Measuring Beliefs under Ambiguity

MOHAMMED ABDELLAOUI
HEC Paris

HAN BLEICHRODT
Erasmus School of Economics
Research School of Economics, Australian National University, Canberra

EMMANUEL KEMEL
HEC Paris

OLIVIER L'HARIDON
University of Rennes I

July 2017

Abstract

This paper presents a simple method to measure the beliefs of a decision maker with non-neutral ambiguity attitudes. Our method require three simple measurements, it is incentive-compatible, and it allows for risk aversion and deviations from expected utility, including ambiguity aversion. An experiment using two natural sources of uncertainty (temperature in Rotterdam and in New York City) showed that the estimated beliefs were well-calibrated, sensitive to the source of uncertainty, and similar to the beliefs that were estimated by more sophisticated but time consuming methods.

KEY WORDS: decision under uncertainty, belief measurement, ambiguity.

Acknowledgements: We are grateful to Aurélien Baillon, David Budescu, Enrico Diecidue, Tom Wallsten, Peter P. Wakker, and Horst Zank for their comments on previous versions.

1. Introduction

According to the intellectual fathers of subjective expected utility (SEU), the standard model of rational choice under uncertainty, the decision maker's beliefs can be measured by observing his choices between uncertain bets (Ramsey 1931, Savage 1954). This position was challenged by Ellsberg's (1961) paradox, which showed that SEU leads to inconsistent probabilities. Ellsberg interpreted his paradox as evidence that preferences depend on a third dimension, in addition to the utility of outcomes and the probability of events, which he called ambiguity and which reflects the reliability, credibility, or adequacy of the decision maker's information. Ellsberg's paradox[1] raises the question how beliefs can be measured when SEU does not hold and decision makers may have non-neutral ambiguity attitudes. The purpose of this paper is to address this question and to present a simple method to achieve this.

We study preferences over two-outcome bets giving payoff x if event E obtains and a lower payoff y otherwise and assume that these preferences can be evaluated as

$$f\big(P(E)\big)U(x) + (1 - f\big(P(E)\big)U(y). \tag{1}$$

In Model (1), U is a (strictly increasing) utility function, P is a probability measure over events that reflects the decision maker's beliefs, and f is a (strictly increasing) function that measures Ellsberg's third dimension, the deviations from subjective expected utility including attitudes towards ambiguity. SEU is the special case of (1) where f is the identity function and the decision maker is neutral towards ambiguity. If f is convex, which reflects pessimism,[2] model (1) is consistent with the paradoxes of Allais

[1] For a recent overview of the experimental evidence see Trautmann and Van de Kuilen 2016).

[2] For instance, in the Ellsberg two color paradox where a ball is randomly drawn from an unknown urn containing 100 Black and Red balls in unknown proportion, most people are indifferent between betting on red or on black, meaning that $P(B) = P(R) = 1/2$. Nevertheless,

(certainty effect) and Ellsberg (ambiguity aversion). Model (1) combines the biseparable preference model of Ghirardato and Marinacci (2001), which contains many of the ambiguity models that have been proposed to explain Ellsberg's paradox as special cases,[3] with the assumption that the decision maker can assign subjective probabilities to events even when he does not maximize SEU. Chew and Sagi (2006) proposed a behavioral foundation for this assumption and Abdellaoui et al. (2011) obtained experimental support for it.

Our method measures beliefs without the interference of ambiguity attitudes based on three measurements only. It is incentive-compatible, uses simple choice-lists as in the measurement of utility under risk (Holt and Laury 2002), and can easily be applied in empirical research. Unlike the method in Abdellaoui et al. (2011), our method does not use chained responses (i.e. previous elicitations are not used in subsequent elicitations) and, hence, it is not vulnerable to error accumulation and the possibility of strategic responding. Our method is based on the elicitation of exchangeable events. Our estimations show that this leads to substantially lower error rates than by the elicitation of probability equivalences.

We applied our method in an experiment using two natural sources of uncertainty, the temperatures in Rotterdam and in New York City at a future date. The elicited subjective distributions of beliefs were well-calibrated, sensitive to the source of

a pessimistic decision maker (convex f) will behave as if the "probability" of the winning color is less than one half.

[3] Examples are Choquet expected utility (Schmeidler 1989), maxmin expected utility (Gilboa and Schmeidler 1989), α-maxmin expected utility (Ghirardato et al. 2004), contraction expected utility (Gajdos et al. 2008), and prospect theory for gains and losses separately (Tversky and Kahneman 1992). A well-known ambiguity model that is not a special case of biseparable preferences is the smooth ambiguity model (Klibanoff et al. 2005).

uncertainty, and reflected much individual heterogeneity. Our method is deterministic and can measure beliefs without the need to measure utility and ambiguity attitudes. To test the robustness of our measurements we also allowed for the stochastic nature of people's judgments and we performed joint estimations of beliefs, utility, and the function f of model (1). The measured beliefs were similar suggesting that our simple method indeed leads to reliable measurements of people's beliefs.

2. The measurement of beliefs

2.1. Notation

Consider a decision maker who faces uncertainty. Uncertainty is modeled through a *state space* S. Exactly one of the states will obtain, but the decision maker does not know which one. *Events* E are subsets of S. E^c is the complement of event E.

The decision maker chooses between two-outcome bets $x_E y$ that pay €x if event E occurs and a lower payoff €y otherwise. The decision maker's preferences over bets are evaluated by model (1). In Model (1), the *utility function* U is an interval scale and we will set $U(0) = 0$. P is a *probability measure* that represents the decision maker's beliefs. The decision maker's beliefs are transformed by the strictly increasing *distortion function* f, that maps subjective probabilities onto [0,1] and that reflects amongst other things the decision maker's ambiguity attitudes.

This paper concentrates on the measurement of P. We will present three, increasingly sophisticated methods to do so. Our first method is deterministic. It measures the median and the dispersion of a decision maker's beliefs using three simple measurements and uses these to estimate the distribution of P. The second method allows for the stochastic nature of people's preferences. The third method is also stochastic and estimates, besides P, also utility U and the distortion function f.

2.2. Deterministic measurement of the median and the dispersion of beliefs

We first specify an interval $[a, b]$ of possible values of a given random variable. In our experiment we studied two sources of uncertainty, the temperatures (in degrees Celsius) in Rotterdam and in New York City on January 15, 2013 at 2pm, and we used $a = -50, b = +50$. The width of the interval $[a, b]$ is irrelevant as long as it contains all values that the decision maker considers possible.[4]

The measurement then proceeds in three elicitations, which are explained in Table 1. First, we measure the median of the distribution by subdividing $[a, b]$ into two equally likely subintervals. In questions two and three we then measure the dispersion of the distribution by subdividing $[a, 0]$ and $[0, b]$ into two equally likely subintervals.

Table 1: Deterministic measurement of the distribution of beliefs

Question	Assessed quantity	Implication
1 Median	z_M: $x_{[a,z_M]}0 \sim x_{[z_M,b]}0$	$P([a, z_M]) = P([z_M, b])$
2 Dispersion	z_L: $x_{[a,z_l]}0 \sim x_{[z_L,0]}0$	$P([a, z_L]) = P([z_L, 0])$
3	z_R: $x_{[0,z_R]}0 \sim x_{[z_R,b]}0$	$P([0, z_R]) = P([z_R, b])$

Question 1: Measuring the median

To measure the median of the distribution, we elicited z_M such that $x_{[a,z_M]}0 \sim x_{[z_M,b]}0$, where $[a, z_M]$ means the temperature lies between a and z_M. This indifference implies

[4] In recorded history temperatures have never been close to -50^oC or $+50^oC$ on January 15, neither in Rotterdam nor in New York City. No subject believed these values were possible, as reflected by their answers.

that the events $[a, z_M]$ and $[z_M, b]$ are *exchangeable* (Ramsey 1931, de Finetti 1937). Substitution in Model (1) gives $P([a, z_M]) =$

$P([z_M, b])$ because f is increasing and $U(x)$ cancels out. It follows that z_M is the median of the distribution of beliefs.

