Eliciting Belief Distributions: A Comparative Study*

Helen Grapow

RISL$\alpha\beta$, Department of Economics, Ghent University

15 December 2026

Abstract

This study compares three mechanisms that elicit individual belief distributions and are easy for participants to understand within strategic environments. Two methods are incentive-compatible under most models of decision-making and rely on non-chained binary choices. One mechanism elicits only beliefs while remaining robust to preferences, while the other provides an integrated measurement of beliefs and uncertainty attitudes, albeit under additional assumptions. The third method relies on survey questions without incentivization. Results show that all three approaches reliably recover subjective probabilities, with distinct strengths across performance metrics. The study highlights the role of belief confidence, together with mean belief, in understanding and predicting choices under uncertainty. Moreover, it underscores the importance of accounting for ambiguity attitudes alongside beliefs in models of decision-making.

1 Introduction

Beliefs, defined as subjective probabilities about the likelihood of different outcomes, are key drivers of major decisions under uncertainty. According to the literature on decision-making under risk and ambiguity (Savage, 1954; Ellsberg, 1961; Ghirardato and Marinacci, 2001; Klibanoff et al., 2005), individuals combine beliefs with preferences to reach a decision. Eliciting subjective probabilities jointly with preferences, or estimating one of the two while controlling for the other, has thus become central in the research agenda for economic behavior. In this framework, methods that elicit belief distributions have been developed to capture both mean beliefs and their dispersion, reflecting the role of belief confidence in driving choices (Gilboa and Schmeidler,

*I acknowledge financial support from the FWO Research Foundation-Flanders (project 11P3P24N). The project received approval from the ethics committee at the Faculty of Economics and Business, Ghent University (UG-EB 2025-Y). I am indebted to Levent Gumus for the help with the experiments and to Ferdinand M. Vieider for helpful comments. I am further grateful to Ranoua Bouchouicha and Yassine Kaouane for valuable discussions. All errors remain my own.

1989; Manski, 2004; Abdellaoui et al., 2024b). Nevertheless, eliciting belief distributions remains far from a settled methodological issue.

In this article, I pit three promising methods for eliciting belief distributions against each other, discussing their underlying assumptions and their practical implementation. By measuring full individual-level belief distributions rather than point estimates, I focus on methodologies that account for subjects' uncertainty around their mean subjective probability. The mechanisms are relatively simple for experimental subjects to understand, responding to prior calls for comparative studies of non-complex elicitation methods.

The methods differ in their methodological approach and whether they are incentive-compatible. Two measurements are incentive-compatible under most major decision-making models. One mechanism, referred to as the *Money Method*, elicits belief distributions through choices between a sure amount and an ambiguous lottery. The other, named the *Bet-Based Method* (Abdellaoui et al., 2024b), elicits beliefs through choices between exogenously determined events to win a fixed prize. Both methods rely on non-chained revealed choices, which aim to avoid error accumulation and strategic responding. Additionally, the use of binary choices simplifies the task for participants and minimizes noise, context effects, and the use of heuristics compared to choice list methods (Vieider, 2018; Bouchouicha et al., 2024; Chapman et al., 2025). The cost of this simplicity lies in the rather complex econometrics, as belief recovery relies on assuming a parametric belief distribution, estimated through hierarchical Bayesian techniques. They differ in that, while the Bet-Based Method is robust to utility curvature and probability distortions, the Money Method enables the disentangled measurement of both belief distributions and preferences - albeit at the cost of assuming an underlying model of decision-making and functional forms. Finally, the *Introspective Method* (Manski, 2004; Wiswall and Zafar, 2015; Delavande and Zafar, 2019) elicits belief distributions by asking participants to self-report the percent chance that some uncertain events occur. While relatively quick to implement, this method relies on additional assumptions, including participants' understanding of probability and truthful reporting.

The mechanisms are compared using a between-subject experimental design. To this end, I first implement a Beauty Contest game (Nagel, 1995). After playing the game, subjective probabilities about the likelihood of the target number, defined as two-thirds of the average guess, are elicited using one incentivized method, alongside the Introspective Method. In a second experiment, participants play an Ultimatum game (Guth et al., 1982). The first mover chooses how to split a monetary amount between herself and the second mover from a menu of three allocations. After the game, participants' subjective probabilities that the second mover would accept each allocation are elicited, following the same structure as in Experiment 1. Both settings offer a simple yet informative design for studying how subjective probabilities about others' behavior influence strategic decision-making in a controlled environment. However, while in the Beauty Contest game players' actions are primarily driven by beliefs about others' guesses, the Ultimatum game allows for the study of multiple drivers of strategic behavior, making it a natural extension of the first experiment. To compare the performance of the methods, I rely on five evaluation metrics. Three non-parametric measures - response time, consistency in responses through the experiment, and monotonicity in choices - pit the practical deployment of the methods. Two parametric measures - belief calibration and predictive accuracy - compare the internal validity of the recovered belief measures in explaining participants' choices in the games.

The results indicate that all three methodologies perform well in eliciting reliable beliefs in both strategic environments. The non-parametric analysis reveals no evidence of inattentive responses due to task complexity or survey fatigue across the three elicitation procedures, with the Introspective Method being substantially quicker than the incentivized alternatives. Nevertheless, caution is warranted regarding the assumption that participants understand probabilities when employing this method. Parametric results show that, in the Beauty Contest game, both the mean and confidence of individual beliefs elicited by the three methods significantly predict participants' choices. In the Ultimatum game, belief confidence plays a prominent role in predicting whether decision makers deviate from their subjective best responses computed using mean beliefs. Furthermore, the performance of the Money Method hinges on the assumed model of decision-making, particularly those incorporating Probability Weighting, fitting the data better than Subjective Expected Utility. In conclusion, the findings suggest that if the goal is to recover beliefs with minimal functional assumptions, the Bet-Based Method appears to be the most efficient alternative. By contrast, if objective includes assessing subjects' attitudes toward ambiguity, the Money Method under ambiguity models is preferable. Finally, if response time is an experimental constraint, the Introspective Method constitutes a suitable option.

Related Literature and Contribution

This study contributes to the literature on beliefs by providing the first comparison of methods for eliciting belief distributions that are simple for participants to understand. Prior research has extensively reviewed belief elicitation techniques, with a focus on scoring rules (Schotter and Trevino, 2014; Schlag et al., 2015; Trautmann and Kuilen, 2015). However, scoring rules, even when corrected for deviation from expected value maximization (Offerman et al., 2009; Hossain and Okui, 2013), may be too complex to understand or sensitive to measurement noise (Charness et al., 2021; Danz et al., 2022). In addition, they elicit point estimates, rather than belief distributions. More recent work by Richard et al. (2025) compares methods that elicit distributions based on exchangeability (Baillon, 2008; Abdellaoui et al., 2011), matching probabilities (Baillon et al., 2018b), and non-incentivized methodologies in terms of stability. The methodologies selected in the current study differ from these works as they are not implemented through the bisection method, which may raise incentive compatibility concerns. Moreover, studies have described several implementations of probability matching, such as those used in Holt and Smith (2009); Karni (2009); Hao and Houser (2012), as difficult for experimental subjects to understand (Schotter and Trevino, 2014; Charness et al., 2021). I further take a different approach by evaluating how the elicited beliefs perform in predicting behavior, rather than how consistently they reveal latent beliefs.

Additionally, this work contributes to the ongoing debate in the literature comparing incentivized methods with introspective surveys. While some studies suggest that surveys perform comparably to incentivized approaches (Trautmann and Kuilen, 2015; Hollard et al., 2015; Burfurd and Wilkening, 2022; Kemel et al., 2025), others find that properly designed incentives improve elicited beliefs (Palfrey and Wang, 2009; Schotter and Trevino, 2014; Gachter and Renner, 2010; Wang, 2011; Richard et al., 2025). However, this discussion remains silent on techniques that elicit belief distributions. By providing the first comparative study of this kind, this work shows that incentivized methods perform comparably to the Introspective Method, particularly in Experiment 2. Nevertheless, due to the Introspective Method's limitations in capturing beliefs across the full subject pool, further research is needed to compare more complex elicitation methods with self-reported approaches.

Finally, this study aligns with a growing body of research that integrates subjective beliefs with ambiguity attitudes in decision-making models. While this framework has gained theoretical traction, empirical support for models other than SEU incorporating uncertainty attitudes alongside subjective probabilities remains limited. Consistent with recent findings (Trautmann and Kuilen, 2015; Delavande et al., 2024; Kemel and Mun, 2024), this study finds that jointly measuring beliefs and ambiguity attitudes enhances the predictive power of models of decision-making.

2 Experimental Design

I compare three methods for eliciting individual-level subjective belief distributions, which differ in their methodological approaches and in whether they are incentive-compatible, using two experiments. In Experiment 1, participants played a Beauty Contest game (Nagel, 1995), while in Experiment 2, they played an Ultimatum game (Guth et al., 1982).

The experiments were implemented in Qualtrics, and participants were recruited from a U.S.-based subject pool via Prolific. Respondents received a participation fee of approximately \$3 for Experiment 1 and \$6 for Experiment 2, and had the opportunity to earn a bonus payment based on one randomly selected choice. On average, participants earned a bonus of \$8.4 in

Treatment I, where beliefs were elicited via the Bet-Based method, and \$11.9 in Treatment II, where beliefs were elicited via the Money Method. Participants were randomly assigned to one of the two treatments and were required to pass a comprehension test after receiving the task instructions. Upon passing the test, they completed three tasks: (1) the game; (2) an incentive-compatible belief elicitation procedure specific to their assigned treatment; (3) the Introspective Method. Participants played the game first, and the order of the remaining tasks was randomized. Finally, participants filled out a questionnaire that included demographic questions. In Experiment 2, I further elicited fairness and risk preferences, following the Global Preferences Survey methodology (Falk et al., 2018), detailed in Appendix 6.

2.1 The Beauty Contest Game

In Experiment 1, the belief elicitation procedures were implemented within the standard Beauty Contest (BC) game (Nagel, 1995). Subjects were asked to state a number within the interval [0, 100]. The participant whose guess was closest to the target number, defined as the mean of all chosen numbers multiplied by $p = \frac{2}{3}$, won a prize of \$20, split evenly in case of a tie.

This experimental framework was selected because it provides a well-established and simple design for studying subjective probabilities about others' behavior, which constitutes an integral component of action choice. The Beauty Contest game is also straightforward to explain, yet it reveals insight into individual strategic reasoning in a controlled environment.

2.2 The Ultimatum Game

At the outset of Experiment 2, participants were asked to play a two-person Ultimatum game (henceforth UG). In this dynamic game, there were two roles that subjects took on: the first mover, or proposer, and the second mover, or responder. The proposer was presented with \$20 to split between herself and the responder, with three possible allocations to choose from, as shown in Table 1. The responder could either accept the proposed allocation, in which case it determined their final payoffs, or reject it, in which case both received nothing. I employed the strategy method, where the second mover stated whether she would accept or reject each of the three offers. Each subject played both roles sequentially. The payoff-relevant role was randomly determined at the end.

Role	**Allocation 1**	**Allocation 2**	**Allocation 3**
Proposer	\$19	\$16	\$10
Responder	\$1	\$4	\$10

Table 1: Ultimatum Game

This game provides a simple tool to study the different drivers of strategic decision-making in a controlled environment. The first mover does not know the probability that the responder will accept each allocation and therefore relies on her subjective beliefs about the likelihood of acceptance for the three possible offers to determine her choice. Compared to the Beauty Contest game, where players' actions are primarily driven by the decision maker's belief about others' guesses, choices in the Ultimatum game hinge on the beliefs over multiple events and on individuals' attitudes towards risk. Other-regarding preferences may also play a role (Fehr and Schmidt, 1999). This makes the UG a natural follow-up to the BC, as it enables a richer analysis of the determinants of decisions under uncertainty. The central aim of Experiment 2 is to compare how the different elicitation methods account for these aspects.

2.3 The Elicitation Methods

2.3.1 The Bet-Based Method

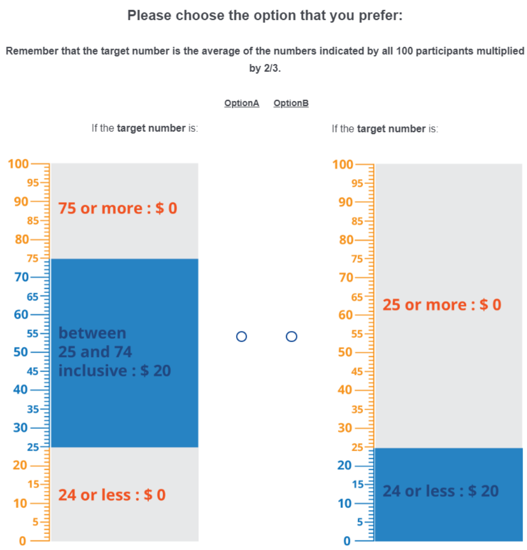


Figure 1: A choice task for the Bet-Based Method

In Treatment I, subjective belief distributions were elicited using the mechanism proposed by Abdellaoui et al. (2024b), referred to as the Bet-Based (BB) method in this study. Emulating methods based on exchangeability (Baillon, 2008; Abdellaoui et al., 2011; 2024a), the Bet-Based method builds on a binary choice procedure that enables collecting information on which of two events is considered more likely. Unlike previous methods, this methodology relies entirely on independent binary choices, circumventing the issues of incentive compatibility and error propagation. When events are generated by the same source of uncertainty, the method remains valid under a range of non-expected utility models, requiring only minor assumptions and approximations. Such models include the Biseparable Preference Model[1] (Ghirardato and Marinacci, 2001) and the Smooth Model (Klibanoff et al., 2005).

