59	0.69	0.85
w_5 (Individual point estimates distri.)	0.28	0.55	0.68	0.78	0.89
w_6 (Individual point estimates distri.)	0.32	0.63	0.77	0.88	0.95
w_7 (Individual point estimates distri.)	0.41	0.75	0.88	0.94	0.98
$log(\xi)$ (Individual point estimates distri.)	-4.09	-3.63	-3.28	-2.97	-2.62
ω (Posterior distribution)	0.06	0.09	0.11	0.13	0.16
DIC	9128.00				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameters that are common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the test set only.

B.2.3 Unique PWF RDU model

Table (21) Unique PWF RDU model, Non-Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
u_1 (Individual point estimates distri.)	0.00	0.19	0.29	0.36	0.49
u_2 (Individual point estimates distri.)	0.01	0.35	0.47	0.57	0.69
u_3 (Individual point estimates distri.)	0.03	0.51	0.59	0.69	0.87
u_4 (Individual point estimates distri.)	0.08	0.62	0.70	0.78	0.91
u_5 (Individual point estimates distri.)	0.25	0.74	0.80	0.85	0.94
u_6 (Individual point estimates distri.)	0.51	0.87	0.90	0.93	0.97
u_7 (Individual point estimates distri.)	0.89	0.95	0.96	0.97	0.99
w_1 (Posterior distribution)	0.52	0.59	0.61	0.64	0.68
w_2 (Posterior distribution)	0.53	0.58	0.62	0.64	0.69
w_3 (Posterior distribution)	0.54	0.60	0.63	0.65	0.70
w_4 (Posterior distribution)	0.64	0.70	0.72	0.72	0.78
w_5 (Posterior distribution)	0.65	0.71	0.74	0.74	0.79
w_6 (Posterior distribution)	0.68	0.73	0.76	0.76	0.81
w_7 (Posterior distribution)	0.69	0.74	0.76	0.76	0.82
$log(\xi)$ (Individual point estimates distri.)	-5.32	-3.76	-2.38	-1.09	0.26
ω (Posterior distribution)	0.12	0.15	0.17	0.19	0.22
DIC	9665.14				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameters that are common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the test set only.

B.2.4 Dual model

Table (22) Dual model, Non-Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
w_1 (Individual point estimates distri.)	0.08	0.24	0.39	0.55	0.75
w_2 (Individual point estimates distri.)	0.14	0.34	0.48	0.63	0.82
w_3 (Individual point estimates distri.)	0.21	0.45	0.59	0.71	0.86
w_4 (Individual point estimates distri.)	0.26	0.55	0.69	0.81	0.91
w_5 (Individual point estimates distri.)	0.39	0.64	0.77	0.86	0.94
w_6 (Individual point estimates distri.)	0.43	0.71	0.86	0.93	0.97
w_7 (Individual point estimates distri.)	0.50	0.83	0.93	0.97	0.99
$log(\xi)$ (Individual point estimates distri.)	-3.80	-3.45	-3.16	-2.90	-2.59
ω (Posterior distribution)	0.06	0.09	0.11	0.14	0.18
DIC	9310.45				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameter that is common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the test set only.

B.2.5 EU model

Table (23) EU model, Non-Parametric Study

	2.5 %	25 %	50 %	75 %	97.5 %
u_1 (Individual point estimates distri.)	0.41	0.60	0.66	0.74	0.82
u_2 (Individual point estimates distri.)	0.54	0.70	0.77	0.84	0.94
u_3 (Individual point estimates distri.)	0.65	0.79	0.84	0.90	0.98
u_4 (Individual point estimates distri.)	0.72	0.84	0.89	0.93	0.98
u_5 (Individual point estimates distri.)	0.81	0.90	0.93	0.95	0.99
u_6 (Individual point estimates distri.)	0.92	0.95	0.96	0.97	0.99
u_7 (Individual point estimates distri.)	0.97	0.98	0.98	0.99	1.00
$log(\xi)$ (Individual point estimates distri.)	-5.59	-3.74	-3.09	-2.20	-0.94
ω (Posterior distribution)	0.10	0.15	0.17	0.20	0.25
DIC	10514.32				

Notes: This table introduces the 2.5%, 25%, 50%, 75%, and 97.5% quantiles of the individual point estimates distribution (for the parameters that are estimated at a subject-level) and of the posterior distribution (for the parameter that is common to all subjects). The point estimates of the individual parameters corresponds to the median of the posterior distribution. The DIC corresponds to the Deviance Information Criterion. The estimates are obtained using the test set only.