Questions 2 and 3: Measuring the dispersion

To measure the dispersion of the distribution, we need two indifferences. We measured z_L such that $x_{[a,z_L]}0 \sim x_{[z_L,0]}0$, and $\mathrm{z_R}$ such that $x_{[0,\mathrm{z_R}]}0 \sim x_{[\mathrm{z_R},b]}0$. From (1), it follows that $P([a, z_L]) = P([z_L, 0]) = \frac{P([a,0])}{2}$ and $P([0, \mathrm{z_R}]) = P([\mathrm{z_R}, b]) = \frac{P([0,b])}{2}$. Adding $P([a, 0])$ to all terms in the second equality and using the additivity of P gives $([a, \mathrm{z_R}]) = \frac{P([a,b])+P([a,0])}{2}$. Because $P([a, z_L]) = \frac{P([a,0])}{2}$ we obtain $P([a, \mathrm{z_R}]) - P([a, z_L]) = \frac{P([a,b])}{2} = 1/2$, which gives a measure of the dispersion of the distribution of beliefs.[5]

We could have immediately measured the dispersion of the distribution of beliefs from z_M by asking for the indifferences $x_{[a,z_L]}0 \sim x_{[z_L,z_M]}0$ and $x_{[z_M,\mathrm{z_R}]}0 \sim x_{[\mathrm{z_R},b]}0$. Then z_L and $\mathrm{z_R}$ equal the 25% and the 75% quantiles of the distribution of beliefs. However, this would make our method chained and may lead to error accumulation (errors made in one questions affect the responses to later questions) and strategic responding (subjects can affect the questions they face at a later stage). We wanted to avoid these problems and therefore used only non-chained measurements.

[5] The above analysis assumes that 0 is in $[a, b]$ and that both $[a, 0]$ and $[0, b]$ have nonzero probability mass. In our setting this makes sense as January temperatures around 0 degrees Celsius (32 degrees Fahrenheit) are common in Rotterdam and in New York City. If 0 may not be in $[a, b]$ then we could use intervals $[a', b']$ and $[a'', b'']$ with $a \leq a', a''$ and $b', b'' \leq b$ and elicit indifferences $x_{[a',z_L]}0 \sim x_{[z_L,b']}0$ and $x_{[a'',z_R]} \sim x_{[z_R,b'']}0$ to obtain information on the dispersion of the distribution of beliefs.

The median and the dispersion permit measuring each individual belief distribution using only three indifferences. To account for skewness in the distribution we assumed a beta distribution, which is consistent with many distributions. To estimate the beta distribution requires specifying the minimum and the maximum possible temperatures. For these bounds we used -50 and +50. The assumption of a common minimum and maximum was necessary to be able to aggregate estimates across individual subjects. The values of the estimated beta distribution are closely associated with the bounds of the distribution. In the online appendix we present two robustness checks, one in which the bounds are elicited deterministically for each subject separately and one in which the bounds were part of the estimated stochastic model.

2.3. Stochastic measurement of beliefs

The method described in Section 2.2. is deterministic and assumes that decision makers make no errors. To account for the stochastic nature of human decision making, we added an error term ϵ to the indifference values z predicted by Model (1), with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In practice, for a given choice list, with an interval $[a, b]$ of possible values , indifference value z was bracketed by two values z^+ and z^-such that z_L: $x_{[a,z^+]}0 \succ x_{[z^+,b]}0$ and $x_{[a,z^-]}0 \prec x_{[z^-,b]}0$. The error specification implies that the likelihood of the observations provided by a given choice list k is equal to:[6]

$$\ell(\theta|z_k^+, z_k^-, X_k) = \Phi\left(\frac{z_k^+ - z_k(\theta, X_k)}{\sigma}\right) - \Phi\left(\frac{z_k^- - z_k(\theta, X_k)}{\sigma}\right) \tag{2}$$

[6] In case no switch between the two options occurred, the likelihood was equal to either $\Phi\left(\frac{z_k^+ - z_k(\theta, X_k)}{\sigma}\right)$ or $1 - \Phi\left(\frac{z_k^- - z_k(\theta, X_k)}{\sigma}\right)$.

where the index k refers to values in choice list k and $z_k(\theta, X_k)$ is the indifference value predicted by Model (1) for a distribution of beliefs with parameters θ and choice list characteristics $X_k = (x_k, y_{k,} a_k, b_k)$. As long as the same x and the same y are used in all choices then their utilities cancel out and we can still estimate beliefs separately (as in the deterministic approach).

2.4. The complete measurement of Model (1)

The measurements described in Sections 2.2. and 2.3 are robust to utility and ambiguity aversion. However they do not estimate U and f separately. The third method measures P, U, and f jointly. In addition to the measurements reported in Section 2.3, we elicited a series of values z such that the decision maker was indifferent between $x_{[z,b]}y$ and money amount c for sure, $x \geq \text{c} \geq y$.[7] From these indifferences, we could estimate U and f by imposing parametric assumptions. The required number of questions depends on the parametric assumptions made.

We measure U and f by varying the size of the events $[z_i, b]$. Because $P([z_i, b]) \geq P([z_i^{'}, b])$ if $z_i \leq z_i^{'}$, our measurements can be interpreted as eliciting indifferences by varying subjective probabilities. Probability equivalences (eliciting indifferences by varying objective probabilities) are widely-used in decision under risk (Farquhar 1984, Holt and Laury 2002). For example, in health economics quality of life is often measured using probability equivalences (Drummond et al. 2015). Our measurements extend the probability equivalence method to decision under uncertainty.

[7] Our method remains valid if c is an uncertain bet instead of a certain outcome.

Under expected utility (f the identity function), we can measure utility using one single indifference by assuming, for example, a constant relative risk aversion (CRRA) utility function. Hence, we can measure utility under uncertainty the same way that Holt and Laury (2002) measured utility under risk.[8]

3. Experiment

3.1. Design

The experiment was run in December 2012. We recruited 82 students from Erasmus University. They received a show-up fee of €5. In addition, each subject played out one of his choices for real. Subjects' total payoffs ranged from €5 to €55 with an average of just over €30. The experiment lasted about one hour.

The experiment was computer-run in small sessions of three subjects. We used small sessions to improve the quality of the data. Subjects first received instructions and then answered several training questions. The training questions took 10 minutes. We included this extensive training to make sure that subjects understood the tasks. We told subjects that there were no right or wrong answers and that we were just interested in their preferences. We encouraged them to ask questions at any time they wished.

We used two sources of uncertainty: the temperatures in Rotterdam and in New York City on 15 January 2013 at 2pm local time. We expected that subjects would have a better knowledge of the temperature in Rotterdam. Previous evidence suggests that people are less ambiguity averse for sources of uncertainty that they feel more competent about (Heath and Tversky 1991, Abdellaoui et al. 2011).

[8] A CRRA utility function with power γ gives $\gamma = log\left(P([z,b])\right)/log\ \left(\frac{c-y}{x-y}\right)$.

Table 2 shows the experimental choice questions. The first 3 choices were used to measure the distribution of beliefs under the assumption of deterministic preferences. The first choice determines the median of the belief distribution, the second and third the dispersion. The fourth and fifth determined the minimum and maximum values of the temperatures in Rotterdam and in New York that the subject considered possible, which were used in the robustness checks reported in the online appendix.

Choices 1, 2, 3, and 6-12 were used to measure beliefs allowing for stochastic preferences. In these choices the outcomes were always €50 and €0 and, by setting $U(€50) = 1$ and $U(€0) = 0$, they cancelled out. Moreover, in these choices we always measured exchangeable events and, because the probability distortion function is increasing, it follows that that the subjective probabilities of these exchangeable events are equal. Choices 4,5, and 14-24 measured the complete stochastic Model (1).

We randomized the order of the questions except that the belief questions always came before the questions that measured the utility and the distortion functions.[9] To test for consistency, we repeated choices 1 and 18 for both the Rotterdam and the New York temperature.

Table 2. The experimental questions and the mean responses.

Choice	Indifference	Purpose	Mean z Rotterdam	Mean z New York
1	$50_{[-50,z]}0 \sim 50_{[z,+50]}0$	Median	2.60	0.23
2	$50_{[-\infty,z]}0 \sim 50_{[z,0]}0$	Dispersion	−3.65	−5.00
3	$50_{[0,z]}0 \sim 50_{[z,+50]}0$		4.91	4.73
4	$49 \sim 50_{[z,50]}0$	Minimum	−21.06	−24.60
5	$1 \sim 50_{[z,50]}0$	Maximum	15.65	16.89
6	$50_{[-10,z]}0 \sim 50_{[z,10]}0$	Stochastic	1.67	−0.24
7	$50_{[-50,z]}0 \sim 50_{[z,-5]}0$		−6.80	−8.18

[9] Hence, questions 4 and 5 always came before questions 14-24 in the estimation of the complete Model (1). The results were not affected when these two questions were removed.