It is assumed that there exists a subjective probability measure $P(.)$ on the state space, i.e [0, 100], that holds without committing to expected utility maximization. All bets are placed to obtain a fixed prize of \$20 or else \$0. A bet on an uncertain event E consequently corresponds to a gamble that pays \$20 if the event occurs and nothing otherwise, denoted by 20_E0. The use of such bets serves to keep the task simple for participants, who could focus on the cutoff points of the two events, as all other features are held constant. Figure 1 displays a screenshot of one of the tasks[2]. A total of 39 choice tasks for Experiment 1 and 78 for Experiment 2 were selected through simulations to recover a broad range of belief parameters and ensure comparability across methods. Additionally, four tasks were repeated to allow for better quantification of errors. Compared to recent methods, the BB relies on fewer stimuli. For instance, the approach by Abdellaoui et al. (2021), which employs choice lists with varying event cutoffs, and the matching-probabilities method by Baillon et al. (2018a) required participants to make 468 and 720 active choices, respectively. This behaviorally simple and quick elicitation comes at the cost of greater econometric complexity required to fit parametric belief distributions.

2.3.2 The Money Method

In Treatment II, subjects' belief distributions were elicited using the Money Method (MM). This mechanism enables the disentangled measurement of both beliefs and uncertainty attitudes, under the assumption of specific functional forms for the relevant quantities. The method is valid under the Biseparable Model (Ghirardato and Marinacci, 2001) and is therefore incentive compatible under most major models of decision-making, except under the Smooth model of Klibanoff et al. (2005).

Upon assuming the existence of a subjective probability distribution $P(.)$ over events generated by a given source on the state space [0,100], each task in this method consists of a binary choice between a sure amount and a bet on an uncertain event E. The bet pays $x \in \{\$5, \$10, \$15, \$20\}$ if the event occurs and $y \in \{\$0, \$5, \$10\}$ otherwise, denoted by $x_E y$. To guarantee the validity of the method under a range of non-expected utility models, the MM necessitates the inclusion of bets on complementary events. Specifically, for non-internal events, i.e. $E \notin (0, 100)$, there was a corresponding choice situation in which the bet was defined on E^c. Further, variation across the outcome dimension was introduced for eliciting preference parame-

[1]The method is valid under most major models of decision-making making such as Subjective Expected Utility (SEU; Savage (1954)); Prospect Theory (PT; Tversky and Kahneman (1992)), Choquet EU (CEU; Schmeidler (1989); Wakker (2010)), and Rank-Dependent Utility (RDU; Quiggin (1982); Wakker (1994)); Maxmin EU (MEU; Gilboa and Schmeidler (1989)), α-Maxmin (Ghirardato et al., 2004), and probability intervals models (Manski, 2004).

[2]The sample tasks for Experiment 2 and the list of binary choices for all methods are in the Appendix 6.

ters. In total, 161 unique binary choice tasks were used in Experiment 1 and 195 in Experiment 2, along with four repeated questions. Figure 2 shows a sample task from Experiment 1.

The Money Method emulates the methodology of Kemel and Mun (2024), who developed an econometric and experimental framework to jointly estimate beliefs and attitudes from binary choices between a sure outcome and a binary prospect under a general non-EU model. However, while their measurement relies on certainty equivalents derived through bisection, my approach introduces a novel feature, as the estimates are recovered solely from non-chained choices. Com-

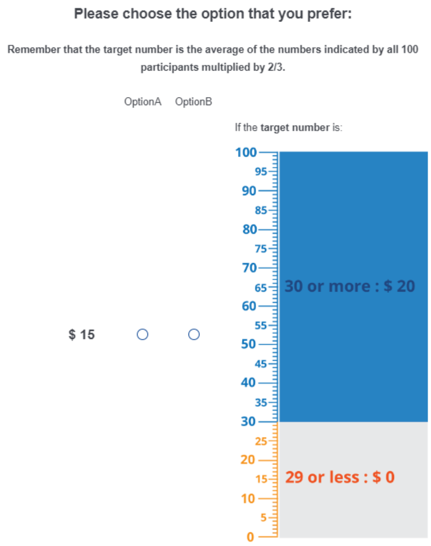


Figure 2: A choice task for the Money Method

pared to the Bet-Based method, this methodology recovers additional drivers of decisions under uncertainty, such as ambiguity attitudes within the same source of uncertainty. By jointly estimating beliefs and attitudes within a unified experimental task, the MM further overcomes the risk of order effects (Schlag et al., 2015). The cost of eliciting such a rich set of measures lies in the high econometric complexity of the approach and the relatively large number of choices required from participants, which may raise concerns about survey fatigue and practical implementation.

2.3.3 The Introspective Method

The Introspective Method elicits beliefs through direct and unincentivized survey questions. I employ the percent chance formulation, in line with recommendations from the literature on eliciting subjective probabilities through surveys (see Manski (2004); Hurd (2009); Delavande et al. (2011); Delavande (2014)). The Method elicits subjective probabilities of an uncertain event by asking the percent chance that the event will occur. For example, values like 2 or 5 may indicate "almost no chance", 45 or 55 "about even", and 95 or 98 "almost certain."

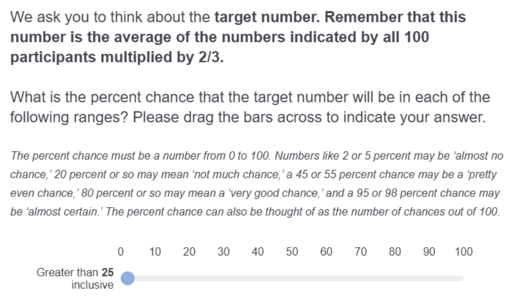


Figure 3: A choice task for the Introspective Method

Following Delavande and Zafar (2019) and Wiswall and Zafar (2015), respondents reported a number between 0 and 100 to indicate the percent chance of two events. Specifically, they stated the percent chance that the target number in Experiment 1 and the proportion of acceptors of a specific allocation in Experiment 2 would fall within 2 ranges, namely $E \in \{[25, 100], [75, 100]\}$. Participants indicated their answer by dragging a bar ranging from 0 to 100, designed to help them visualize their response. Figure 3 shows a screenshot of a task.

Compared to the Bet-Based Method and Money Method, this approach requires significantly fewer questions to estimate beliefs, making it relatively quick and simple for both the experimenter and the survey respondents. The method's practicality has contributed to its wide use in empirical analysis. This efficiency comes at the cost of relying on behavioral assumptions, such as that individuals understand probabilities and report their beliefs truthfully. Including this mechanism allows assessing how these methodological simplifications affect elicited beliefs compared with measures that do not rely on such assumptions.

3 Econometrics

This section details the econometric recovery of belief distributions. Table 2 summarizes the main assumptions, recovered measures, and theoretical robustness of each method. The Table highlights the trade-off between the theoretical validity, the amount of information acquired, and the ease of practical deployment.

Method	Main Assumptions	Number of tasks	Recovered Measures	Incentive Compatible under
Bet-Based Method	Beta distribution of beliefs; Probabilistic sophistication within source	39	Belief distributions	Ghirardato and Marinacci (2001); Klibanoff et al. (2005)
Money Method	Beta distribution of beliefs; Probabilistic sophistication within source; Decision-making model and functional forms	161	Belief distributions; Preferences	Ghirardato and Marinacci (2001)
Introspective Method	Log-normal distribution of beliefs; Truthful reporting; Understanding of probability concepts	2	Belief distributions	-

Note: The model of Ghirardato and Marinacci (2001) encompasses SEU, PT, CEU, RDU, Maxmin EU, α-Maxmin, and Manski (2004).

Table 2: Comparison of Elicitation Methods

3.1 Recovery of Parametric forms for the incentivized methods

For the incentivized methods, beliefs are assumed to follow a Beta distribution due to its flexibility in capturing various belief patterns. Let T denote the uncertain event of interest. The probabilistic belief of subject i about T is expressed as follows:

$$P_i(T \mid \alpha_i, \beta_i) = (\alpha_i, \beta_i), \tag{1}$$

where α_i and β_i denote the non-negative shape parameters. The mean m_i and the concentration k_i of the subjective beliefs distribution are obtained by parameterizing these parameters as $m_i = \frac{\alpha_i}{\alpha_i+\beta_i}$ and $k_i = \alpha_i + \beta_i$, where higher values of k denote lower confidence around the mean of the Beta distribution[3].

Let $P_i(E_j^B)$ denote participant i's subjective probability of the event entailed by "Option B" in the j^{th} binary choice within the set of elicitation tasks. This belief reflects the probability mass attributed to the event according to i's underlying distribution and is defined as:

$$P_i(E_j^B) = \mathcal{F}(u_B \mid \alpha_i, \beta_i) - \mathcal{F}(\ell_B \mid \alpha_i, \beta_i), \tag{2}$$

where $\mathcal{F}$ denotes the cumulative distribution function of the Beta distribution, and u and ℓ represent the upper and lower bounds of the event E_j^B. Thus, through revealed choices between binary options, the methods elicit the underlying distribution governing subjects' choices. The following subsections detail the procedures used by each elicitation method.

The estimation of the individual-level parameters relies on relatively few data points, which results in an increased uncertainty surrounding the estimates. To address the resulting uncertainty, a measurement error model is employed through a Bayesian hierarchical approach, wherein individual-level parameters are treated as samples from an overarching distribution that characterizes the entire experiment. The method is geared at maximizing the predictive performance of the estimates (see Vieider 2024). The model is estimated in Stan using Hamiltonian Monte Carlo simulations. Stan is launched from an algorithm of R using CmdStanR.

3.1.1 The Bet-Based Method

Let $P_i(E_j^A)$ and $P_i(E_j^B)$ denote subject i's belief about the likelihood of the event in Option A and Option B, respectively, in the jth binary choice within the set of elicitation tasks. $P_i(E_j^B)$

[3] For computation convenience, the logit-normal approximation to the Beta distribution is used to recover the parameters of the Beta distribution, as in (Abdellaoui et al., 2024b). This approach is more efficient and avoids issues of numerical stability that can occur when estimating cumulative distribution functions of the Beta distribution. The beliefs are further elicited assuming they follow a Kumaraswamy distribution, which also constitutes an approximation of the Beta distribution, as a robustness check. Results available upon request.

is defined as in equation (2) and $P_i(E_j^A)$ is expressed as:

$$P_i(E_j^A) = \mathcal{F}(u_A \,|\, \alpha_i, \beta_i) - \mathcal{F}(\ell_A \,|\, \alpha_i, \beta_i), \tag{3}$$

with notation consistent with equation (2). Given the setup of the elicitation mechanism, subject i chooses Option A over Option B whenever $P_i(E_j^A) \geq P_i(E_j^B)$, or else displays the opposite choice pattern. The probability of subject i choosing Option A over Option B in choice j is estimated using the following logistic regression:

$$Pr_i^j[A \succ B] = \text{Bern}\left[\text{logit}^{-1}\left(\frac{P_i(E_j^A) - P_i(E_j^B)}{\sqrt{2}\tau}\right)\right] \tag{4}$$

Here, *Bern* denotes the Bernoulli distribution, and $logit^{-1}$ represents the inverse-logit link function. The scaling by $\sqrt{2}\tau$ standardizes τ, the standard deviation of the residual (Train, 2009).

3.1.2 The Money Method

In the Money Method, identification of belief distributions relies on subjects' choices between a sure amount and a lottery. This requires additional assumptions about the representation of subjects' preferences over the two options. Let c_j denote the sure amount in Option A, and $L_j := (x_{E_j^B} y)$ the bet in Option B for the j^{th} choice within the elicitation task. Let $V_i^{L_j}$ and $V_i^{c_j}$ be subject i's valuation of the bet and the sure amount, respectively. The subjective probability $P_i(E_j^B)$ enters the valuation of the bet and is defined in equation (2). Subject i will choose Option A over Option B whenever $V_i^{c_j} \geq V_i^{L_j}$. The probability of subject i choosing Option A over Option B in choice j is then estimated using the following logistic regression:

$$Pr_i^j[A \succ B] = \text{Bern}\left[\text{logit}^{-1}\left(\frac{V_i^{c_j} - V_i^{L_j}}{\sqrt{2}\tau}\right)\right] \tag{5}$$

Depending on the modeling choice, the MM requires different assumptions on the functional forms of the utility and probability distortions. To this end, I evaluate subjects' preferences according to three major models of decision-making under uncertainty: *Subjective Expected Utility* (SEU; Savage (1954)); *Probability Weighting*, which includes PT (Tversky and Kahneman, 1992), CEU (Schmeidler, 1989; Wakker, 2010), and RDU (Quiggin, 1982; Wakker, 1994); and the *Multiple Priors*, encompassing MEU (Gilboa and Schmeidler, 1989) and Alpha-Maxmin (Ghirardato et al., 2004). The functional forms assumed in each model are detailed in Appendix 6.