C Robustness to the noise structure

Table (24) Out-of-sample predictive power (additional noise structures)

Robustness Check	Model	Constraint	Kappa	MCC	Acc.	Median kappa	Log-likelihood	Vuong test (p-value)
Cont. Utility	RDU	None	0.33	0.33	0.70	0.21	−1101.93	
		Unique Utility	0.35	0.35	0.70	0.19	−1092.48	0.53
		Unique PWF	0.21	0.23	0.66	0.00	−1132.67	0.07
Logit	RDU	None	0.32	0.32	0.69	0.14	−1120.86	
		Unique Utility	0.33	0.33	0.70	0.16	−1103.86	0.22
		Unique PWF	0.19	0.21	0.65	0.00	−1159.9	0.05

Notes: This table summarizes the indicators of predictive power obtained with different noise structure, a contextual utility model and a Logit model respectively. The Vuong test is realized in comparison with a baseline RDU model with the same noise structure.

D Robustness to additional specifications

Table (25) Out-of-sample predictive power, RDU unique PWF models

Robustness Check Specification	Change	Kappa	MCC	Acc.	Median kappa	Log-likelihood
Prior	Gamma	0.19	0.22	0.66	0.00	−1132.56
	Truncated Normal	0.16	0.18	0.65	0.00	−1161.69
Utility function	Saha (1993)	0.17	0.20	0.66	0.00	−1157.37
	Prelec (1998)	0.18	0.20	0.65	0.00	−1168.74

Notes: This table summarizes the indicators of predictive power obtained with additional parametric assumptions for the model with a unique probability weighting function. We changed the specification first of the prior on the r parameter and then, the parametric form given to the utility function.

E Comments on Baillon et al. (2020)

This section aims to discuss the results from Baillon et al. (2020), and the relevance for us to adopt a simple Rank-Dependent Utility (RDU) model to describe the data rather than reference-dependent behavior models.

As already mentioned in our paper, Baillon et al. (2020) elicit a Prospect Theory model with heterogeneous reference points (rp from now). Then, they classically suppose the prospect theory's value function :

$$v(x) = \begin{cases} (x - rp)^r \text{ if } x < rp \\ -\lambda(rp - x)^r \text{ otherwise,} \end{cases} \tag{19}$$

with λ the classical parameter of loss aversion, a r a positive parameter that determines the curvature of the value function. Subjects are supposed to follow a particular rule to fix their reference point, with :

- Rule 1 : the status quo, with a reference point at 0 (implicitly an RDU model, Quiggin, 1982).
- Rule 2 : the "MaxMin" rule, where the reference point is the maximum of the minimal outcomes of the two lotteries.
- Rule 3 : the "MinMax" rule, where the reference point is the minimum of the maximal outcomes of the two lotteries.
- Rule 4 : the "X at P max" rule, where the reference point is the outcome that has the highest probability in the two lotteries.
- Rule 5 : the EV rule, where the Expected Values of the lotteries are also their reference points.
- Rule 6 : the KR rule, where the reference point is the lottery itself, using a stochastic reference point (Kőszegi and Rabin, 2006)

The prospect theory's value function is then added to the Expected Value of the lottery to obtain the total utility — that depends, consequently, on the sum of a "rational part" and of a reference-dependent part.

We tested the robustness of the results from Baillon et al. (2020) by comparing the goodness-of-fit of their mixture model with our simple RDU framework. To have comparable results, we changed the noise structure of the Baillon et al. (2020) model and we supposed a logit model with a tremble parameter instead of the simpler logit model chosen in the original paper (as we can see below, this noise specification tends to better fit the data, since the tremble parameter we obtain is far from 0). We also choose a more general PWF parametric form, with the version from Prelec (1998) with two parameters presented above instead of the version with only one parameter used in the original paper, at least concerning the main specification.

According to our results, the posterior distribution indicates that half of the individuals follow the RDU model, and this model seems overwhelmingly dominant in the population. Moreover, we see that the goodness-of-fit of an RDU model, as measured by the deviance information criterion, is higher for an RDU model than for the mixture model Baillon et al. (2020) propose. While this short comment is not sufficient to conclude to the absence of reference-dependent behaviors in this context, the RDU hypothesis we use seems to be at least a relevant approximation of the true decision model adopted by most subjects.