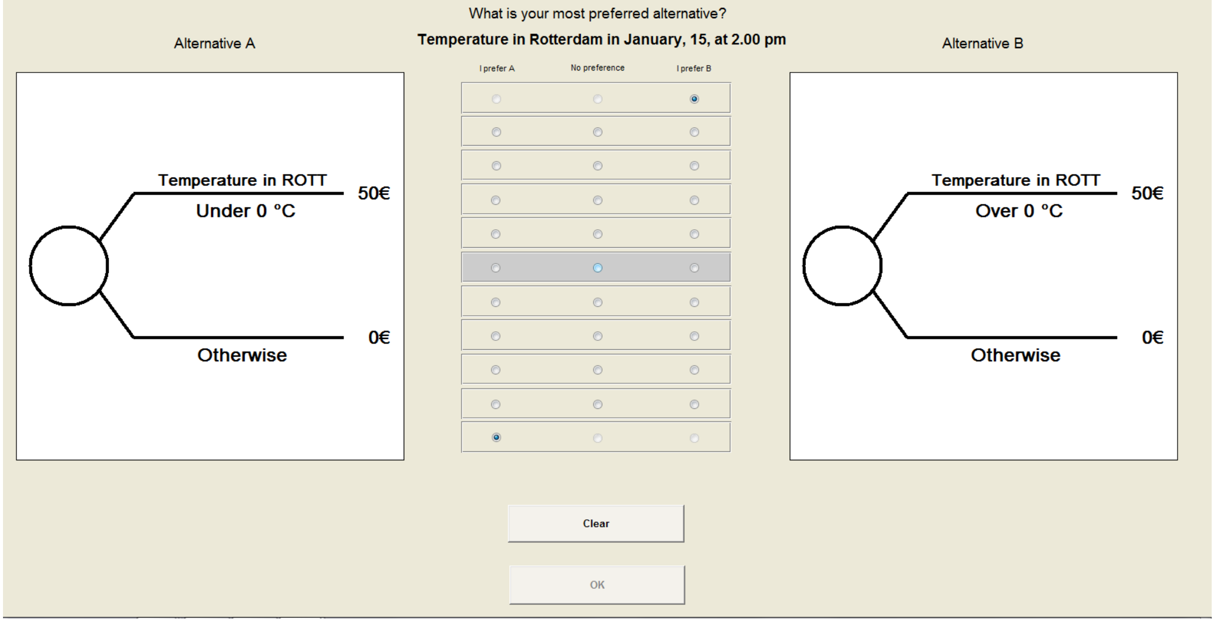
8	$50_{[+5,z]}0\sim 50_{[z,+50]}0$	Beliefs	7.85	7.84
9	$50_{[-15,z]}0\sim 50_{[z,+5]}0$		−0.77	−2.61
10	$50_{[-5,z]}0\sim 50_{[z,+15]}0$		3.16	2.34
11	$50_{[-50,z]}0\sim 50_{[z,-10]}0$		−11.94	−12.46
12	$50_{[+10,z]}0\sim 50_{[z,+50]}0$		11.72	12.34
13	$10_{[-50,z]}0\sim 10_{[z,+50]}0$	Test Model (1)	2.57	−0.21
14	$5\sim 50_{[z,+50]}0$	Utility and Weighting Functions	6.44	6.20
15	$10\sim 50_{[z,+50]}5$		3.59	2.37
16	$25\sim 50_{[z,+50]}15$		−0.44	−2.65
17	$20\sim 50_{[z,+50]}5$		−1.67	−4.70
18	$25\sim 50_{[z,+50]}0$		−3.44	−6.68
19	$35\sim 50_{[z,+50]}20$		−3.23	−6.20
20	$5\sim 10_{[z,+50]}0$		−1.05	−2.38
21	$10\sim 15_{[z,+50]}0$		−4.39	−8.55
22	$40\sim 50_{[z,+50]}20$		−8.13	−12.49
23	$40\sim 50_{[z,+50]}5$		−10.35	−15.09
24	$45\sim 50_{[z,+50]}0$		−16.13	−19.50
1 rep.	$50_{[-50,z]}0\sim 50_{[z,+50]}0$	Consistency	3.05	−0.02
18 rep.	$25\sim 50_{[z,+50]}0$		−2.49	−6.07

Note: 1 rep. (18 rep.) denotes the repetition of choice 1 (18). The first column denotes the number of the choice, the second the indifference that was measured, the third the purpose of this indifference (e.g. median means it served to elicit the median of the distribution of beliefs), and the fourth and fifth column denote the mean values of z in Rotterdam (fourth column) and New York City (fifth column).

Several choices gave additional information about subjects' preferences. Question 13 tested Model (1), which underlies all our measurements. It was similar to question 1 except that the highest payoffs were €10 instead of €50. The elicited indifference $10_{[-50,z]}0\sim 10_{[z,+50]}0$ implies by Model (1) that $P([-50,z]) = P([z,50])$. Consequently, Question 13 gave another measurement of the median of the distribution of beliefs, which, provided Model (1) is correct, should be equal to the measurement of the median obtained in Question 1, except for random error. Different responses would signal a violation of Model (1).

Choices 18 and 20 tested whether utility belonged to the power (constant relative risk aversion) family. The payoffs of choice 18 were five times as large as those of choice 20 and under constant relative risk aversion the elicited z-values should be the same.

Figure 1: Display of the choice lists



What is your most preferred alternative?

Temperature in Rotterdam in January, 15, at 2.00 pm

We used choice lists to elicit the indifference values. Each choice list contained 9 choices. Figure 1 gives an example of a choice list used in the exchangeability questions. In the first choice on the list subjects compared $A = x_{[a',a']}y$ and $B = x_{[a',b']}y$, i.e. z was equal to a', and subjects should prefer bet B. The value of z increased by $\frac{b'-a'}{8}$ in each following choice on the list, increasing the attractiveness of A, until z was equal to b' in the final choice where subjects should prefer A. We imposed the choice in the first and final choice of the list and asked subjects to complete the choice list. The program imposed monotonicity: if a subject preferred $A = x_{[a',z]}y$ to $B = x_{[z,b']}y$ for some value of z then the program automatically selected A for all $z' > z$. Figure 1 shows a choice where the value of z is zero. Subjects could either choose A or B or click on "no preference". If subjects selected "no preference" for some value of z then we used this value as their indifference value. However, this happened in only 0.8% of the choice lists. Most subjects switched at some point from B to A. We then used the midpoint of the highest value of z^{+} for which they preferred B and the lowest value of z^{-} for which they preferred A as their indifference value.

Subjects made 364 choices in total. At the end of the experiment, one of these 364 choices was selected randomly and played out for real. If a subject had chosen A or B in this choice then we played out his preferred bet. If the subject had selected "no preference" then we randomly selected one of the two bets to be played out for real. Ambiguity averse subjects may prefer this randomization to one of the two bets. However, as mentioned before, very few subjects chose the option "no preference" and there is no reason to suspect that such a preference for randomization has affected the results.

4.2. Analysis

We assumed that the beliefs followed a beta distribution. We also tried other distributions (Gaussian, triangular) but these led to a worse fit.

In the stochastic approach, we assumed that utility belonged to the exponential (constant absolute risk aversion) family. We also performed the analyses using the CRRA family, but this led to a decrease in goodness of fit both at the aggregate level and for 57% of our subjects. Moreover, the responses to choices 18 and 20 indicated that constant relative risk aversion did not hold. We used the same utility function for both sources of uncertainty. Abdellaoui et al. (2011) found support for the assumption that utility is relatively stable across sources of uncertainty.