3.2 The Introspective Method

Respondents' subjective probabilities of the event of interest T are elicited through an Introspective Method established in previous literature (e.g. Delavande and Zafar (2019); Wiswall and Zafar (2015)). This elicitation requires asking the percent chance that T will be greater than two thresholds. For each individual, the following data points are recovered: i) belief that T would exceed 25, ii) belief that T would exceed 75. Given the limited number of data points, beliefs could not be estimated under the Beta distribution and were instead assumed to follow a log-normal distribution, which is less flexible in capturing belief patterns. Individual i beliefs about T are expressed as:

$$ln(T_i) \sim N(\mu_i, \sigma_i^2), \tag{6}$$

where μ_i and σ_i represent the individual-level mean and standard deviation of the belief distribution, respectively. Through simulation, I generate a sequence of draws from the assumed distribution, compute the simulated counterparts of the two belief statistics, and then minimize the quadratic distance between simulated and observed data to identify the best-fitting parameters of the distribution[4].

Finally, to obtain the 95% highest-density intervals around the belief parameters, I applied a bootstrap procedure to the fitted estimates and took the 2.5^{th} and 97.5^{th} percentiles of the bootstrap distributions. The estimations are done in Matlab.

4 Results

Experiment 1 was conducted in October 2024. In Treatment I, beliefs were elicited via the Bet-Based Method (N=98), and in Treatment II, beliefs were elicited via the Money Method (N=92). Experiment 2 took place in May 2025, and 100 subjects were assigned to Treatment I and 98 to Treatment II. In this section, I firstly describe players' behavior in the Beauty Contest and Ultimatum games across treatments, which serves as a benchmark for analyzing the elicited subjective beliefs. I then conduct a comparative analysis of the performance of the elicitation methods through five main metrics. Three metrics are non-parametric -response time, consistency of answers throughout the experiment, and monotonicity in choices- while the remaining two are parametric, comparing the calibration and predictive accuracy of individual-level belief estimates relative to game behaviors. Table 3 provides a summary of key results.

[4] We employ two questions as in Delavande and Zafar (2019). This procedure is a simplified version of the estimation detailed in Appendix E of the work by Wiswall and Zafar (2015), who elicit individual-level belief distributions about future earnings, and additionally allow the parameters of the distribution to grow with age, which is beyond the scope of this study.

Performance measures	Bet-Based Method	Money Method	Introspective Method
Non-Parametric			
Response Time	✓	✗	✓
Monotonicity	✓	✓	✗
Consistency	✓	✓	–
Parametric			
Calibration	✓	✓	✗/✓
Explanatory Power	✓	✓	✓

Note: Ticks denote relatively high performance; Crosses denote relatively low performance. Split indicators (e.g., ✓/✗) refer to differences between BC and UG, respectively. Dashes indicate not applicable. The MM is under ambiguity models.

Table 3: Performance of elicitation methods

4.1 Descriptive Statistics

4.1.1 Choices in the Game

The Beauty Contest game

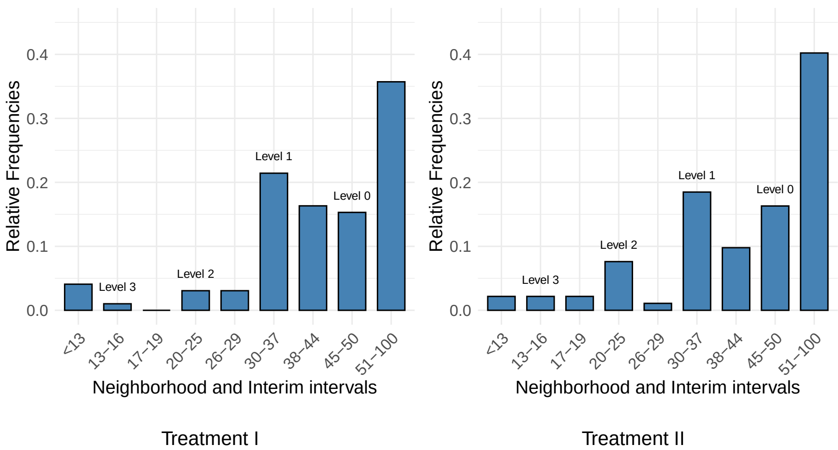


Figure 4: Levels of reasoning in Beauty Contest game.

Note: In Treatment I, beliefs are elicited via the BB Method; in Treatment II, via the MM.

The target number in the BC, defined as two-thirds of the average number chosen by participants, equals 31 in both treatments, consistent with previous findings (Mauersberger and Nagel, 2018). Figure 4 shows the distribution of guesses for Treatment I and Treatment II, grouped by level of reasoning and corresponding interim ranges, following the standard *level-k* model (Nagel, 1995). Relative to the guesses in Treatment I, the distribution of reasoning levels in Treatment II appears slightly more polarized, with a higher proportion of subjects showing no strategic reasoning (guessing values above 50) or Level-2 or higher (values below 25).

Insights into subjects' game behavior are also obtained by examining their choices in the BB elicitation tasks, reported in Table 9 of the Appendix. For instance, 53% of participants bet on the target number being below 50 rather than 50 or more, 35.7% on 30-40 rather than 40-50, and 18.4% on below 25 rather than 25-50, indicating that subjects' beliefs concentrate around values consistent with low reasoning levels. By contrast, players' guesses in the game cannot be revealed non-parametrically from choices in the MM and Introspective Method, making the BB mechanism more informative in this respect.

The Ultimatum Game

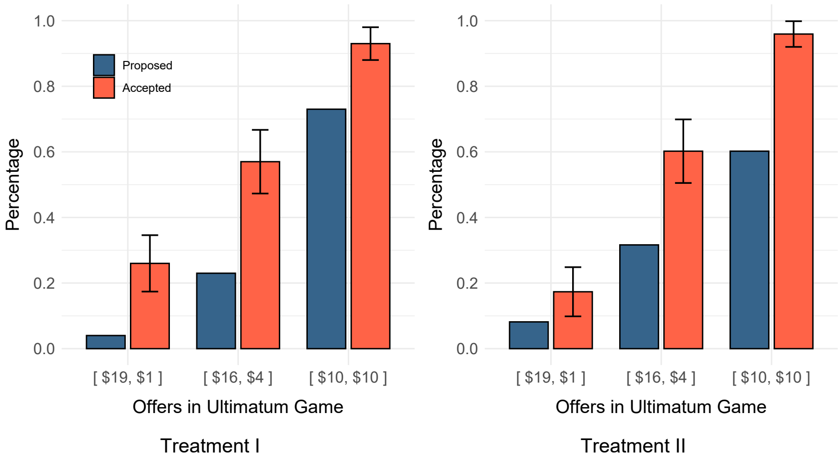


Figure 5: Choices in Ultimatum game

Note: In Treatment I, beliefs are elicited via BB; in Treatment II, via MM. Error bars show 95% CI.

Figure 5 presents the distribution of subjects' proposals and acceptances for each allocation in the Ultimatum game, separately for Treatment I and Treatment II. Fewer than 10% of the proposers chose the most unequal allocation [$19, $1], which corresponds to the subgame perfect Nash equilibrium, while around 27% proposed the second allocation [$16, $4], and the modal and median choice was the equal split [$10, $10], as typically reported in the literature (Kagel and Roth, 1995; Blanco et al., 2010; Trautmann and Kuilen, 2015). Among responders, about 20% accepted the most unequal offer, 58% accepted allocation 2, and nearly all participants accepted the equal split. These acceptance rates align with previous findings that a small number of available options increases acceptance rates of unequal offers (Oosterbeek et al., 2004). Turning to a between-treatment comparison, a slightly higher proportion of subjects in Treatment I made fairer proposals and accepted the most unequal allocation. Thus, the observed proposals and acceptance rates are consistent with documented regularities in ultimatum bargaining, with subjects in Treatment I displaying slightly more altruistic behavior than those in Treatment II.

4.1.2 Response Time and Consistency

Response time and consistency analysis aim to evaluate and compare the practical implementability of the methodologies. Participants in Treatment I required less *time* to complete the BC (UG) experiment compared to those in Treatment II, with average completion times of 20 (39) and 36 (54) minutes, respectively. This difference is likely due to the greater number of questions in the MM (165) compared to the BB Method (44). Looking at the average time spent per choice[5], participants took slightly more time to answer belief questions in the BB (3.9 seconds) than in the MM (3.1 seconds). The considerable total time spent on the Money Method, combined with the relatively short time per binary task, signals upside risks of less deliberate behavior. This could stem from the greater complexity of the decision problem or from survey fatigue, as the tasks are asymmetric and numerous. This issue is examined further in the next paragraphs. Lastly, the average response time for the introspective questions was 13 seconds. Although this figure is not directly comparable due to the different nature of the elicitation task, the Introspective Method stands out relative to the others by enabling a quicker elicitation.

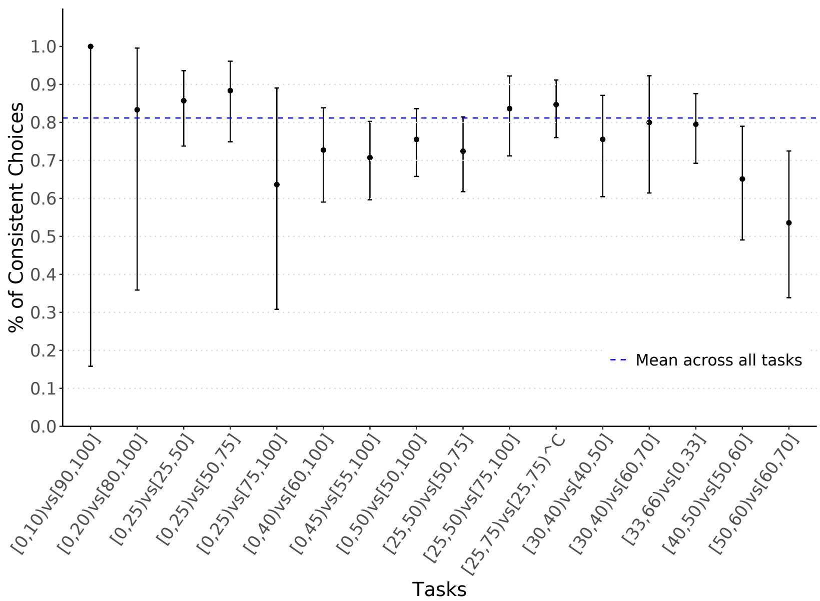


Figure 6: Proportion of consistent subjects in the Bet-Based Method - Experiment 1

Consistency of subjects' choices, an indicator of internal validity for elicitation mechanisms (Manski, 2004; Costa-Gomes and Weizsacker, 2008; Trautmann and Kuilen, 2015; Schlag et al., 2015), is evaluated through two measures: participants' consistency over the belief tasks in repeated choices, and consistency across tasks. The limited data obtained from the Introspective Method do not allow for this non-parametric analysis. The test-retest results show that subjects

[5]The time spent per question is calculated as the difference between the timestamp of the first click and that of the submission. This may underestimate the actual timings, but it provides the active time spent across tasks.

gave identical answers to the four repeated questions in 83% of cases in the MM and 81% in the BB Method across both experiments, indicating high consistency in the belief tasks and no signs of random choice behavior. Notably, in the BB, for Experiment 2, two repeated questions per allocation show higher inconsistency for allocation [\$16, \$4] (23%) than for allocation [\$19, \$1] (17%). This may reflect higher confidence about the acceptance of the latter offer, a pattern that will be confirmed by the next results.

Consistency across tasks is defined by the relation between the lotteries subjects select in the elicitation task and their stated mean belief. These beliefs correspond to the participant's guess in the BC experiment and to the self-reported point beliefs about acceptance of each allocation in the UG[6]. Figure 6 and Table 4 report the consistency in each method for the first Experiment[7]. In the BB treatment, the weighted average consistency is 81% in Experiment 1, and 70% in Experiment 2. In the MM treatment, it is 68% in Experiment 1, and 89% for allocation 1 and 68% for allocation 2 in Experiment 2. The high consistency across tasks registered in both treatments confirms that choices are not random. Thus, the relatively longer time of the incentivized methods seems to pose a challenge mainly for experimental time constraints rather than respondent attention. Moreover, consistency across tasks signals the confidence subjects have in their beliefs. In Experiment 1, Table 4 shows that only subjects who declared a guess within the interval $[70, 90)$ exhibit considerable inconsistency, indicating a positive relationship between confidence and strategic reasoning for subjects in Treatment II. In Experiment 2, the lower consistency for allocation 2 similarly reflects the lower confidence about its acceptance, a finding already signaled in the test-retest results.