Table (26) Results from the mixture model and from the RDU model

	Mixture Model	RDU
r (Median of the point estimates)	0.51	0.31
λ (Median of the point estimates)	1.90	-
β (Median of the point estimates)	0.82	1.22
γ (Median of the point estimates)	0.22	0.54
ξ (Median of the point estimates)	14.35	32.96
ω (Median of the posterior distribution)	0.08	0.10
$p(Model = 1)$ (Median of the posterior distribution)	0.49	-
$p(Model = 2)$ (Median of the posterior distribution)	0.30	-
$p(Model = 3)$ (Median of the posterior distribution)	0.03	-
$p(Model = 4)$ (Median of the posterior distribution)	0.03	-
$p(Model = 5)$ (Median of the posterior distribution)	0.05	-
$p(Model = 6)$ (Median of the posterior distribution)	0.10	-
DIC	10877.49	10773.26

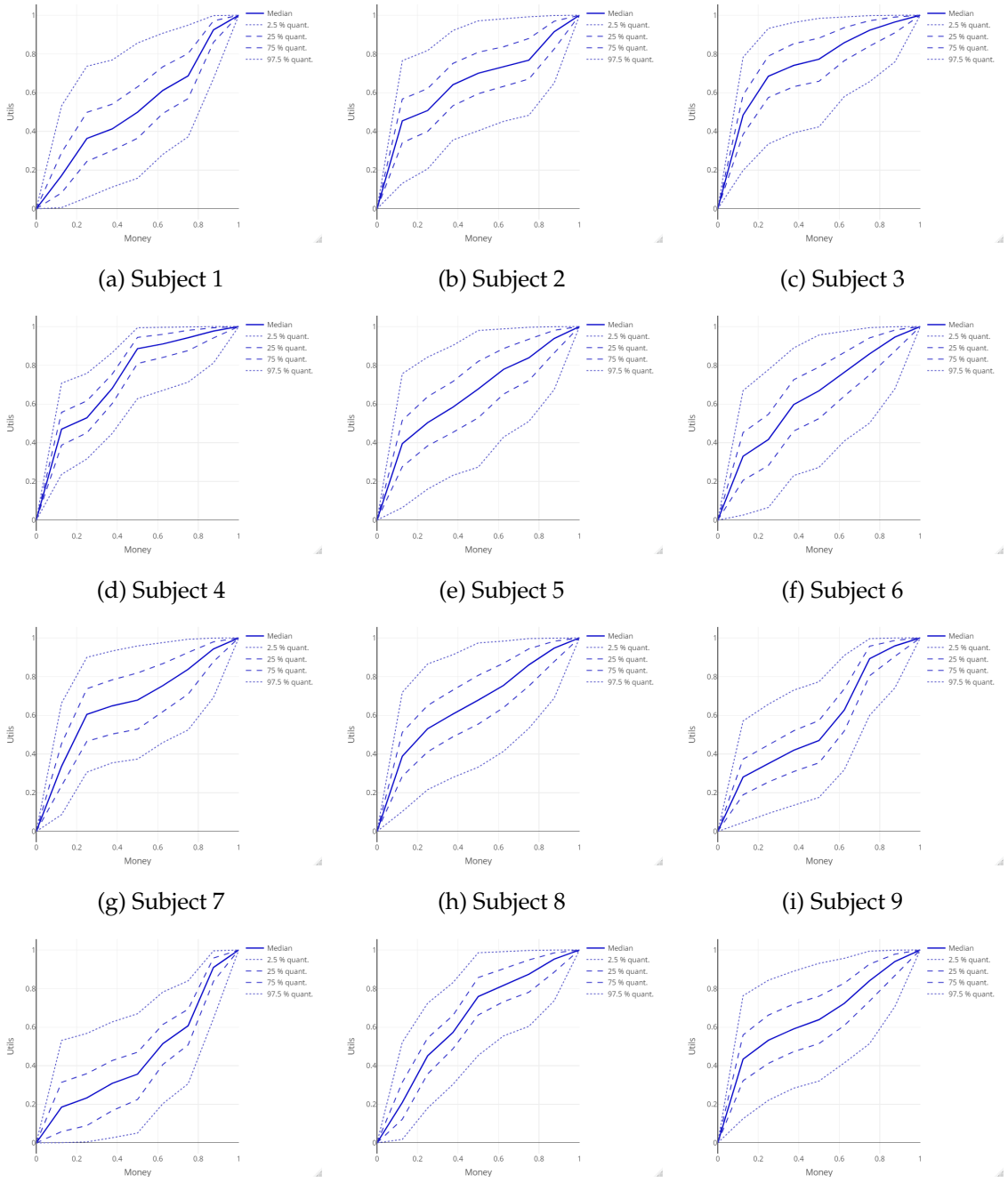
F Details of the MCMC

We use JAGS in R to compute our estimation of the Bayesian models. As one could expect, the models that exhibit poor predictive power may also give convergence issues. This is why we adopt different numbers of iterations in the Gibbs sampling used in JAGS depending on the model under consideration.