The distortion function was estimated using Prelec's (1998) two-parameter specification:

$$f\big(P(E)\big) = \exp\,(-d(-\ln(P(E)))^{g}). \tag{3}$$

The parameter g mainly reflects the insensitivity of the decision maker to likelihood information (Abdellaoui et al. 2011). The parameter d measures the decision maker's pessimism/optimism with higher values indicating more pessimism. Dummies were added to test for source dependence of g and d. We used different error terms for the exchangeability tasks 1,2,3, 6-13 and for the probability equivalence tasks 4,5, 14-24. At the individual level, we estimated the model parameters by maximizing the sum of the log likelihoods defined by Eq. (2) over all choices, using the BFGS algorithm. To prevent that the estimated values were based on a local optimum, we used 50 randomly distributed sets of starting values for the parameters. We did this both at the aggregate level by pooling the individual choices and for each subject separately. The individual results are reported in the Appendix.

At the aggregate level we estimated a random coefficients model. Instead of estimating each individual parameter separately, the random coefficients model estimates the parameters of the population-level distribution from which the individual parameters are drawn. Hence, the estimation for each subjects borrows strength from other subjects to obtain a more powerful analysis leading to more precise and less biased estimates (Kreft & de Leeuw 1998). The online appendix describes the random coefficients estimation in detail.

We used likelihood ratio tests to test the goodness of fit of nested models and the Bayes Information Criterion otherwise.

An important question that we seek to address is whether our simple method elicits the same beliefs as the more sophisticated stochastic methods. If so, this would make the method suitable to use in empirical elicitations of beliefs. Hence, our main interest is to test for equalities of parameters. Classic significance tests are less suitable for this as they do not allow to state evidence for the null and they overstate the evidence against the null (Rouder et al. 2009). Hence, we used Bayes factors (BF) instead. Bayes factors indicate how much more likely the alternative is than the null. A Bayes factor of 10 indicates that the alternative is 10 times as likely as the null given the data. Conversely, a Bayes factor of 0.10 indicates that the null is 10 times as likely as the alternative given the data. We used the common interpretation that a Bayes factor larger than 3 signals some support for the alternative over the null, a Bayes factor larger than 10 signals strong support for the alternative over the null, and a Bayes factor larger than 30 signals very strong support for the alternative over the null. Similarly, a Bayes factor less than 0.33 signals some support for the null over the alternative, a Bayes factor less than 0.10 signals strong support for the null over the alternative, and a Bayes factor less than 0.03 signals very strong support for the null over the alternative. To check for robustness we also performed classic statistical tests. These led to the same conclusions and are reported in the online appendix. The online appendix also contains the results from other statistical tests that we compared and that are not reported in the paper.

5. Results

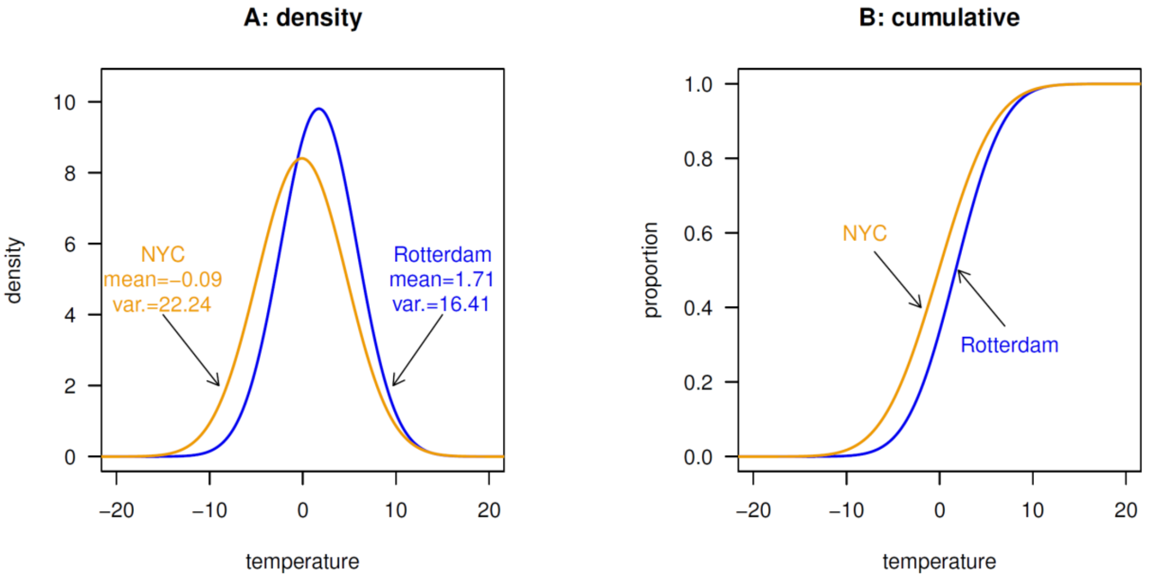
5.1. Consistency

Table 2 shows the mean indifference values that we elicited. The consistency was good. We repeated four elicitations, one exchangeability question and one probability equivalence question for both sources, temperature in Rotterdam and temperature in New York. A Bayesian Anova showed support for the null that the original and the repeated elicited values were the same over the alternative that they differed (all $BF < 0.28$). The Spearman correlations between the original and the repeated measurements were substantial (all $\varrho > 0.67$).

A critical assumption underlying our analysis is that Model (1) holds. Under Model (1) we should find the same values of z in Question 1 and 13 and, indeed, Bayesian t-tests revealed support for the null that the elicited values of z were the same in these choices (both $BF < 0.15$).

Choices 18 and 20 tested the CRRA utility function, which is widely used in empirical economics. For Rotterdam a Bayesian t-test was inconclusive ($BF = 1{,}7$) but for New York temperature we found very strong evidence that the CRRA utility function did not hold (BF=482.7).

Figure 2: Density and cumulative distribution of beliefs. Deterministic approach



5.2. Deterministic measurement of beliefs

The deterministic approach uses only the first three elicitations of Table 2 to estimate the distribution of beliefs. The first elicited value of z, which measures the median of the distribution, was higher for Rotterdam than for New York (BF=9.4) signaling that subjects took account of the differences between the sources and expected higher temperatures in Rotterdam than in New York City. The second and third values of z measure the dispersion of the distributions. The test of the second value was inconclusive ($BF = 1.1$), but we found support for the null that the third value of z was the same for Rotterdam and New York ($BF = 0.16$).

Figure 2 shows the estimated density (panel A) and distribution functions (panel B) of the beliefs for Rotterdam and New York temperatures under the deterministic approach. based on the mean values obtained in the first three measurements. The estimated distributions reflect that subjects expected lower temperatures in New York than in Rotterdam. A Bayesian analysis showed very strong support for the hypothesis that the means of the distributions for Rotterdam and New York differed ($BF = 39.6$). However, we also found support for the null that the variances were the same ($BF = 0.21$).