Event (Main)	Event (Compared)	Consistent Choices (%)	Tot observations
$[0, 10)$	$[40, 50), [50, 60)$	100.0	4
$[10, 30)$	$[70, 90)$	75.0	12
$[0, 30)$	$[70, 100]$	78.6	14
$[40, 50)$	$[0, 10), [50, 60)$	55.9	34
$[0, 50)$	$[50, 100]$	75.5	49
$[50, 100]$	$[0, 50)$	65.1	43
$[50, 60)$	$[0, 10), [40, 50)$	64.3	28
$[70, 100]$	$[0, 30)$	85.7	7
$[70, 90)$	$[10, 30)$	28.6	7

Table 4: Consistency of Choices in the Money-Method - Experiment 1

[6]For example, suppose a participant's guess in the BC game is 17. In a subsequent BB task, she is asked to choose between Option A, a bet on the number being $[0, 50)$, and Option B, a bet on the number being $(50, 100]$. The participant is consistent if she selects Option A. In the MM, a participant whose guess is 17 is consistent if she chooses the lottery over the sure outcome more frequently when betting on the event $[0, 50)$, than on $[50, 100]$ over the 11 questions per lottery. This measure considers only choices involving equally sized events.

[7]For the BB, the percentage of consistent participants is shown for choices involving equally-sized events. Figure 14 in the Appendix shows data for all BB binary choices in Experiment 1, while Figures 19 and 20 report consistency analysis for allocation 1 and 2, respectively.

4.1.3 Monotonicity

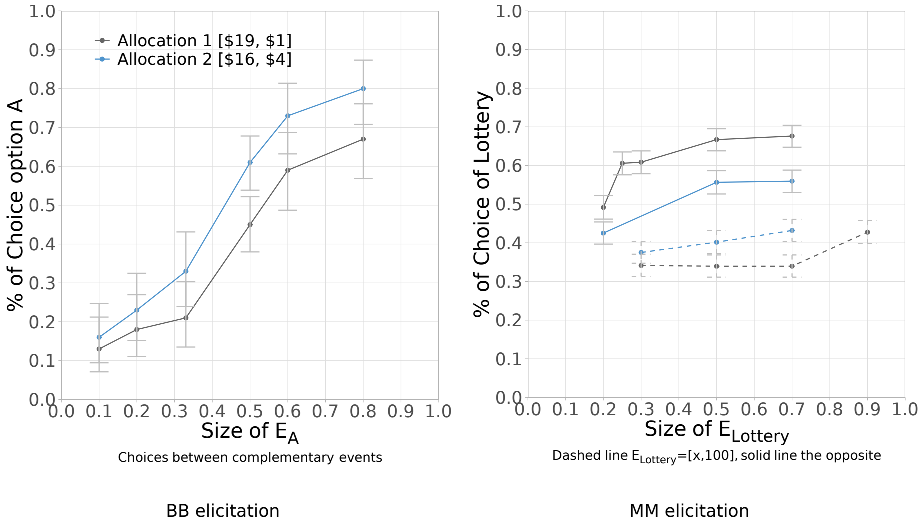


Figure 7: Monotonicity in Choices - Experiment 2

In the Bet-Based Method, monotonicity is satisfied if subjects choose Option A more frequently as E_A, the event determining the positive prize for Option A, increases relative to the complementary event $E_B = E_A^c$, following Abdellaoui et al. (2024b)[8]. Monotonicity in the Money Method is satisfied if the proportion of choices of the bet (Option B) out of 11 binary choices per lottery increases as the size of E_B increases.

The left and right panels of Figure 7 display monotonicity in the BB and the MM, respectively, for Experiment 2. The patterns are shown separately by allocation and, for the MM, also by event type[9]. Results for Experiment 1 are reported in the Appendix. Findings indicate that participants adhere to monotonicity in both experiments, consistent with the earlier evidence that responses are not random. The only exception concerns bets on allocation [$19, $1] being accepted by a high proportion of participants. This likely reflects polarized and stable beliefs that this allocation has a low likelihood of being accepted. Indeed, when the bets on allocation [$19, $1] involve events in the lower-bound of the state space, subjects choose the lottery more frequently - even more than in Experiment 1. Turning to a between-method comparison, monotonicity is weaker in the MM than in the BB, likely because preferences also influence choices in the Money Method.

Lastly, in the Introspective Method, participants report the percent chance that the event of interest exceeds 25 and, separately, 75. Responses violate the monotonicity property of a cumulative distribution function if the reported probability for the second question exceeds that of

[8]In these tasks, the whole state space $[0, 1]$ is covered, and $E_A = [0.5 - \frac{s}{2}, 0.5 + \frac{s}{2}]$, where s is the size of E_A.

[9]Upper-bound events (dashed lines) are $E_B \in [30, 100], [50, 100]$, lower bound events (solid lines) are $E_B \in [0, 50], [0, 70]$).

the first. This occurs for about 35% of the participants in both experiments, a higher proportion than that reported in previous studies (Delavande and Zafar, 2019). In these cases, fitting a parametric belief distribution is not possible. This flags the downside of methods that assume that subjects are able to reason in terms of probability distributions.

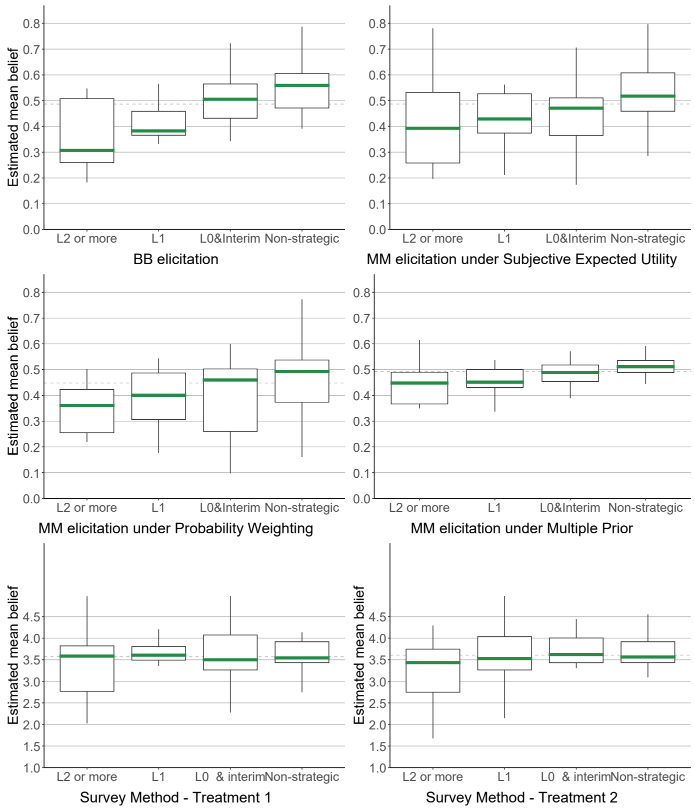
4.2 Parametric Analysis

The parametric analysis relies on comparing belief parameters with observed behaviors, which differ across experiments. For this reason, the results are presented separately for the two games.

4.2.1 The Beauty Contest

Calibration

Calibration of the belief distribution parameters is assessed by analyzing the alignment of the mean beliefs with individual behavior and the uncertainty surrounding the estimated measures.



Note: The dashed gray lines display the median estimates across sample.

Figure 8: Estimated average mean beliefs by level of reasoning

The Panels in Figure 8 display box plots of the estimated means of the belief distributions about the target number in the Beauty Contest game, shown for each method and, for the MM, under each specification. The individual estimates are grouped into four reasoning levels, based on players' declared guesses in the game. This categorization enables a meaningful comparison between methods, particularly for the incentivized ones, as the proportion of individuals across categories is similar between treatments. Conversely, comparability of the estimates *within* each treatment is limited because the number of individuals and their guesses in each category differ[10]. The medians of the mean beliefs, indicated by the dashed gray horizontal line in the figures, are similar across treatments and methods, consistent with the findings of Section 4.1, as the target number is identical in both treatments. For the incentivized methods, the average estimated mean beliefs, shown by the green lines inside the box plots, decrease with lower guesses in the game, indicating good calibration (Seidenfeld, 1985). Conversely, the mean estimates of the log-normal distribution from the Introspective Method, shown in the bottom row of Figure 8, appear less well calibrated, as the downward trend across higher reasoning levels is less evident, particularly in Treatment I. Across specifications, the estimates under PW align most closely with participants' guesses in the game, particularly compared to SEU[11].

The forest plots in Figure 9, which report the mean and 95% highest density interval of the mean and the confidence parameters for 28 randomly selected subjects[12], confirm that uncertainty surrounding the elicited measures is greatest for the Money Method under SEU. Furthermore, a positive and statistically significant within-subject correlation (at the 10% level) is found between the mean beliefs elicited through the Introspective Method and those obtained from the BB. The correlation between the Introspective measures and the MM is significant under PW, but not under SEU and MP, although these remain positive[13].

Finally, the correlation between the belief parameters and ambiguity attitudes is examined to gain further insight into the behavioral drivers in the BC game. Under Probability Weighting (Goldstein and Einhorn, 1987), both preference parameters are significantly correlated with the belief measures (Spearman correlation at the 1% level). Likelihood insensitivity (γ) is positively correlated with both mean belief and belief concentration, a finding consistent with evidence that insensitivity reflects irrationality[14] (Baillon et al., 2018b). Optimism (δ) is negatively correlated with mean beliefs and positively with belief concentration, indicating that more optimistic subjects display less strategic reasoning and lower confidence in their beliefs. Under Multiple Priors, ambiguity attitudes are positively associated with the belief parameters (Spearman correlation at the 10% level or more). Hence, non-strategic individuals and those less confident in their

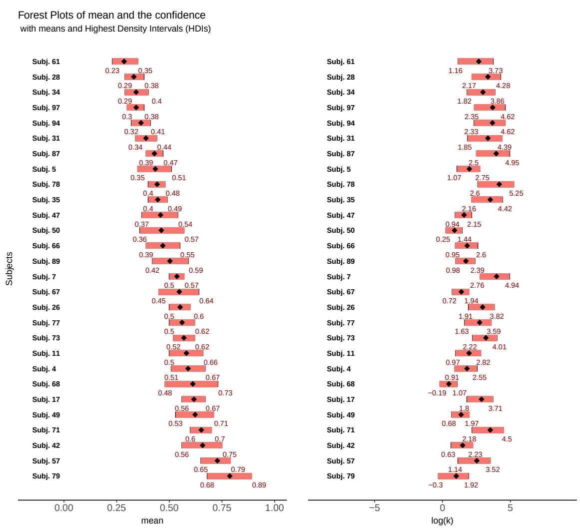
[10] In both treatments, approximately 38% of participants are classified as non-strategic, 29%, named "L0 & Interim", declared a number between 38 and 50, and 20% guessed a number between 30 and 37, aligning with L1 reasoning. Around 13% of players fall into the "L2 or more" group, as they guessed a number lower than 30.

[11] The lower heterogeneity across categories in the MM MP likely reflects the parametric constraints imposed by that specification.

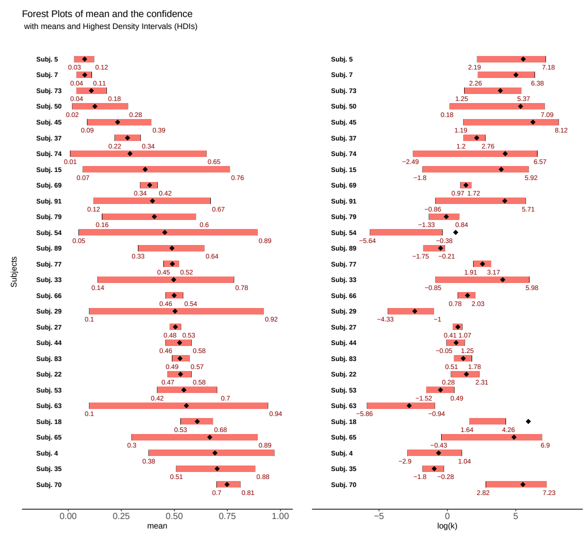
[12] For brevity, I report the remaining plots in Appendix 6.

[13] Within-subject correlations between the confidence measures from the Introspective and incentivized methods are not reported as the variance of the log-Normal and the concentration of the Beta might be structurally different.

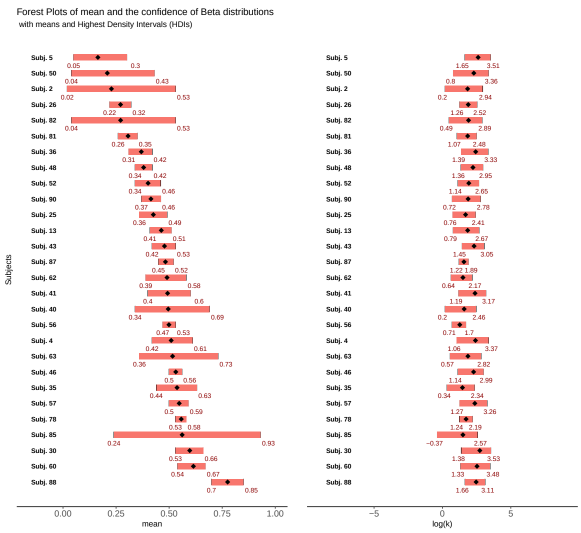
[14] Around 20% of subjects display oversensitivity ($\gamma > 1$), which is thus also positively correlated with the mean.