- Parametric studies (main paper):
  - Baseline RDU model, RDU with a unique utility function model, RDU with a unique utility function and a unique sensitivity parameter model, and Dual model: 3 chains, 6000 iterations with 2000 iterations in burn-in and one iteration out of four kept in the final sample.
  - RDU with a unique PWF model, RDU with a unique utility function and a unique elevation parameter model, EU model: 3 chains, 30000 iterations, with 2000 iterations in burn-in and one iteration out of tweenty kept in the final sample.
- Parametric studies (robustness checks):
  - Baseline RDU model, RDU with a unique utility function model: 3 chains, 6000 iterations with 2000 iterations in burn-in, and on iteration out of four kept in the final sample.
  - RDU with a unique PWF model: 3 chains, 24000 iterations, with 2000 iterations in burn-in and one iteration out of sixteen kept in the final sample.
- Non-parametric:
  - Baseline RDU model, RDU with a unique utility function model, Dual model: 3 chains, 8000 iterations with 2000 iterations in burn-in and one iteration out of six kept in the final sample.
  - RDU with a unique PWF model and EU model: 3 chains, 16000 iterations, with 4000 iterations in burn-in and one iteration out of twelve kept in the final sample.
- Mixture models (the Baseline RDU model against RDU with a unique utility function model and the replication of Baillon et al., 2020): 5 chains, 5000 iterations with 2000 iterations in burn-in and one iteration out of three kept in the final sample.

At the end of the MCMC process, we verified that the Gelman-Rubin statistic of individual estimates is under or close to 1.20 for all the subjects.

G Examples of non-parametric curves



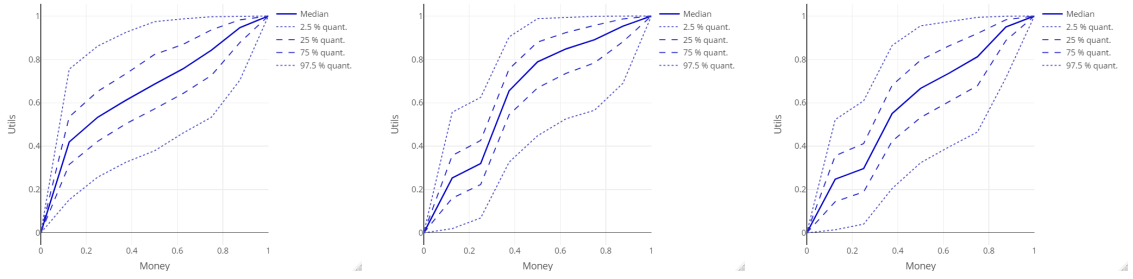
(a) Subject 1 (b) Subject 2 (c) Subject 3

(d) Subject 4 (e) Subject 5 (f) Subject 6

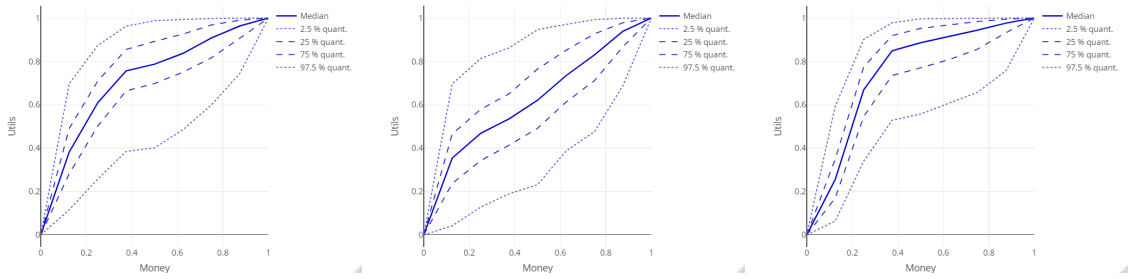
(g) Subject 7 (h) Subject 8 (i) Subject 9

(j) Subject 10 (k) Subject 11 (l) Subject 12

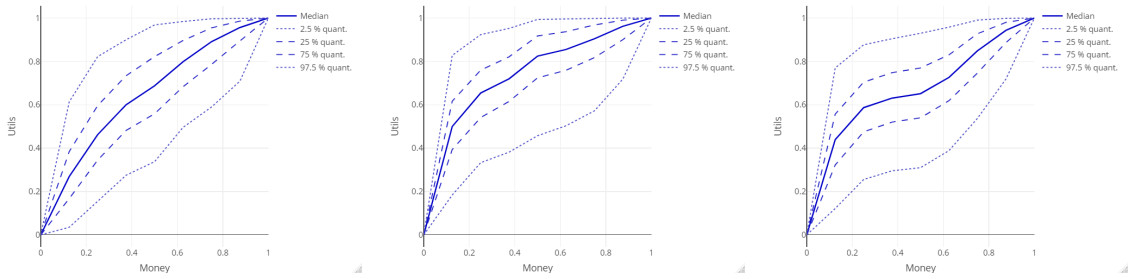
Figure (6) Utility functions (1)