Figure 3: Comparison between beliefs and the empirical distribution function

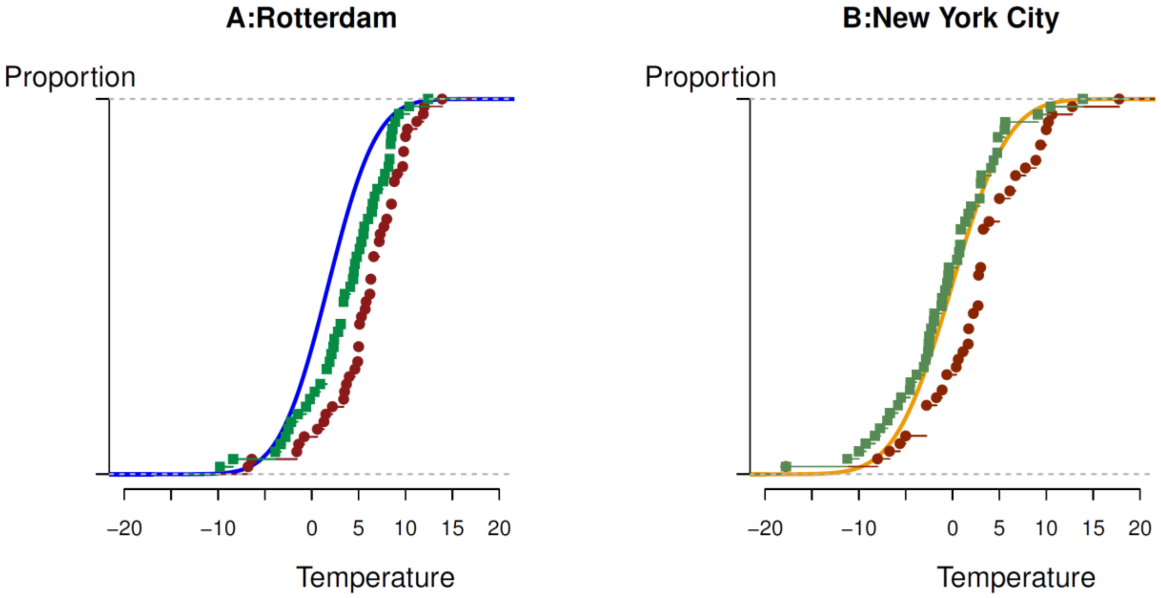
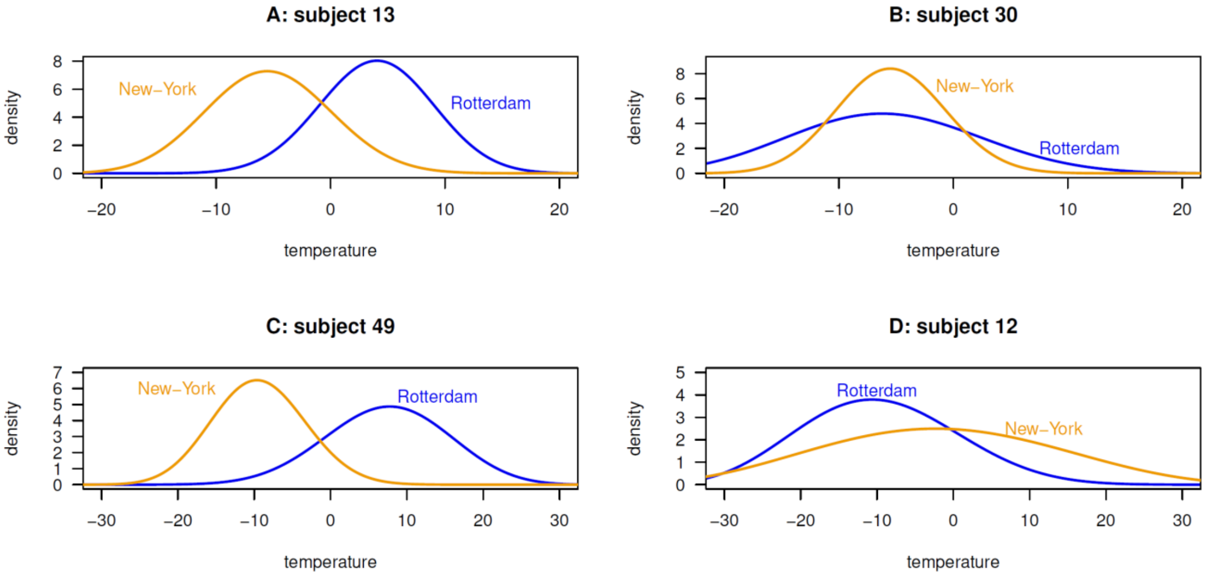


Figure 3 compares the estimated distribution functions and the empirical distribution functions of the average and maximum temperatures over the past 50 years for Rotterdam (Panel A) and New York (Panel B). We did not have the exact historical data of the temperature on 2pm, but it is probably close to the maximum temperature. Subjects' beliefs were generally well-calibrated. Both the means and the medians of the elicited distributions were close to their historical values. The shapes of the elicited distributions were also similar to those of the historical distributions. If subjects answered randomly or if our method would introduce biases then we would expect systematically different shapes. Panel A shows that beliefs about the temperature in Rotterdam were slightly low compared with the historical data. This may be explained by availability bias (Tversky and Kahneman 1973) as the winters in Rotterdam of 2009-2012 were relatively cold. For New York City (Panel B) beliefs were close to the historical data.

Figure 4: Individual belief distributions



The individual belief distributions differed considerably. To illustrate, Figure 4 shows the belief distributions for temperatures in Rotterdam and New York of four of our subjects. Panel A shows the distributions of subject 13, which have the same variance but different means. These distributions look like Gaussian distributions. For subject 30 (Panel B) the distributions have almost equal means, but the variance is higher for Rotterdam temperature. Subject 49 (Panel C) has negatively skewed distributions of beliefs. Subject 12 (Panel D) has a Gaussian distribution of beliefs about temperature in Rotterdam, but a close to uniform distribution of beliefs for temperature in New York City.

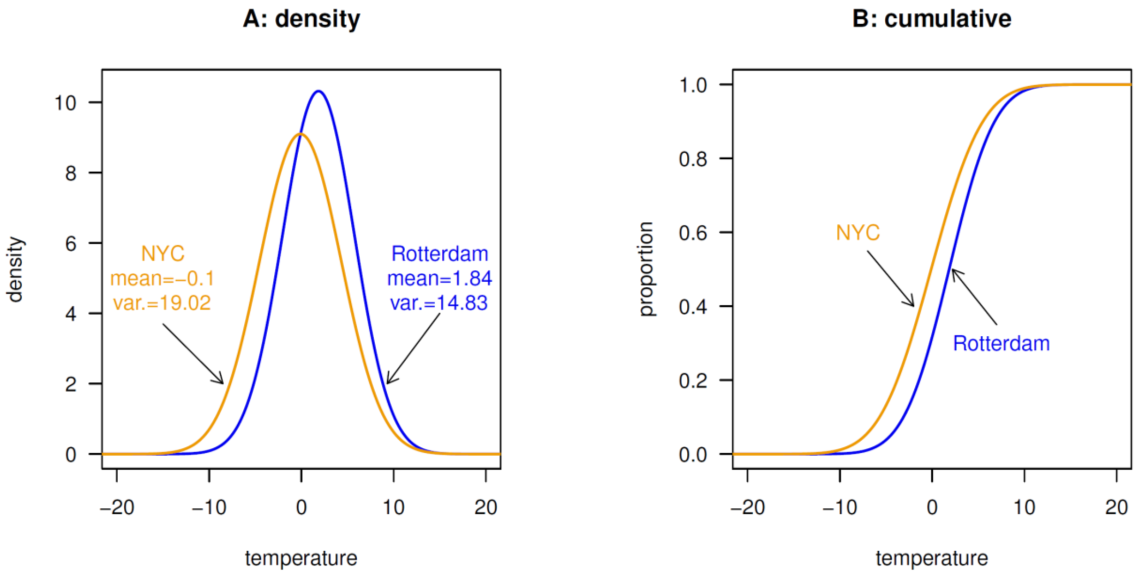
5.3. Stochastic measurement of beliefs

The stochastic approach used 10 elicitations to measure the beliefs distribution. The main purpose of this analysis was to test the robustness of the results from the deterministic approach when we allowed for response error.

We observed no substantial differences with the results from the deterministic approach. Figure 5 shows the density and cumulative distribution function based on the means of the individual estimated parameters of the beta distribution. They are close to those that were obtained under the assumption of deterministic preferences. For Rotterdam we found support for the null that the estimated parameters of the beta distribution were the same under the deterministic and the stochastic approach (both $BF < 0.21$). For New York we also found support or the null but it was slightly weaker (both $BF = 0.33$ for α and $BF = 0.39$ for β). For both Rotterdam and New York, we found support for the null that the means of the “deterministic” and the “stochastic” distributions were the same ($BF = 0.13$ for Rotterdam, $BF = 0.24$ for New York). We also found support for the null that the variance of the New York distribution was the same under the deterministic and the stochastic approach ($BF = 0.25$). However, for Rotterdam the evidence was inconclusive ($BF = 0.84$).

The evidence that subjects expected lower temperatures in New York City was very strong under the stochastic approach ($BF = 61534.3$). We also found very strong evidence that the variance of the Rotterdam beliefs distribution was lower than that of the New York distribution ($BF = 11673.8$).