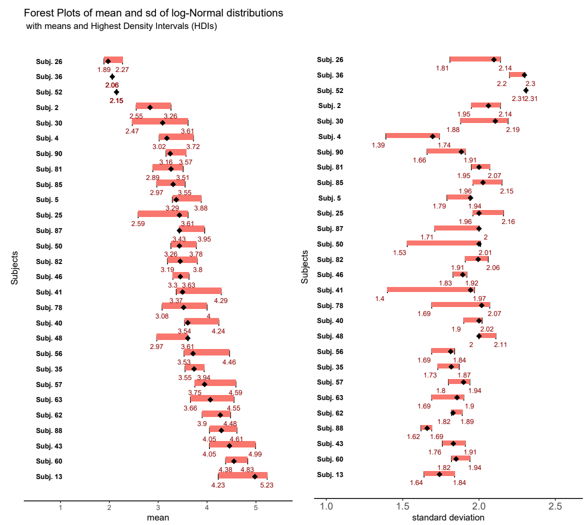
Bet-Based Method



Money Method under SEU



Money Method under PW



Introspective Method - Treatment II

Figure 9: Forest plots: - BC Experiment

beliefs display greater ambiguity aversion and perceive more ambiguity.

Taken together, the calibration analyses show that the BB and the MM under ambiguity models stand out in recovering belief measures that closely align with individuals' declared target numbers in the BC and exhibit low uncertainty. The significant correlations between preferences and belief estimates under both ambiguity models further indicate that in MM these models provide a more reliable framework than SEU. Finally, the relatively weaker calibration of the mean estimates of the Introspective Method may reflect the lower flexibility of the log-normal distribution in capturing heterogeneous beliefs compared to the Beta. Nevertheless, there is considerable variation and low uncertainty around these estimates.

Explaining the Behavior in the Beauty Contest game

To compare the predictive power of the measures elicited by each method, the recovered measures are regressed on the guesses in the BC game using outlier-robust regression[15].

Table 5 and Table 6 report the results for the incentivized methods and the Introspective Method, respectively. For each method, Regression I includes only the estimated mean belief as regressor (e.g., BB I for the Bet-Based Method, SEU I for the MM under SEU, S1 I for the Introspective Method in Treatment I). Regressions II adds the measure of belief confidence to the mean belief (e.g., BB II, SEU II, S II). For the MM under ambiguity models, a third specification (PW III, MP III) adds the ambiguity-attitude parameters to mean beliefs[16].

	Bet-Based		Money Method							
			SEU		PW			MP		
Dep Var: BC guesses	BB I	BB II	SEU I	SEU II	PW I	PW II	PW III	MP I	MP II	MP III
Mean Belief	17.16***	14.39***	10.23**	9.81**	9.01***	9.01*	8.61***	6.14***	6.14***	5.17***
	(4.75)	(4.80)	(4.54)	(4.75)	(3.94)	(4.71)	(2.61)	(3.77)	(3.77)	(3.52)
Belief Concentration		-1.10***		0.02**		0.28***			1.63***	
		(0.38)		(0.02)		(0.08)			(0.08)	
Likelihood Insensitivity							1.86***			
							(2.18)			
Optimism							3.25***			
							(2.88)			
Ambiguity Aversion										3.43
										(3.60)
Ambiguity Perception										8.02***
										(4.44)
Constant	17.9***	2.46***	14.71***	14.32***	12.25***	15.13**	7.16***	14.85***	2.83	7.07***
	(3.62)	(3.77)	(4.09)	(5.13)	(3.73)	(8.25)	(6.30)	(3.78)	(5.14)	(3.04)
Observations	4214	4214	15180	15180	15180	15180	15180	15180	15180	15180
Subjects	98	98	92	92	92	92	92	92	92	92
Bayes R^2	0.122	0.326	0.148	0.203	0.159	0.255	0.256	0.163	0.413	0.167

Note: *** p<0.01, ** p<0.05, * p<0.1, standard errors in parenthesis. In each regression, I control for age and gender, and, in Treatment I, for interaction effects between the mean and confidence. The effects, not reported for parsimony, are positive and statistically significant (95% CI), except for Gender in Reg. PW II and MP II. Bayesian R^2 is computed following Gelman et al. (2019).

Table 5: Regression of Beauty Contest Guesses on elicited Measures

Three main findings emerge from these regressions. First, across all methods, both mean beliefs and belief confidence are highly statistically significant predictors of guesses. This confirms the non-parametric evidence that participants who do not engage in strategic reasoning tend to report higher and more uncertain beliefs. In Treatment I, the interaction between mean beliefs

[15] An econometric outline of the outlier-robust regression is provided in Appendix 6.

[16] A fourth regression (PW IV, MP IV), which also incorporates belief concentration to the regressors of Reg III is not reported, as collinearity among predictors limits the interpretability of the individual coefficients. These and additional regression tables used as robustness checks are available upon request.

and belief uncertainty is positive, indicating that uncertainty matters more for individuals with higher beliefs[17]. In Treatment II, belief uncertainty shows a uniformly positive effect on choices, and the interaction term is omitted due to insignificance, consistent with the more concentrated distribution of guesses at higher levels documented in Section 4.1.

	Introspective Method Treatment I		**Introspective Method Treatment II**	
Dep var: BC guesses	Reg. S1 I	Reg. S1 II	Reg. S2 I	Reg. S2 II
Mean Belief	8.91***	10.17***	11.20***	8.32***
	(1.30)	(1.49)	(1.41)	(2.05)
Belief Variance		6.76***		3.74**
		(1.92)		(1.92)
Age, Gender	NO	NO	NO	NO
Constant	9.47**	3.70	5.09	8.32
	(4.58)	(4.62)	(4.58)	(2.05)
Observations	69	69	55	55
Subjects	69	69	55	55
R^2	-0.044	0.09	0.223	0.216

Note: *** p<0.01, ** p<0.05, * p<0.1, standard errors in parenthesis. Effects of age and gender are excluded as not statistically significant. In Treatment I, I control for interaction effects between mean and variance, which are statistically significant (95% CI) but not reported for parsimony. $R^2 = 1 - \omega_m^2/\omega_0^2$, where ω_0^2 captures the variance in a model with intercept only, and ω_m^2 is the variance in the model with covariates included as data.

Table 6: Regression of BC guesses on Survey Belief Measures

Second, consistent with Abdellaoui et al. (2024b), regressions that incorporate belief confidence achieve substantially higher explanatory power than those relying only on mean beliefs and controls, often with a R^2 that roughly doubles or even triples. Finally, in line with earlier analysis, the explanatory power is highest for the Bet-Based Method and the Money Method under ambiguity attitudes: in these specifications, the mean and concentration of beliefs, along with demographic controls, account for 33% in Regression BB II, 26% in MM PW II, and 41% in MM MP II of the variation in guesses about the target number. These values are higher than in the SEU specification [18], where the explained variance is only 21%. A similar value is found for the explained variance in the Introspective Method, which performs slightly worse in Treatment I than II, in line with calibration analysis.

Regression PW III and MP III also explain sizable variation (26% and 17%), though not exceeding their counterparts PW II and MP II. In the PW specification, the size of the decision maker's set of priors representing ambiguity perception is positively associated with guesses, while ambiguity aversion is not statistically significant. This absence of ambiguity aversion aligns with evidence from studies employing uncertainty sources other than Ellsberg urns (Charness et al., 2013; Trautmann and Kuilen, 2015; Baillon et al., 2018b; Kocher et al., 2018).

[17] Uncertainty also lowers guesses at low mean belief levels (a polarizing effect). Because few subjects fall in this region, the relevant marginal effect is positive.

[18] This comparison is restricted to internally elicited measures, as R^2 across studies is not directly comparable due to differences in experimental tasks and regression analyses.

4.2.2 The Ultimatum game

Estimation of Belief Distributions

Figure 10 displays box plots of the elicited mean beliefs about the likelihood that allocation $[19, 1]$ and $[16, 4]$ will be accepted, shown separately for each elicitation method and, for the Money Method, under each specification. The individual-level estimates are grouped by proposer type, that is, by participants' allocation choice as first movers in the UG[19]. Compared to the actual acceptance frequencies, participants hold conservative beliefs: they are optimistic about the acceptance rates of allocation $[19, 1]$ and pessimistic about responses for allocation $[16, 4]$, in line with previous findings (Huck and Weizsacker, 2002; Costa-Gomes and Weizsacker, 2008; Bellemare et al., 2008; Trautmann and Kuilen, 2015).

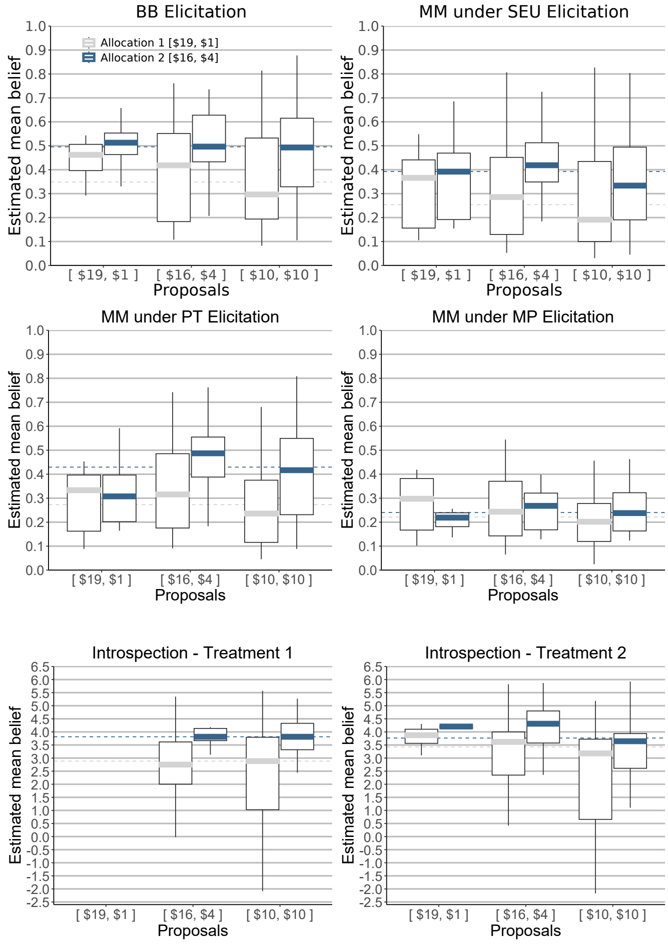


Figure 10: Estimated mean beliefs about the acceptance of each allocation, by proposer type.

The estimates appear well-calibrated with participants' behavior in the game across all elicitation methods. The average mean beliefs, displayed by the horizontal dashed lines, show that

[19]For the introspective estimates from Treatment I, the analysis includes only two proposer types because participants who chose allocation 1 violated the monotonicity condition required to fit their belief distributions.

players generally believe that allocation 1 $[19, 1]$ is less likely to be accepted than allocation 2 $[16, 4]$, a finding in line with actual proposals and monotonicity analysis. This difference is smallest under the Money Method assuming Multiple Prior Models and the Introspective Method for Treatment II, indicating lower calibration under that specification. Beliefs about the likelihood of acceptance vary systematically by proposer type[20], especially for mean estimates of allocation 1. Proposers of allocation 1 hold the highest and most confident probability of its acceptance, and proposers of allocation 2 assign a higher and more confident probability of acceptance of that offer[21] than proposers of allocation 3.Finally, the forest plots in Appendix 6 show that the highest density intervals around belief measures are wider under the Money Method estimated assuming SEU than under the alternative MM specifications, consistent with the findings from Experiment I.

Explaining the Behavior in the Ultimatum game

This subsection compares the performance of the belief measures elicited through the different methods in explaining proposers' choices in the Ultimatum game. To evaluate the explanatory power of the elicited mean, I follow previous studies (Costa-Gomes and Weizsacker, 2008; Trautmann and Kuilen, 2015) and evaluate the proportion of proposers' offers that are predicted as best responses by the mean beliefs of each method. In addition, I compute the subjective valuation loss, defined as the difference between the belief-based optimal valuation and the valuation of the observed offer, under evaluation frameworks specific to each model[22]. Table 7 reports that mean beliefs predict between 60% and 70% of choices, significantly outperforming random prediction across all methodologies ($p < 0.01$). These rates exceed those reported in Trautmann and Kuilen (2015), who find that beliefs elicited via scoring rules account for 30-40% of choices in a Ultimatum game, and are higher than those documented in Costa-Gomes and Weizsacker (2008) and comparable to findings of (Rey-Biel, 2009) for normal-form games. Valuation losses are relatively small across all elicitation methods, indicating that deviations from belief-consistent behavior are limited on average, with higher losses observed under the Money Method when estimated under ambiguity specifications.