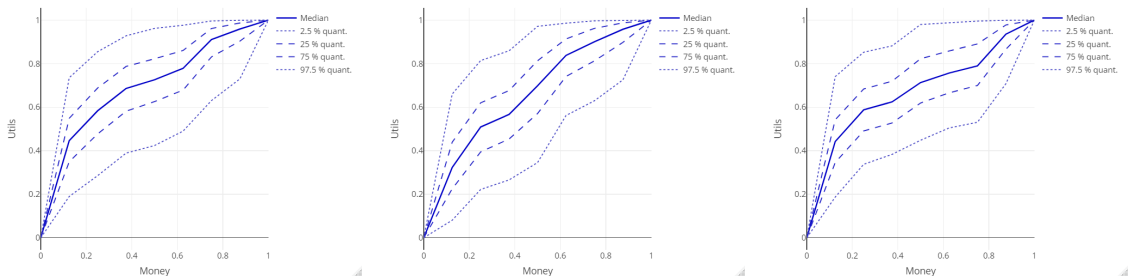
(a) Subject 13 (b) Subject 14 (c) Subject 15



(d) Subject 16 (e) Subject 17 (f) Subject 18

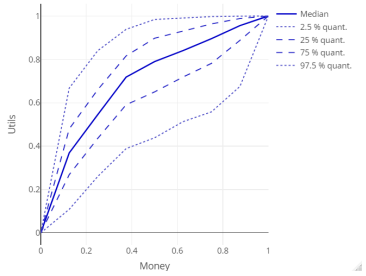


(g) Subject 19 (h) Subject 20 (i) Subject 21

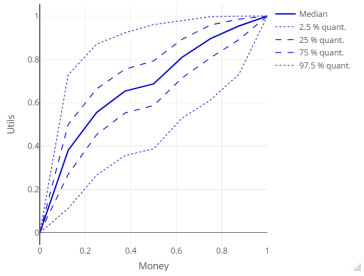


(j) Subject 22 (k) Subject 23 (l) Subject 24

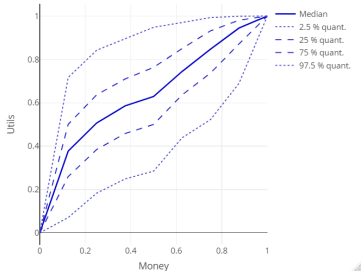
Figure (7) Utility functions (2)