Figure 5: Density and cumulative distribution of beliefs. Stochastic approach based on the means of the individual values



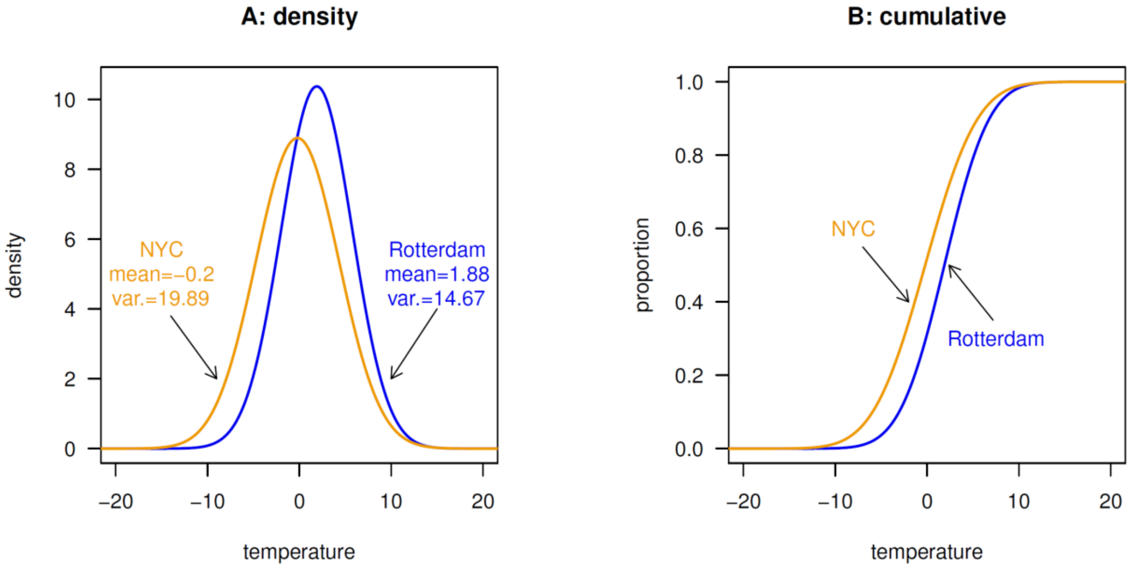
5.4. Beliefs and attitudes

We now turn to the results of the complete elicitation of Model (1) including the measurement of utility and ambiguity attitudes. Figure 6 shows the estimated beliefs distributions for Rotterdam and New York based on the means of the individually estimated parameters of the beta distribution. They are again similar to the distributions elicited under the deterministic approach and under the stochastic approach. A Bayesian Anova supported the null that the parameters of the beta distribution were the same under the three approaches (both $BF < 0.09$ for Rotterdam and both $BF < 0.16$ for New

York.). The parameters were also strongly correlated (all $\varrho > 0.73$ for Rotterdam and all $\varrho > 0.66$ for New York).

The data also supported equality of the means of the belief distributions under the three approaches for both Rotterdam ($BF = 0.05$) and New York ($BF = 0.14$). We found support for the null that the variances were the same for New York under the three approaches ($BF = 0.19$), but for Rotterdam the data were inconclusive about the equality of the variances ($BF = 1.59$). The comparison between the stochastic approach and the complete elicitation supported the null of equal variances ($BF = 0.12$), but the comparison with the deterministic approach was inconclusive ($BF = 0.82$) just like we saw in the comparison between the deterministic and the stochastic approach.

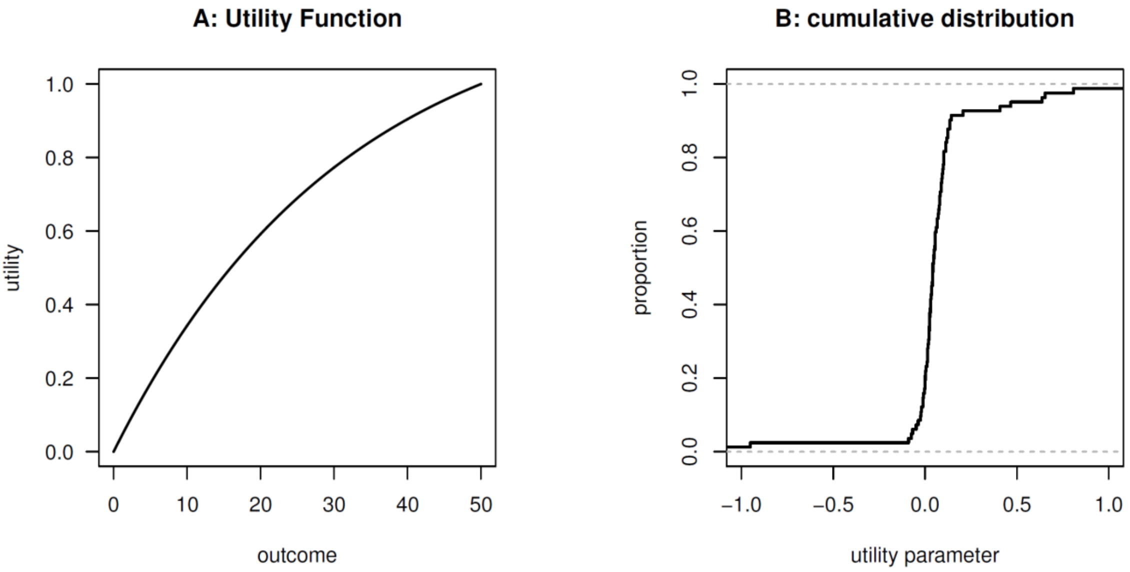
Figure 6: Density and cumulative distribution of beliefs. Complete approach based on the means of the individual values



5.4.1. Utility

Figure 7 shows the estimated utility function based on the mean of the individual estimates and the cumulative distribution function of the individual estimates. We found support for the null that utility was linear ($BF = 0.14$). Panel B shows that this held for a large proportion of our subjects.

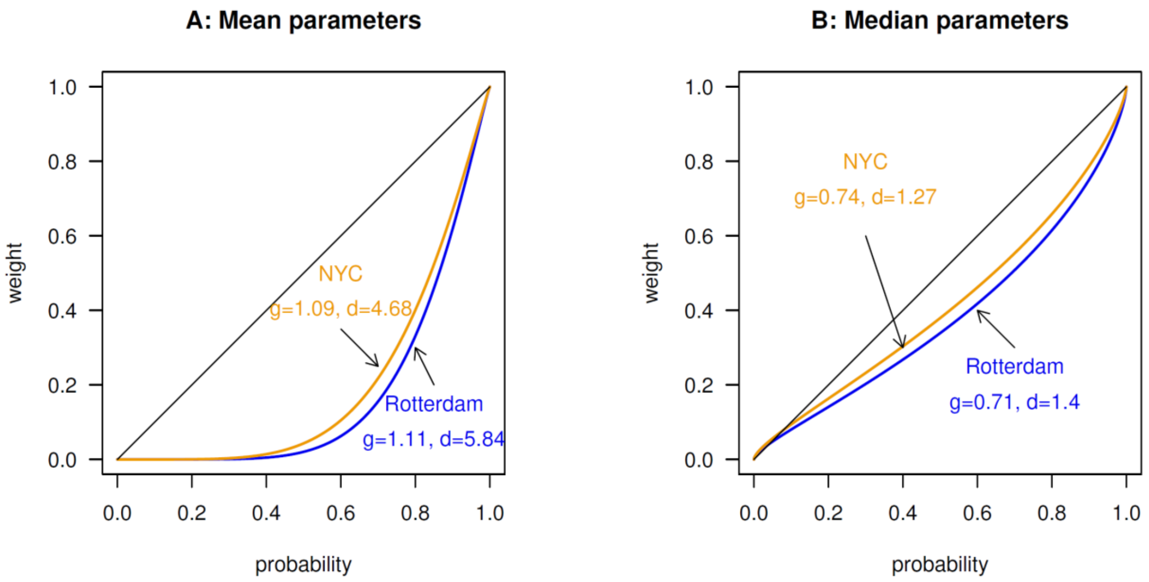
Figure 7: Estimated utility function using the mean of the individual estimates



5.4.2 Distortion functions

Figure 8 shows the estimated distortion functions for Rotterdam and New York based on the mean and median individual estimations. The means of the estimated parameters suggest a very pessimistic attitude in line with Gilboa and Schmeidler's (1989) maxmin expected utility model. The medians signal a less pessimistic attitude although the distortion functions lie for the most part below the diagonal reflecting underweighting of subjective probabilities. The observed median parameters were similar to those found by van de Kuilen and Wakker (2011). Compared to Abdellaoui et al. (2011) we found the same likelihood insensitivity, but more pessimism.