[20] There is lower between-subject variability of median mean beliefs about allocation 2 than 1. This aligns with the non-parametric finding that beliefs are more stable for allocation 1 than 2, especially in Treatment I. Taken together with the higher average elicited altruism in Treatment I, this suggests that factors beyond mean beliefs -such as social preferences- may have played a more prominent role in the proposers' decisions in Treatment I.

[21] One exception are the beliefs of proposers of allocation 1 for the MM under PW. However, overall, the sample holds higher and more confident beliefs about allocation 1 than 2, in line with non parametric analysis about consistency.

[22] More details are outlined in the Appendix. Robustness checks available upon request

Measure	Bet-Based Method	Money Method			Introspective Method	
		SEU	PW	MP	Treatment I	Treatment II
Explained Offers (%)	68***	61***	60***	60***	57***	69***
Valuation Loss	0,25	0,31	0,40	0,38	0,22	0,21
Subjects	100	98	98	98	56	58

Note: Fehr and Schmidt (1999) utility and mean beliefs are used to calculate explained offers the UG. Valuation Loss is normalized and calculated under each model's specification. *, **,*** indicates significant better prediction than random at 10%, 5%, 1% significance level, one-sided binomial test

Table 7: Explained Offers in the Ultimatum Game

What explains proposers' valuation losses? I address this question using a fractional logit regression, allowing for the role of factors not entering the valuation function, such as demographics, preferences, and belief confidence. Table 8 shows that these regressors have a small but significant effect on behavior. In particular, belief confidence negatively affects losses in both treatments, indicating that higher belief uncertainty is consistently associated with lower losses. Together with belief uncertainty, in Treatment I, negative reciprocity significantly increases losses for both the Introspective and Bet-Based Methods. In Treatment II, social preferences play no significant role, while risk aversion significantly contributes to reducing valuation losses, as anticipated by previous analysis. Only in the MM under SEU and MP the regression have no explanatory power. Taken together, parametric results indicate that belief measures perform well in explaining behavior across all methodologies, with the Money Method performing similarly to the Bet-Based and the Introspective Method in terms of predictive power only under decision models under PW.

	Bet-Based Method	Money Method			Introspective Method	
Dep var: Valuation Loss		SEU	PW	MP	Treatment I	Treatment II
Belief uncertainty	-0.18**	-0.00	-0.01**	-0.00	-5.32**	-3.70**
	(1.39)	(0.26)	(0.01)	(0.00)	(2.34)	(1.81)
Risk aversion	-0.00	-1.14	-2.71***	-0.16*	-0.07	-0.43***
	(0.09)	(0.84)	(1.00)	(1.25)	(0.11)	(0.15)
Negative reciprocity	1.00***	0.32	0.22	0.37	0.63*	-0.38
	(0.32)	(0.31)	(0.29)	(0.32)	(0.36)	(0.52)
AGE, GENDER	YES	YES	YES	YES	YES	YES
Constant	-0.70	0.225	0.70	-2.63*	1.13	-1.81
	(1.39)	(0.17)	(1.30)	(1.60)	(2.31)	(2.04)
Subjects	100	98	98	98	56	58
$R^2 McFadden$	0.184	0.03	0.096	0.044	0.242	0.283

Note: Belief uncertainty about the acceptance rate of allocation 1 and 2 is calculated as the concentration in the Beta distribution (BB and MM) or variance in the log-normal distribution (Introspection). *** p<0.01, ** p<0.05, * p<0.1, standard errors in parenthesis. Effect of Age and Gender are not reported for parsimony. The dependent variable is normalized, scaled by the difference between the individual maximum and minimum valuations across allocations.

Table 8: Regression of Valuation Losses on Belief Confidence in the Ultimatum game

5 Discussion and Conclusion

This study contributes to the literature on belief elicitation by conducting the first systematic comparison of simple methods for eliciting belief distributions, providing insights into which approach is best suited for different experimental objectives and the trade-offs they entail. The non-parametric results show that the Introspective Method stands out in terms of response time, particularly when compared to the Money Method. Nevertheless, participants remain consistent in their choices throughout the experiment in both incentivized methods, providing little evidence of random responding or survey fatigue resulting from the relatively long elicitation procedure. Monotonicity in choices aligns with the overall analysis for the incentivized methods, whereas for the Introspective Method, it reveals that a non-negligible share of participants may not fully understand probabilities, calling into question the core behavioral assumption of this approach.

Parametric results indicate that both mean beliefs and confidence around them vary systematically with the type of behavior exhibited by subjects across all elicitation methods. In Experiment I, both subjects' mean beliefs and the confidence they hold in these beliefs positively predict strategic reasoning in the Beauty Contest game, particularly when beliefs are elicited through incentivized methods. Experiment II further shows that beliefs about the acceptance of offers in the Ultimatum game differ not only by the strategy chosen by the first mover, but also by the type of action available, with no elicitation method clearly dominating in terms of predictive power. Notably, confidence around mean beliefs emerges as the most robust predictor of the losses subjects incur when deviating from (mean) belief-consistent best responses, across methods. This confirms that belief confidence captures additional aspects driving choices that are not accounted for by mean beliefs alone. Taken together, these findings show the applicability of models of individual decision making to strategic environments, highlighting the explanatory role of belief distributions and their associated confidence.

The results further underscore the importance of uncertainty attitudes across both experiments. Under the Money Method, ambiguity attitudes not only significantly affect choices but also improve the model's explanatory power relative to the standard SEU framework. However, outlier-robust regressions in Experiment I show that ambiguity, together with mean beliefs, explains a substantial yet smaller share of choice variation than confidence combined with mean beliefs. This again highlights the explanatory power of belief confidence in decision making under uncertainty, beyond the role of elicited or controlled preferences.

Ultimately, the choice among methods should be guided by the specific research objective. If the focus is on belief elicitation, the Bet-Based Method appears the most suitable option, as it controls for ambiguity attitudes and minimizes behavioral assumptions. When the aim includes jointly measuring beliefs and preferences, the Money Method under Probability Weighting should be preferable. Lastly, if researchers face constraints on survey length, the Introspective Method may offer the most efficient recovery of beliefs. However, the experimenter may verify participants' understanding of probabilities through comprehension checks or design the tasks so as to circumvent reliance on such assumptions. A comparison of these options is left to future research.

Overall, the results confirm that beliefs are a fundamental component of strategic behavior and that elicitation methods that are easy for participants to understand are effective. Finally, this study provides practical guidance for making informed design choices by evaluating elicitation methods along multiple dimensions and highlighting how to address, rather than overlook, their most critical assumptions.

References

Abdellaoui, M., A. Baillon, L. Placido, and P. P. Wakker (2011): "The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation," *American Economic Review*, 101, 695â723.

Abdellaoui, M., H. Bleichrodt, and C. Gutierrez (2024a): "Unpacking Overconfident Behavior When Betting on Oneself," *Management Science*, 70, 7042–7061.

Abdellaoui, M., H. Bleichrodt, E. Kemel, and O. lâHaridon (2021): "Measuring Beliefs Under Ambiguity," *Operations Research*, 69, 599–612.

Abdellaoui, M., L. Gumus, Y. Kaouane, E. Kemel, and F. M. Vieider (2024b): "Beliefs in Reciprocity, Confidence, and Trust," *Working Paper*, https://www.dropbox.com/scl/fi/jqdzoodxzsrlsfi8a6no3/Beliefs_trust_game.

Baillon, A. (2008): "Eliciting Subjective Probabilities Through Exchangeable Events: An Advantage and a Limitation," *Decision Analysis*, 5, 76–87.

Baillon, A., H. Bleichrodt, U. Keskin, O. l'Haridon, and C. Li (2018a): "The Effect of Learning on Ambiguity Attitudes," *Management Science*, 64, 2181–2198.

Baillon, A., Z. Huang, A. Selim, and P. Wakker (2018b): "Measuring Ambiguity Attitudes for All (Natural) Events," *Econometrica*, 86, 1839–1858.

Bellemare, C., S. Kröger, and A. van soest (2008): "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities," *Econometrica*, 76, 815–839.

Blanco, M., D. Engelmann, and H. Normann (2010): "A Within-Subject Analysis of Other-Regarding Preferences," *Games and Economic Behavior*, 72, 321–338.

Bouchouicha, R., R. Oprea, F. M. Vieider, and J. Wu (2024): "Is Prospect Theory Really a Theory of Choice?" Working Paper Series of Faculty of Economics and Business Administration, Ghent University, Belgium 23/1068.

Bruhin, A., H. Fehr-Duda, and T. Epper (2010): "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion," *Econometrica*, 78, 1375–1412.

Burfurd, I. and T. Wilkening (2022): "Cognitive heterogeneity and complex belief elicitation," *Experimental Economics*, 25, 557–592.

Chapman, J., P. Ortoleva, E. Snowberg, L. Yariv, and C. Camerer (2025): "Reassessing Qualitative Self-Assessments and Experimental Validation," NBER Working Papers 33520, National Bureau of Economic Research, Inc.

Charness, G., U. Gneezy, and V. Rasocha (2021): "Experimental methods: Eliciting beliefs," *Journal of Economic Behavior & Organization*, 189, 234–256.

Charness, G., E. Karni, and D. Levin (2013): "Ambiguity attitudes and social interactions: An experimental investigation," *Journal of Risk and Uncertainty*, 46, 1–25.

Chateauneuf, A., J. Eichberger, and S. Grant (2007): "Choice under uncertainty with the best and worst in mind: Neo-additive capacities," *Journal of Economic Theory*, 137, 538–567.

Costa-Gomes, M. and G. Weizsacker (2008): "Stated Beliefs and Play in Normal-Form Games," *The Review of Economic Studies*, 75, 729–762.

Danz, D., L. Vesterlund, and A. J. Wilson (2022): "Belief Elicitation and Behavioral Incentive Compatibility," *American Economic Review*, 112, 2851–2883.

Delavande, A. (2014): "Probabilistic Expectations in Developing Countries," *Annual Review of Economics*, 6, 1–20.

Delavande, A., J. Ganguli, and F. Mengel (2024): "Uncertainty Attitudes, Subjective Expectations and Decisions under Uncertainty," *R&R Management Science*.

Delavande, A., X. Gine, and D. McKenzie (2011): "Measuring subjective expectations in developing countries: A critical review and new evidence," *Journal of Development Economics*, 94, 151–163.

Delavande, A. and B. Zafar (2019): "University Choice: The Role of Expected Earnings, Nonpecuniary Outcomes, and Financial Constraints," *Journal of Political Economy*, 127, 2343 – 2393.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011): "Individual Risk Attitudes: Measurement, Determinants, And Behavioral Consequences," *Journal of the European Economic Association*, 9, 522–550.

Ellsberg, D. (1961): "Risk, Ambiguity, and the Savage Axioms," *The Quarterly Journal of Economics*, 75, 643–669.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018): "Global Evidence on Economic Preferences," *The Quarterly Journal of Economics*, 133, 1645–1692.

Fehr, E. and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114, 817–868.

Gachter, S. and E. Renner (2010): "The effects of (incentivized) belief elicitation in public goods experiments," *Experimental Economics*, 13, 364–377.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995): *Bayesian data analysis*, Chapman and Hall/CRC.

GHIRARDATO, P., F. MACCHERONI, AND M. MARINACCI (2004): "Differentiating ambiguity and ambiguity attitude," *Journal of Economic Theory*, 118, 133–173.

GHIRARDATO, P. AND M. MARINACCI (2001): "Risk, Ambiguity, and the Separation of Utility and Beliefs," *Mathematics of Operations Research*, 26.

GILBOA, I. AND D. SCHMEIDLER (1989): "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, 18, 141–153.

GOLDSTEIN, W. M. AND H. J. EINHORN (1987): "Expression Theory and the Preference Reversal Phenomena," *Psychological Review*, 94, 236–254.

GUTH, W., R. SCHMITTBERGER, AND B. SCHWARZE (1982): "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior & Organization*, 3, 367–388.

HAO, L. AND D. HOUSER (2012): "Belief elicitation in the presence of naïve respondents: An experimental study," *Journal of Risk and Uncertainty*, 44, 161–180.

HOLLARD, G., S. MASSONI, AND J.-C. VERGNAUD (2015): "In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments," *Theory and Decision*, 80.

HOLT, C. AND A. M. SMITH (2009): "An update on Bayesian updating," *Journal of Economic Behavior & Organization*, 69, 125–134.

HOSSAIN, T. AND R. OKUI (2013): "The Binarized Scoring Rule," *The Review of Economic Studies*, 80, 984–1001.

HUCK, S. AND G. WEIZSACKER (2002): "Do players correctly estimate what others do?: Evidence of conservatism in beliefs," *Journal of Economic Behavior Organization*, 47, 71–85.

HURD, M. (2009): "Subjective Probabilities in Household Surveys," *Annual Review of Economics*, 1, 543–564.