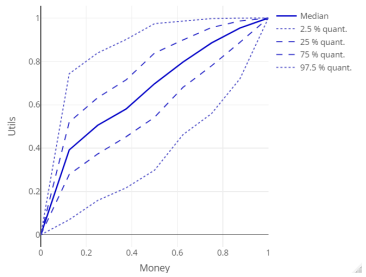
(a) Subject 25



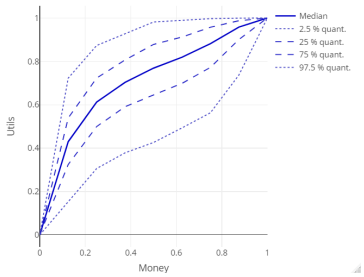
(b) Subject 26



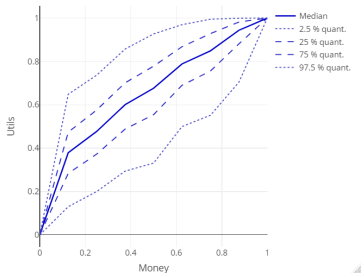
(c) Subject 27



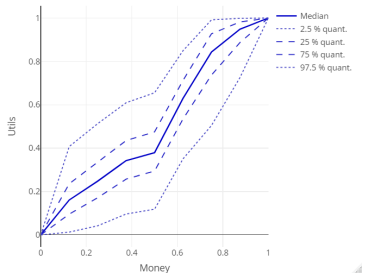
(d) Subject 28



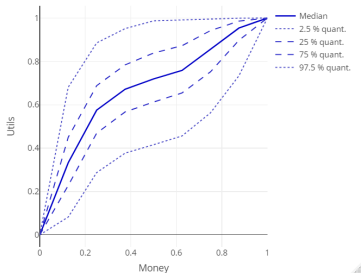
(e) Subject 29



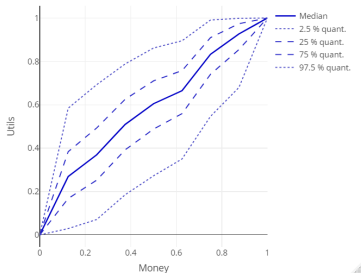
(f) Subject 30



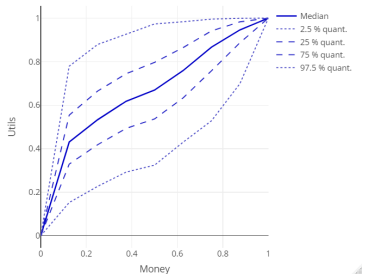
(g) Subject 31



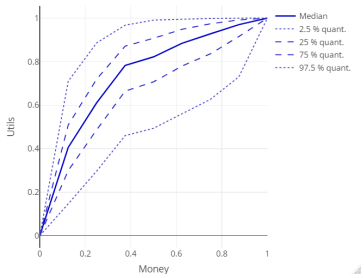
(h) Subject 32



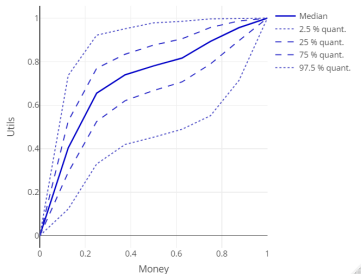
(i) Subject 33



(j) Subject 34

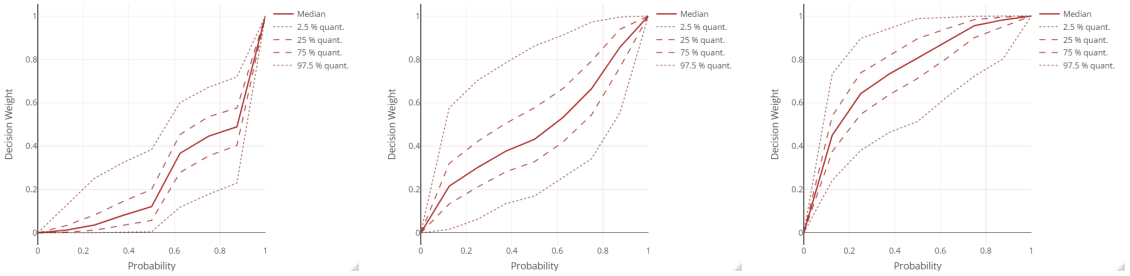


(k) Subject 35

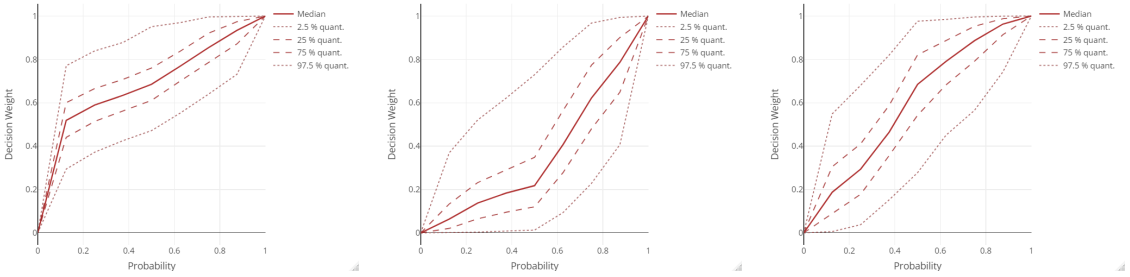


(l) Subject 36

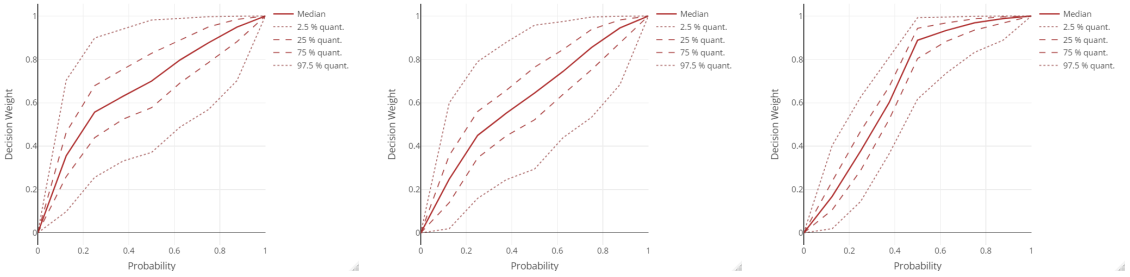
Figure (8) Utility functions (3)



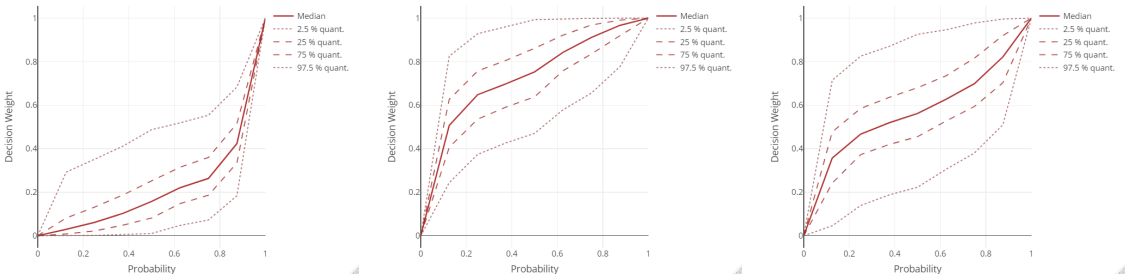
(a) Subject 1 (b) Subject 2 (c) Subject 3