Figure 8: Distortion functions using the means of the estimated parameters



The distortion functions for Rotterdam and New York look similar and, indeed, the data supported the null that both likelihood insensitivity $(BF = 0.12)$ and pessimism $(BF = 0.14)$ were the same for Rotterdam and New York. Hence, we found no evidence that the distortion function depended on the source of uncertainty.

Figure 9: Cumulative distribution functions of the individual parameters for likelihood insensitivity and pessimism.

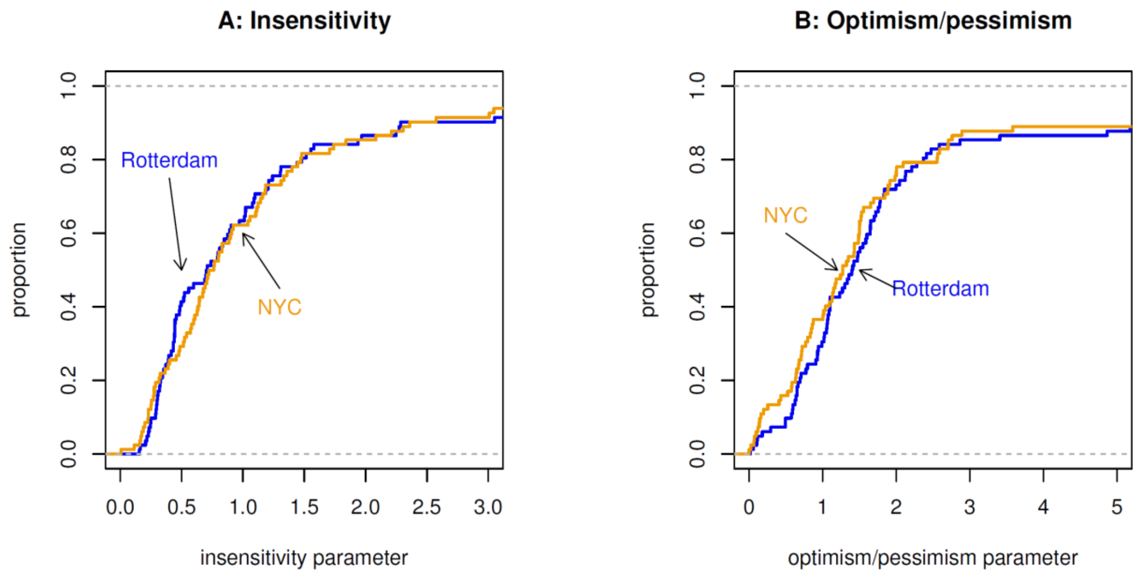


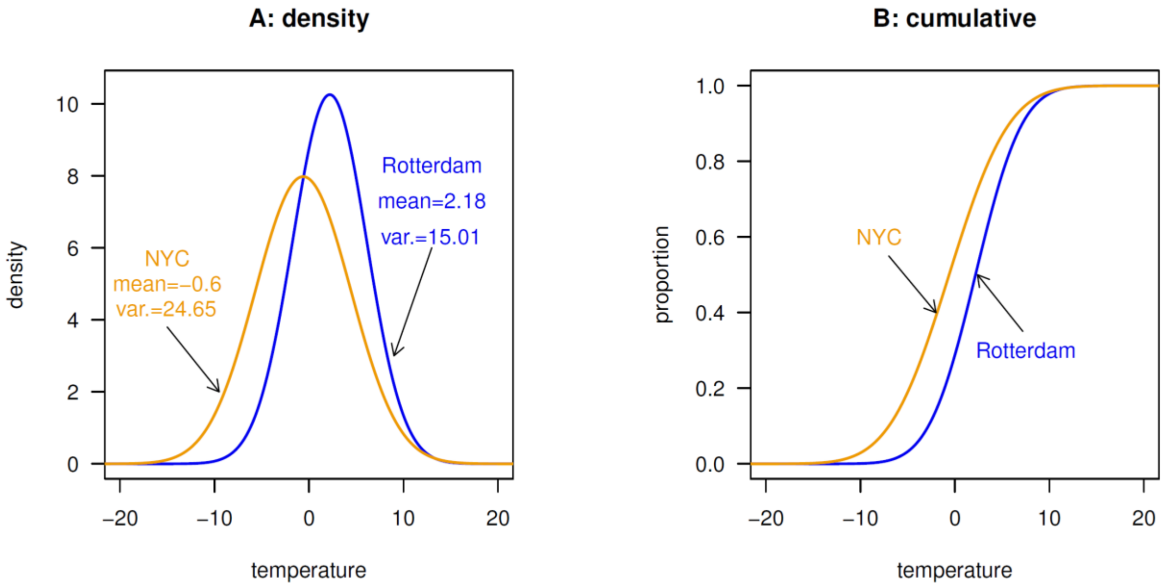
Figure 9 shows the cumulative distribution functions of the individual likelihood insensitivity and pessimism parameters for Rotterdam and New York. The figure shows that the functions for Rotterdam and New York and that there is a lot of individual heterogeneity. This large heterogeneity implied that the evidence on pessimism was inconclusive ($BF = 1.2$ for Rotterdam and $BF = 0.61$ for New York) and that we actually found support for the null that there was perfect likelihood sensitivity ($BF = 0.17$ for Rotterdam and $BF = 0.16$ for New York).

Given the above results we also explored whether our data were consistent with subjective expected utility. We re-estimated Model (1) with the restriction that the parameters d and g in Eq.(3), Prelec's (1998) specification of the distortion function, were both equal to 1. Model (1) with d and g unrestricted fitted significantly better for 45 of our 82 subjects. For the remaining subjects we could not reject expected utility.

5.5 Random coefficients estimation

To check for robustness we also estimated a random coefficients model. Figure 10 shows the estimated belief distributions under the random coefficients model for Rotterdam and New York. They are again similar to what we observed before. For Rotterdam, a Bayesian Anova gives very strong support for the null that all four methods that we considered (deterministic, stochastic, complete estimation, and random coefficients) give the same parameters of the beta distribution and the same mean ($BF < 0.02$). However, we also found strong support for the alternative that the variances differ ($BF = 19.9$). This is mainly due to the higher variance for the deterministic method. For New York, all tests supported the null that the parameters of the beta distribution, the mean, and the variance of the belief distributions elicited by the four methods were the same (all $BF < 0.32$). The correlations between the parameters of the beta distribution were very high (both $\varrho > 0.97$).

Figure 10: Density and cumulative distribution of beliefs.
Random coefficients model using the mean of the individual values



Turning now to attitudes, if we compare the estimates of the individual elicitations of Model (1) with those from the random coefficients model, we found support for the null that utility is the same ($BF = 0.12$). However, we also found strong support for the alternative that likelihood insensitivity differed (both $BF > 12.0$). We found more likelihood insensitivity in the random coefficients estimation. The comparison of pessimism was inconclusive for both Rotterdam and New York ($BF = 0.75$ for Rotterdam and $BF = 0.45$ for New York). In the random coefficients model we found some support for the hypothesis that the distortion functions differed across sources ($BF = 7.4$ for g and $BF = 5.5$ for d).

6. Discussion

Many decisions are made under uncertainty. To analyze these decisions requires measuring beliefs. This is complex due to the interactions between beliefs, risk attitudes, and ambiguity attitudes. Existing methods have solved this interaction problem by imposing simplifying assumptions. The popular proper scoring rules (Brier 1950, Good 1952) assume risk and ambiguity neutrality and only measure beliefs correctly if the decision maker maximizes expected value (Winkler and Murphy 1970). While refinements that allow for risk aversion and deviations from expected utility exist (Offerman et al. 2009, Kothiyal et al. 2010, Hossain and Okui 2013), systematic biases remain (Kadane and Winkler 1988, Armantier and Treich 2013). Moreover, psychologists have questioned whether people can sensibly respond to proper scoring rules (Erev et al. 1993).