KAGEL, J. H. AND A. E. ROTH (1995): *The Handbook of Experimental Economics*, Princeton University Press.

KARNI, E. (2009): "A Mechanism for Eliciting Probabilities," *Econometrica*, 77, 603–606.

KEMEL, E., T.-E. KAOUANE, Y., AND V. TEIXEIRA (2025): "Measuring Perceived Inequality: an empirical comparison of methods," *Forthcoming*.

KEMEL, E. AND S. MUN (2024): "(In)Consistency of Beliefs and Attitudes Under Uncertainty: Every Cloud has a Silver Lining," *Working Paper*.

KLIBANOFF, P., M. MARINACCI, AND S. MUKERJI (2005): "A smooth model of decision making under ambiguity," *Econometrica*, 73, 1849–1892.

KOCHER, M. G., A. M. LAHNO, AND S. T. TRAUTMANN (2018): "Ambiguity aversion is not universal," *European Economic Review*, 101, 268–283.

LATTIMORE, P. K., J. R. BAKER, AND A. D. WITTE (1992): "The influence of probability on risky choice: A parametric examination," *Journal of Economic Behavior & Organization*, 17, 377–400.

MANSKI, C. F. (2004): "Measuring Expectations," *Econometrica*, 72, 1329–1376.

MAUERSBERGER, F. AND R. NAGEL (2018): "Levels of reasoning in Keynesian Beauty Contests: a generative framework," in *Handbook of computational economics*, Elsevier, vol. 4, 541–634.

Nagel, R. (1995): "Unraveling in guessing games: An experimental study," *The American economic review*, 85, 1313–1326.

Offerman, T., J. Sonnemans, G. V. D. Kuilen, and P. P. Wakker (2009): "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes ," *The Review of Economic Studies*, 76, 1461–1489.

Oosterbeek, H., R. Sloof, and G. Kuilen (2004): "Cultural Differences In Ultimatum Game Experiments: Evidence From A Meta-Analysis," *Experimental Economics*, 7, 171–188.

Palfrey, T. R. and S. W. Wang (2009): "On eliciting beliefs in strategic games," *Journal of Economic Behavior Organization*, 71, 98–109.

Papke, L. E. and J. M. Wooldridge (1996): "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates," *Journal of Applied Econometrics*, 11, 619–632.

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

Rey-Biel, P. (2009): "Equilibrium play and best response to (stated) beliefs in normal form games," *Games and Economic Behavior*, 65, 572–585.

Richard, T., B. Loic, T. Epper, M.-J. Montoya-Villalobos, I. Aydogan, U. Turmunkh, and E. Kemel (2025): "Eliciting Beliefs: A Multi-Method Comparison," *Wroking Paper.*

Savage, L. J. (1954): *The foundations of statistics*, John Wiley & Sons.

Schlag, K. H., J. Tremewan, and J. J. Van der Weele (2015): "A penny for your thoughts: A survey of methods for eliciting beliefs," *Experimental Economics*, 18, 457–490.

Schmeidler, D. (1989): "Subjective Probability and Expected Utility without Additivity," *Econometrica*, 57, 571–87.

Schotter, A. and I. Trevino (2014): "Belief Elicitation in the Laboratory," *Annual Review of Economics*, 6, 103–128.

Seidenfeld, T. (1985): "Calibration, Coherence, and Scoring Rules," *Philosophy of Science*, 52, 274–294.

Train, K. E. (2009): *Discrete choice methods with simulation*, Cambridge university press.

Trautmann, S. T. and G. Kuilen (2015): "Belief Elicitation: A Horse Race among Truth Serums," *Economic Journal*, 125, 2116–2135.

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

Vieider, F. M. (2018): "Violence and Risk Preference: Experimental Evidence from Afghanistan: Comment," *American Economic Review*, 108, 2366â82.

——— (2024): *Bayesian Estimation of Decision Models.*

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

WAKKER, P. (1994): "Separating marginal utility and probabilistic risk aversion," *Theory and Decision*, 36, 1â44.

WAKKER, P. P. (2010): *Prospect theory: For risk and ambiguity*, Cambridge university press.

WANG, S. (2011): "Incentive effects: The case of belief elicitation from individuals in groups," *Economics Letters*, 111, 30–33.

WISWALL, M. AND B. ZAFAR (2015): "Determinants of College Major Choice: Identification using an Information Experiment," *The Review of Economic Studies*, 82, 791–824.

WU, G., J. ZHANG, AND R. GONZALEZ (2004): "Decision under risk," *Blackwell handbook of judgment and decision making*, 399–423.

6 Appendix

A. The Experiment

The number of decision tasks for the MM and BB was determined with hierarchical econometric tools to recover belief distributions through simulations. It represents the optimal amount of stimuli that ensures comparability between the methods, maximizes the power to identify the belief distribution over a wide range of outcomes, and minimizes the burden of the numerous tasks for the respondents. Particularly, the Spearman correlation between simulated and estimated values is 0,5-0,8, depending on the simulated parameters.

Experiment 2

Example of belief tasks

In Experiment 2, I elicit the proposer's subjective probability that the responder will accept allocations 1 and 2. To avoid assuming participants' understanding of percentages, I formulated all choices based on the number of *acceptors* out of 100, i.e., the number of responders out of 100 participants accepting each offer. Figures 11, 12, and 13 illustrate a task for each method.

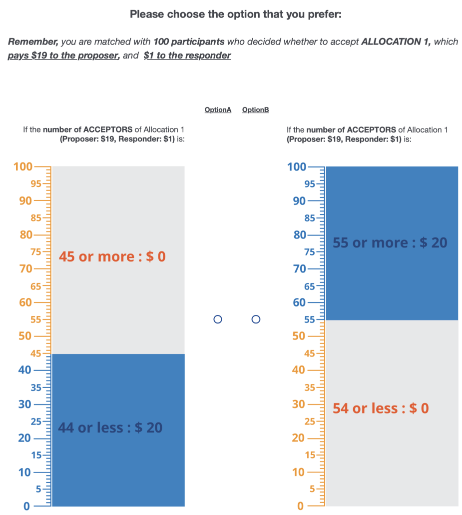


Figure 11: Ultimatum game - a choice task for the BB method

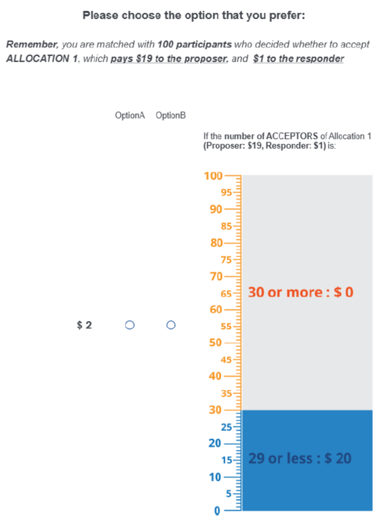


Figure 12: Ultimatum game - a choice task for the MM method

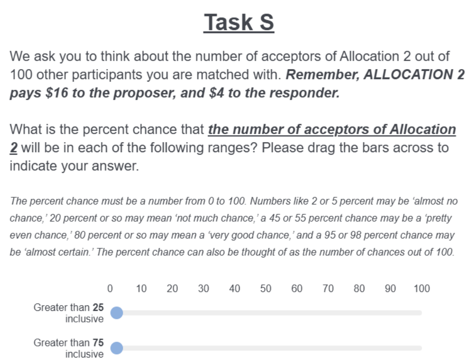


Figure 13: Ultimatum game - choice tasks for the Introspective Method

In the Introspective Method, differing from Experiment 1, I further added a reminder in the instructions about the inconsistency of responses that violate the monotonicity property of a cumulative distribution function. This approach follows previous studies in which monotonicity is not enforced but encouraged (Trautmann and Kuilen, 2015).

Preferences

1) *Altruism.* The measure of altruism consists of 2 questions, weighted in accordance with the Global Preferences Survey recommendations. One hypothetical donation decision (weight: 0.635) and an 11-point self-assessment of willingness to give to good causes (weight: 0.365).

2) *Risk taking*: Self-assessed willingness to take risks (11-point scale), shown to significantly predict risk-taking in both field settings and incentivized experiments across diverse populations (Dohmen et al., 2011; Vieider et al., 2015).

3) *Negative reciprocity*: Minimum acceptable offer in UG.

B. Econometrics

The Money Method

The Money Method elicits the disentangled measures of belief and attitudes under three classes of models, which are special cases of the Biseparable Model (Ghirardato and Marinacci, 2001). First, I assume that preferences admit a *Subjective Expected Utility* (SEU) representation (Savage, 1954). Subject i's valuation of the sure amount c_j and the lottery $L_j := (x_{E_j^B} y)$ in task j are given by:

$$V_i^{c_j} = u_i(c_j) \tag{7}$$

$$V_i^{L_j} = P_i(E_j)u_i(x_j) + (1 - P_i(E_j))u_i(y_j), \tag{8}$$

where $P_i(E_j)$ represents (mean point) belief about E_j, and $u()$ denotes the utility function, expressed in its power form for flexibility and interpretability (Bruhin et al., 2010):

$$u_i(x) = x^{\phi_i}$$

Here, x is a non-negative monetary payoff, and $\phi_i > 0$ measures utility curvature: $\phi_i > 1$ indicates convex utility, $\phi_i = 1$ linear utility, and $\phi_i < 1$ concave utility.

Next, I extend the SEU model to *Probability Weighting* (PW) Models by allowing for subjective transformations of probabilities into decision weights by means of a probability weighting function $W(P(E))$. Prospect Theory (Tversky and Kahneman, 1992), Choquet Expected Utility (Schmeidler, 1989; Wakker, 2010), and Rank-Dependent Utility (Quiggin, 1982; Wakker, 1994) are special cases embedded in this specification. The certain amount is evaluated as in Equation

(7), while the bet L_j is evaluated as:

$$V_i^{L_j} = W_i(P_i(E_j))u_i(x_j) - (1 - W_i(P_i(E_j)))u_i(y_j) \tag{9}$$

The probability weighting function follows the specification proposed by Goldstein and Einhorn (1987) and Lattimore et al. (1992), which captures individual heterogeneity well (Wu et al., 2004):

$$W_i(P_i(E_j)) = \frac{\delta_i P_i(E_j)^{\gamma_i}}{\delta_i P_i(E_j)^{\gamma_i} + (1 - P_i(E_j))^{\gamma_i}},$$

where δ_i controls the elevation of the curve (optimism), and γ_i the curvature (likelihood-sensitivity). This model coincides with SEU (Savage, 1954) when $\delta_i = \gamma_i = 1$.

Finally, the *Multiple Prior* (MP) specification assumes subjects represent ambiguity through a non-singular set of priors. This interpretation includes the Maxmin EU (Gilboa and Schmeidler, 1989) and Alpha-Maxmin (Ghirardato et al., 2004). The subjective valuation of L_j is:

$$V_i^{L_j} = \alpha_i \min_{\pi_i \in \vartheta_i} [\pi_i(E_j)U_i(x_j) + (1 - \pi_i(E_j))U_i(y_j)] + (1 - \alpha_i) \max_{\pi_i \in \vartheta_i} [\pi_i(E_j)U_i(x_j) + (1 - \pi_i(E_j))U_i(y_j)] \tag{10}$$

Here, α_i represents the relative index of ambiguity aversion, and ϑ_i is the size of the decision maker's set of priors (ambiguity perception). To recover the parameter of this specification, we follow Baillon et al. (2018a) and assume the Neo-additive preferences of Chateauneuf et al. (2007), defined as:

$$W_i(P_i(E_j)) = \begin{cases} 1/2(a_i - b_i) + (1 - a_i)P_i(E_j) & \text{if } 0 \leq P_i(E_j) \leq 1 \\ 1 & \text{if } P_i = 1 \\ 0 & \text{if } P_i = 0 \end{cases}$$

The parameters a_i and b_i are such that $0 \leq a_i \leq 1$ and $-a_i \leq b_i \leq a_i$. Then, $\alpha_i = \frac{1}{2}(1 + \frac{b_i}{a_i})$ if $a_i > 0$ and $\alpha_i = \frac{1}{2}$ if $a_i = 0$ and $\vartheta_i = \{\pi_i : (1 - a_i)P(E_j) \leq \pi(E_j) \leq 1 - (1 - a_i)(1 - P(E_j))\}$. The Multiple Priors interpretation thus does not allow for oversensitivity (a_i>1). The recovery of the parameters is done within Bayesian Hierarchical Econometrics. Further details about the latter two elicitation are outlined in Baillon et al. (2018a)[23].

[23] The main difference is that they use Certainty Equivalents and Nonlinear Least Squares, whereas I econometrically recover the measure within a single model in Stan (see Section 6).