(d) Subject 4 (e) Subject 5 (f) Subject 6

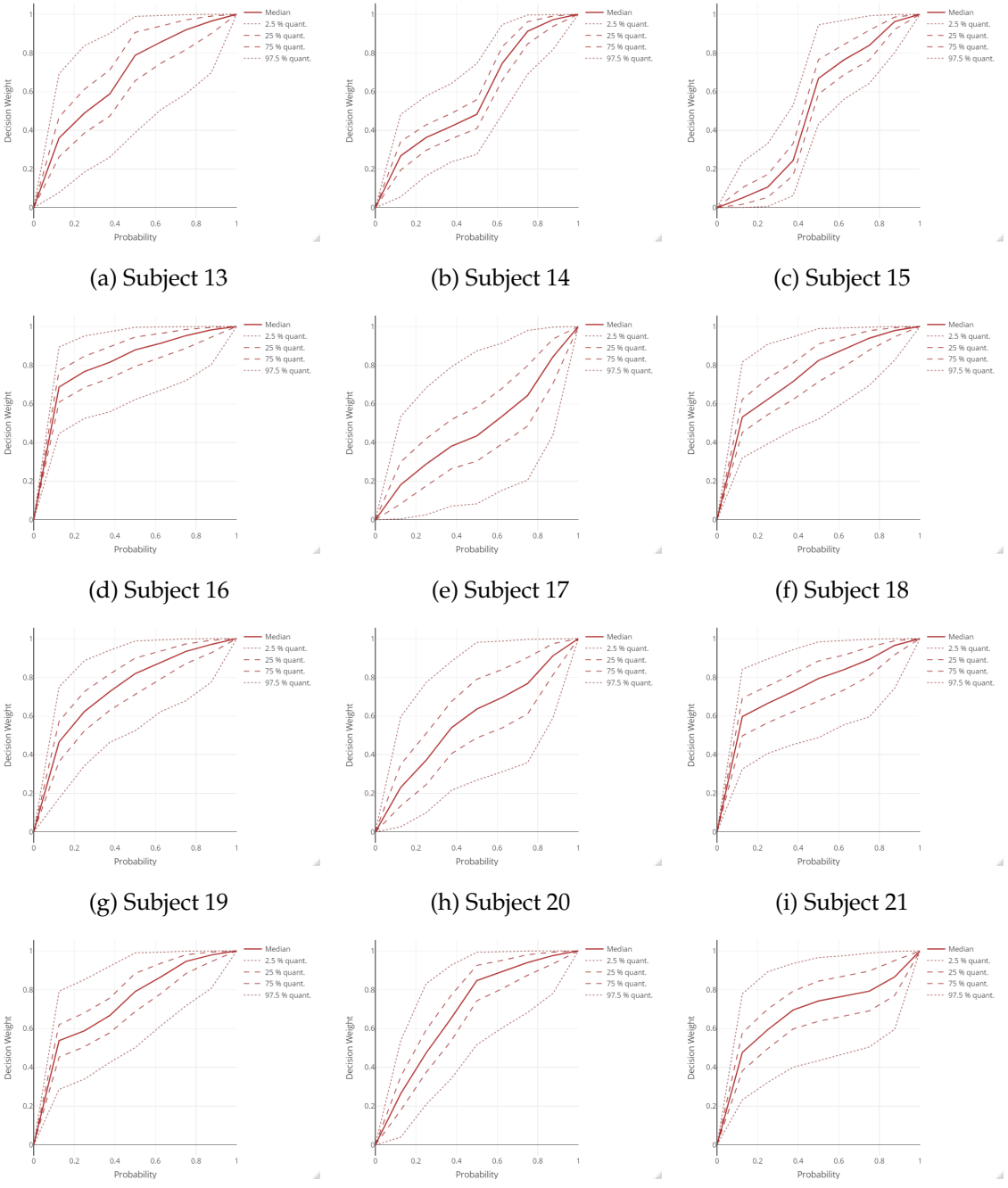


(g) Subject 7 (h) Subject 8 (i) Subject 9



(j) Subject 10 (k) Subject 11 (l) Subject 12

Figure (9) Probability Weighting functions (1)