The limitations of proper scoring rules have led to the use of non-incentivized methods (Manski 2004, Armantier and Treich 2013, Trautmann and Van de Kuilen 2014) and prediction markets (Wolfers and Zitzewitz 2004) in belief elicitation. These methods have limitations of their own. Hypothetical choice may lead to less careful and less truthful responses. Estimations from prediction markets assume risk neutrality and only give information about aggregate beliefs ignoring individual heterogeneity.[10] Prelec (2004) suggested an ingenious method, the Bayesian truth serum, which even permits measuring beliefs for unverifiable events. However, the incentive-compatibility of the Bayesian truth serum might be opaque for subjects and it assumes that decision makers

[10] Karni (2009) suggested measuring beliefs using matching probabilities (see also Spetzler and Stael Von Holstein 1975). This method avoids some violations of expected utility, but is invalid under ambiguity aversion (Budescu et al. 2011). There exist methods that are robust to ambiguity aversion (e.g. Abdellaoui et al. 2005, Diecidue et al. 2007, Baillon 2008, Abdellaoui et al. 2011), but these rely on simplifying assumptions or chained responses.

are Bayesians. Empirical evidence shows that people often deviate from Bayesianism (Grether 1980, El-Gamal & Grether 1995, Charness & Levin 2005, Poinas et al. 2012).

In this paper, we have introduced a simple method to measure beliefs that addresses the abovementioned problems. Our method is incentive-compatible, non-chained, and permits measuring the individual distribution of beliefs from just three measurements. Perhaps most important, our method gives correct measurements even when the decision maker has non-neutral risk and ambiguity attitudes and deviates from expected utility and Bayesianism. Experimental implementation showed that the measured beliefs were well-calibrated and sensitive to the different sources of uncertainty. We tested the robustness of the measurements by our simple method with those from increasingly sophisticated methods that took account of the stochastic nature of people's preferences, that measured both beliefs and ambiguity attitudes, and that allowed the estimates for each subject to benefit from the data of the other subjects (the random coefficients model). We found support for the null that our simple method gave the same estimates as these more sophisticated methods.

Our method is based on Model (1), Ghirardato and Marinacci's (2001) general model of biseparable preferences with probabilistic sophistication within sources of nature. Empirical support for the central condition underlying biseparable preferences was obtained by Abdellaoui et al. (2016). Probabilistic sophistication within sources of nature was tested by Abdellaoui et al. (2011), who could not reject it. We included one test of Model (1) and obtained support for it. More evidence is required to settle the debate on the acceptability of assuming Model (1), but based on what is available Model (1) seems to describe people's preferences rather well.

We fitted the individual belief distributions by the beta distribution. As the beta distribution is general we do not believe that this has introduced substantial distortions.

The beta distribution fitted significantly better than the other distributions that we tried. Fitting a continuous distribution to a small number of individual data points has the advantage that response errors are smoothened out and leads to belief forecasts that are at least as accurate as "raw" data points which are often subject to considerable error.

A final argument in favor of our method is that our method may be less prone to errors. We estimated two error terms, one for the exchangeability questions and one for the probability equivalence questions. We found strong evidence that the estimated errors in the exchangeability questions were lower than those in the probability equivalence questions. The median error was about three times as small in the exchangeability questions. As our method uses only exchangeability questions, this findings suggests that the error in our measurements will be substantially lower than the error in methods based on probability equivalence measurements.

7. Conclusion.

This paper introduces a simple method to measure individual beliefs. Our method is incentive compatible, non-chained, and it estimates the distribution of beliefs using three simple measurements. It is valid under many decision models and is robust to ambiguity aversion. Measured beliefs were well-calibrated, sensitive to the two sources of uncertainty and similar to those obtained using more sophisticated methods. We hope that by providing such a simple method to measure beliefs in decision under ambiguity will help decision analysts to accurately account for uncertainty in their models.

References

Abdellaoui, M., Baillon, A., Placido, L., & Wakker, P. P. (2011). The rich domain of uncertainty: Source functions and their experimental implementation. *American Economic Review, 101*(2), 695-723.

Abdellaoui, M., Bleichrodt, H., L'Haridon, O., & Van Dolder, D. (2016). Measuring loss aversion under ambiguity: A method to make prospect theory completely observable. *Journal of Risk and Uncertainty, 52*(1), 1-20.

Armantier, O., & Treich, N. (2013). Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review, 62*, 17-40.

Baillon, A. (2008). Eliciting subjective probabilities through exchangeable events: An advantage and a limitation. *Decision Analysis, 5*(2), 76-87.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1-3.

Budescu, D., Abbas, A., & Wu, L. (2011). Does probability weighting matter in probability elicitation? *Journal of Mathematical Psychology, 55*(4), 320-327.

Charness, G., & Levin, D. (2005). When optimal choices feel wrong: A laboratory study of Bayesian updating, complexity, and affect. *American Economic Review, 88*(4), 933-946.

Chew, S. H., & Sagi, J. S. (2006). Event exchangeability: Probabilistic sophistication without continuity or monotonicity. *Econometrica, 74*(3), 771-786.

de Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales De L'Institut Henri Poincaré, 7*, 1-68.

Diecidue, E., Wakker, P. P., & Zeelenberg, M. (2007). Eliciting decision weights by adapting de Finetti's betting-odds method to prospect theory. *Journal of Risk and Uncertainty, 34*(3), 179-199.

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes*. Oxford: Oxford University Press.

El-Gamal, M. A., & Grether, D. M. (1995). Are people Bayesian? Uncovering behavioral strategies. *Journal of the American Statistical Association, 90*(432), 1137-1145.

Ellsberg, D. (1961). Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics, 75*, 643-669.

Erev, I., Bornstein, G., & Wallsten, T. S. (1993). The negative effect of probability assessments on decision quality. *Organizational Behavior and Human Decision Processes, 55*(1), 78-94.

Farquhar, P. (1984). Utility assessment methods. *Management Science, 30*, 1283-1300.

Gajdos, T., Hayashi, T., Tallon, J. M., & Vergnaud, J. C. (2008). Attitude toward imprecise information. *Journal of Economic Theory, 140*(1), 27-65.

Ghirardato, P., Maccheroni, F., & Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory, 118*(2), 133-173.

Ghirardato, P., & Marinacci, M. (2001). Risk, ambiguity, and the separation of utility and beliefs. *Mathematics of Operations Research, 26*(4), 864-890.

Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics, 18*, 141-153.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society.Series B (Methodological), ,* 107-114.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics, 95*(3), 537-557.

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty, 4*, 4-28.

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

Hossain, T., & Okui, R. (2013). The binarized scoring rule. *Review of Economic Studies, 80*, 984-1001.

Kadane, J. B., & Winkler, R. L. (1988). Separating probability elicitation from utilities. *Journal of the American Statistical Association, 83*(402), 357-363.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica, 77*(2), 603-606.

Klibanoff, P., Marinacci, M., & Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica, 73*(6), 1849-1892.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage Publications.

Manski, C. F. (2004). Measuring expectations. *Econometrica, 72*(5), 1329-1376.

Offerman, T., Sonnemans, J., Van De Kuilen, G., & Wakker, P. P. (2009). A truth serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies, 76*(4), 1461-1489.

Poinas, F., Rosaz, J., & Roussillon, B. (2012). Updating beliefs with imperfect signals: Experimental evidence. *Journal of Risk and Uncertainty, 44*, 219-241.

Prelec, D. (1998). The probability weighting function. *Econometrica, 66*, 497-528.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science (New York, N.Y.), 306*(5695), 462-466.

Ramsey, F. P. (1931). *The foundations of mathematics and other logical essays*. New York: Harcourt, Brace and Co.

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science, 12*, 185-190.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica, 57*, 571-587.

Spetzler, C. S., & Stael Von Holstein, Carl-Axel S. (1975). Probability encoding in decision analysis. *Management Science, 22*(3), 340-358.

Trautmann, S. T., & Van de Kuilen, G. (2016). Ambiguity attitudes. In G. Keren, & G. Wu (Eds.), *Blackwell handbook of judgment and decision making* (pp. 89-116) Wiley Blackwell.

Trautmann, S., & Van de Kuilen, G. (2014). Belief elicitation: A horse race among truth serums. *Economic Journal, 125*, 2116-2135.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207-232.

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

van de Kuilen, G., & Wakker, P. P. (2011). The midweight method to measure attitudes toward risk and ambiguity. *Management Science, 57*(3), 582-598.

Winkler, R. L., & Murphy, A. H. (1970). Nonlinear utility and the probability score. *Journal of Applied Meteorology, 9*(1), 143-148.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives, 18*(2), 107-126.