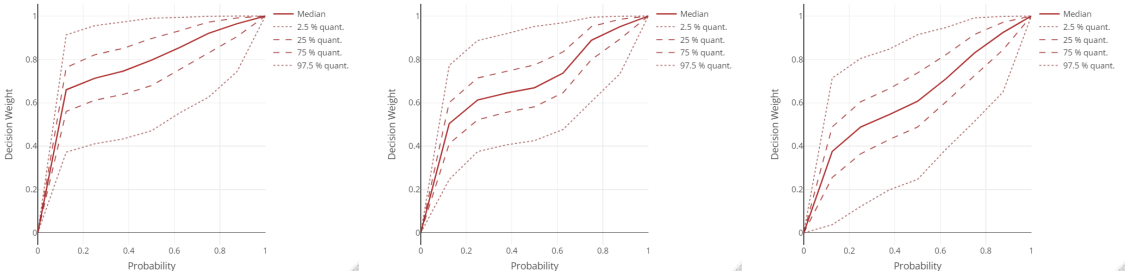
(a) Subject 13 (b) Subject 14 (c) Subject 15

(d) Subject 16 (e) Subject 17 (f) Subject 18

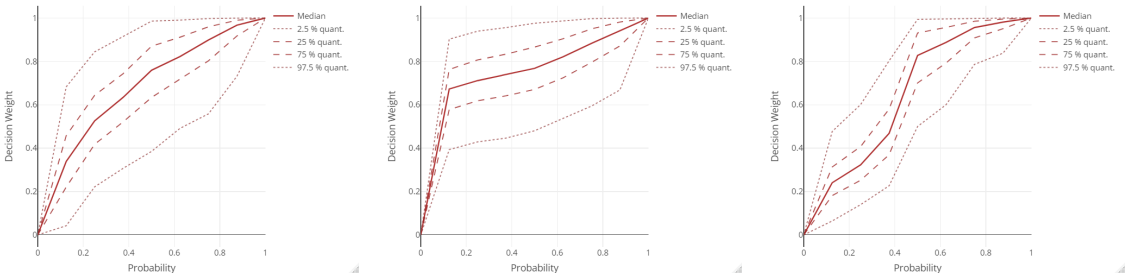
(g) Subject 19 (h) Subject 20 (i) Subject 21

(j) Subject 22 (k) Subject 23 (l) Subject 24

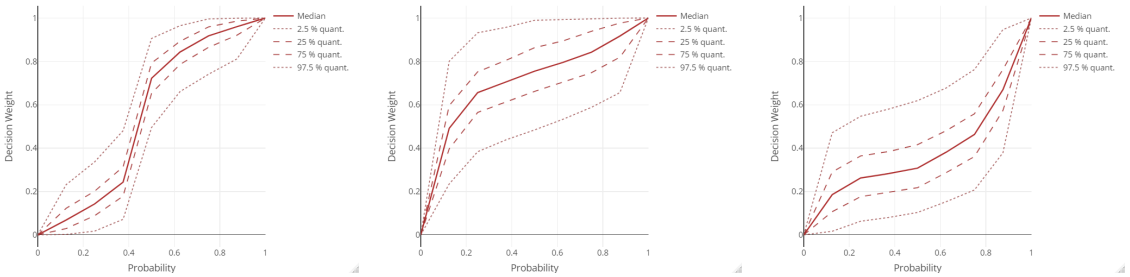
Figure (10) Probability Weighting functions



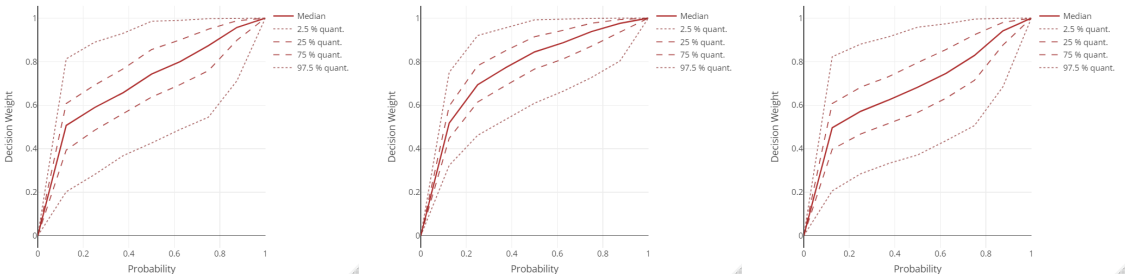
(a) Subject 25 (b) Subject 26 (c) Subject 27



(d) Subject 28 (e) Subject 29 (f) Subject 30



(g) Subject 31 (h) Subject 32 (i) Subject 33



(j) Subject 34 (k) Subject 35 (l) Subject 36

Figure (11) Probability Weighting functions (3)