Explaining Action Choice

The Beauty Contest game

In Experiment 1, the estimated parameters for subjective belief distributions from each method are compared in terms of their performance in explaining subjects' actions in the Beauty Contest game. To achieve this, subjects' guesses for the target number in the game are regressed on the belief measures from each elicitation mechanism within an *outlier-robust regression model* taking the form of Student-t distributions with 3 degrees of freedom[24]. This helps to guard against disproportionate effects of precisely estimated outliers, given that imprecise outliers are discounted within the Bayesian hierarchical setting (see Gelman et al. (1995)). Let BC_i denote subject i's guess in the Beauty Contest game. It is assumed that BC_i follow a normal distribution centered at a single value denoted by μ_{BC_i}, which is given by:

$$\mu_{BC_i} = \gamma_0 \mathbf{X} + \gamma_1 m_i^{\mathcal{M}} \tag{11}$$

I define $m_i^{\mathcal{M}}$ as the mean value of subject i's belief distribution elicited via the belief elicitation mechanism $\mathcal{M}$, defined in Subsection 3.1 and $\mathbf{X}$ is the matrix of control factors (i. e. demographic characteristics such as gender and age). Accordingly, γ_0 and γ_1 denote the parameters of interest. I first ran this regression (Reg. I in the main text) to assess the interpretive value of this parameter. Subsequently, other variables are included in the regression model, such as the confidence around the mean belief distribution (Reg. II) and, for the Money Method, ambiguity attitudes (Reg. III).

The Ultimatum game

In Experiment 2, the performance of the three belief elicitation mechanisms is compared in terms of their ability to explain the first mover's choice of the allocation $k \in \{1, 2, 3\}$ proposed to the responder in the game. Following (Costa-Gomes and Weizsacker, 2008; Trautmann and Kuilen, 2015), I first compute each subject's *best response*, defined as the allocation to propose assuming individual maximize Fehr and Schmidt (1999) social utility function, calculated with the elicited altruism individual measures, and $m_{i,k}^{\mathcal{M}}$ mean belief that allocation k will be accepted[25] under the elicitation mechanism $\mathcal{M}$. This allows comparison of mean performance across elicitation methods. Particularly, I evaluate the proportion of the observed proposals that correspond to the predicted best response.

I then compute the subjective valuation associated to each allocation. For all methods, I employ the Fehr and Schmidt (1999) utility function and acceptance probabilities $P_i(E_k) = m_{i,k}^{\mathcal{M}}$

[24] In the robustness checks available upon request, the degrees of freedom are further calculated endogenously

[25] Over all analysis, we assume that the fair proposal will be accepted for sure, that is $m_{i,3}^{\mathcal{M}} = 1$.

and an evaluation framework that differs by model. Specifically, the Bet-Based method, the Money Method under SEU, and the Introspective Method are evaluated using Equation 8; the Money Method under PW uses Equation 9; and the Money Method under MP uses Equation 10. The *subjective valuation loss* is defined as the difference between the valuation of the belief-based best response and that of the observed choice. This loss is normalized by the subject-specific valuation range, which is the difference between the maximum and minimum valuations across allocations, yielding a measure bounded between 0 and 1.

Finally, subjects' valuation losses are regressed on the mean of the belief confidence about the acceptance of allocation 1 and 2, demographics and preference measures[26] that do not enter the valuation functions. Estimation is conducted using a fractional regression model (Papke and Wooldridge, 1996), with linear regressions used as robustness checks[27].

C. Estimation code for the Money Method

Beliefs and attitudes in the Money Method are estimated using a Bayesian hierarchical model coded in Stan (see Vieider (2024) for a tutorial to recover in the context of decision making under risk). Convergence is carefully checked using best practices recommended by the Stan community. Below, I report the Stan code used in the Ultimatum game to estimate the belief distributions over the acceptance probabilities of Allocations 1 and 2. This formulation captures the correlation between belief and attitude parameters through a unified correlation matrix, allowing joint estimation within a single model. The current implementation assumes the underlying decision model follows Probability Weighting, but the code can be adapted for SEU (assuming damma=delta=1) or Multiple Priors (substituting the Goldstein and Einhorn (1987) weighting function with the one of Chateauneuf et al. (2007)), as detailed in Section 6). The estimation of the model for Experiment 1 - where beliefs about the likelihood of the target number are elicited - can be obtained by simplifying the following code such that `m` and `k` are single vectors of length `Nid`, rather than matrices of dimension $[Nid, 2]$, since only one belief is elicited per individual. Comments are preceded by //.

```
data {
  int<lower=1> N;                        // tot  nr of observations
  int<lower=1> Nid;                      // tot nr of individuals
  array[N] int id;                       // (sequential) individuals
  vector[N] lowa;                        // lower bound event of option A
  vector[N] upa;                         // upper bound
  vector[N] sure;                        // sure amount
```

[26]For Risk attitudes, I use the parameter ϕ_i elicited in the belief elicitation tasks for the MM under SEU and PW, and the inverse of the self-assessed willingness to take risk in the remaining models.

[27]Available upon request

```
  vector[N] high;                            // high payoff amount of lottery
  vector[N] low;                            // low payoff amount of lottery
  array[N] int choice_u;                    //=1 if option B (uncertain) is chosen
  int<lower=1> N_alloc;                     // tot nr allocations (3)
  array[N] int<lower=1,upper=N_alloc-1> al;//(sequential) allocations
}
transformed data {
  vector[N] logit_qa;
  vector[N] logit_pa;
  for (i in 1:N) {
    logit_qa[i] = logit(upa[i] - 0.00001);
    logit_pa[i] = logit(lowa[i] + 0.00001);
  }
}
parameters {
  vector[7] means;                   // vector of means ( first 2 allocation means, then  2 alloc variances, then gamma delta rho)
  vector<lower=0>[7] tau;            // vector of sds of parameters
  cholesky_factor_corr[7] L_omega;   // Cholesky f. for parameters correlation
  array[Nid] vector[7] Z;            // standard normal
  real<lower=0> sigma;               // variance of the error term
}
transformed parameters{
  array[Nid] vector[7] pars;
  matrix[Nid, N_alloc-1] m;          // mean beliefs (for the 2 allocations)
  matrix[Nid, N_alloc-1] k;          // concentration of beliefs
  matrix[Nid, N_alloc-1] m_logit;    // mean logit distribution
  matrix[Nid, N_alloc-1] sd_logit;   // standard deviation logit distribution
  vector<lower=0>[Nid] gamma;        // GE weighting function insensitivity/slope
  vector<lower=0>[Nid] delta;        // GE weighting function optimism/elevation
  vector<lower=0>[Nid] rho;          //  utility task

  for (n in 1:Nid){
    pars[n] = means + diag_pre_multiply(tau, L_omega) * Z[n]; //matrix of hierarchical parameters on correct scale&correlation
    for (x in 1: (N_alloc-1)){
      m[n, x] = inv_logit( pars[n, x] );//ensure values between 0 and 1
      k[n, x] = exp( pars[n, x + 2] ); //exponential to ensure positivity
      m_logit[n, x] = digamma(m[n, x] * k[n, x]) - digamma((1 - m[n, x]) * k[n, x]);
      sd_logit[n, x] = sqrt( trigamma(m[n, x] * k[n, x]) + trigamma( (1 - m[n, x]) * k[n, x]) );
    }
    gamma[n]=exp(pars[n,5]);
```

```
    delta[n]=exp(pars[n,6]);
    rho[n]=exp(pars[n,7]);
  }
}
model {
  vector[N] Pa;
  vector[N] wp;                   //weighting function
  vector[N] pv;                   //prospect value
  vector[N] udiff;               //utility difference between choosing pv and sure

  for (x in 1 : 2 ){
    means[x] ~ normal( 0, 5);    //aggregate mean beliefs (uninformative prior)
  }
  for (x in 3 : 4 ){
    means[x] ~ normal( 2.3, 5); //aggregate k beliefs (uninformative prior)
  }
  means[5] ~ normal(0, 5);
  means[6] ~ normal(0, 5);
  means[7] ~ normal(0, 5);
  for (n in 1:Nid)
  Z[n] ~ std_normal( );       //matrix of uncorrelated samples (z-scores)

  tau ~ exponential(5);            //weakly regularising priors for the SDs
  L_omega ~lkj_corr_cholesky(3); //(weakly informative) prior for correl matrix
  sigma ~ normal( 0, 2);        // aggregate residual variance

  for (n in 1:Nid)
  Z[n] ~ std_normal( );           //matrix of uncorrelated samples (z-scores)

  for (i in 1:N) {
    Pa[i] = normal_cdf(logit_qa[i]| m_logit[id[i],al[i]], sd_logit[id[i],al[i]])
    - normal_cdf( logit_pa[i] | m_logit[id[i], al[i]], sd_logit[id[i], al[i]]);
   }
  wp = ( delta[id] .* Pa.^gamma[id] )
  ./ ( delta[id] .* Pa.^gamma[id] + ( 1 - Pa ).^gamma[id] );
  pv = ( wp .* high.^rho[id]) + ((1 - wp) .* low.^rho[id] );
  udiff = ( pv - sure.^rho[id] );
  choice_u ~ bernoulli_logit(udiff ./ ( sqrt(2)* sigma .* (high-low) ) );
}
```

Listing 1: Stan code for MM belief estimation

D. Results

Experiment 1

Binary Choice Tasks and Choice Proportions

Task ID	E_{A_j}	E_{B_j}	% Option A	Task ID	E_{A_j}	E_{B_j}	% Option A
1	[0,0.10)	[0.9,1]	51.0	2	[0.25,0.75)	[0,0.25)	89.8
2	[0,0.20)	[0.80,1]	55.1	22*	[0.25,0.75)	[0,0.25) ∪ [0.75,1]	85.7
3	[0,0.25)	[0.25,0.5)	18.4	23	[0.25,0.75)	[0.75,1]	90.8
4	[0,0.25)	[0.50,0.75)	31.6	24	[0.30,0.40)	[0.40,0.50)	35.7
5*	[0,0.25)	[0.75,1]	57.1	25	[0.30,0.40)	[0.60,0.70)	60.2
6	[0,0.33)	[0.66,1]	52.0	26	[0.33,0.66)	[0,0.33)	74.5
7	[0,0.40)	[0.60,1]	47.9	27	[0.33,0.66)	[0,0.33) ∪ (0.66,1]	50.0
8	[0,0.45)	[0.55,1]	56.1	28	[0.33,0.66)	[0.66,1]	76.5
9*	[0,0.50)	[0.50,1]	53.0	29	[0.40,0.50)	[0.50,0.60)	62.2
10*	[0.10,0.30)	[0,0.10)∪ [0.30,1]	10.2	30	[0.40,0.60)	[0,0.40)	33.7
11	[0.10,0.50)	[0,0.10)∪ [0.50,1]	44.9	31	[0.40,0.60)	[0,0.40) ∪ [0.60,1]	30.6
12	[0.10,0.90)	[0,0.10)	94.9	32	[0.40,0.60)	[0.60,1]	55.1
13	[0.10,0.90)	[0,0.10)∪ [0.90,1]	93.9	33	[0.45,0.55)	[0,0.45)	23.5
14	[0.10,0.90)	[0.90,1]	93.9	34	[0.45,0.55)	[0,0.45)∪[0.55,1]	18.4
15	[0.20,0.40)	[0,0.20)∪ [0.40,1]	24.5	35	[0.45,0.55)	[0.55,1]	36.7
16	[0.20,0.80)	[0,0.20)	85.7	36	[0.50,0.60)	[0.60,0.70)	68.4
17	[0.20,0.80)	[0,0.20)∪ [0.80,1]	85.7	37	[0.50,0.90)	[0,0.50)∪ [0.90,1]	28.6
18	[0.20,0.80)	[0.80,1]	89.8	38	[0.60,0.80)	[0,0.60)∪ [0.80,1]	14.3
19	[0.25,0.50)	[0.50,0.75)	61.2	39	[0.70,0.90)	[0,0.70)∪ [0.90,1]	13.3
20	[0.25,0.50)	[0.75,1]	73.5				

Superscript * indicates that the task has been shown to subjects twice.

Table 9: Binary Choices in Experiment 1 - BB elicitation

Task ID	E_L	% Lottery Choice	Task ID	E_L	% Lottery Choice
E_1	[0,0.10)	0.32 (0.03)	E_1^c	[0.10,1.00)	0.60 (0.03)
E_2	[0,0.30)	0.40 (0.03)	E_2^c	[0.30,1.00)	0.44 (0.03)
E_3	[0,0.50)	0.46 (0.03)	E_3^c	[0.50,1.00)	0.40 (0.03)
E_4	[0,0.70)	0.58 (0.03)	E_4^c	[0.70,1.00)	0.31 (0.03)
E_5	[0.40,0.50)	0.32 (0.02)	E_7	[0.10,0.30)	0.60 (0.03)
E_6	[0.50,0.60)	0.33 (0.03)	E_8	[0.70,0.90)	0.33 (0.03)

Standard errors in parentheses. Each task involves 11 choices with sure amounts varying from 0 to 20. E_L^c denotes the complement of E_L. Choosing the lottery yields (20,0): \$20 if E_L occurs, \$0 otherwise. Additionally, 29 choices for task E_3 with payoffs (10,0), (20,5), (20,10), and (15,5) elicited preferences.

Table 10: List of Binary Tasks and Proportion of Lottery Choices in Experiment 1 - MM elicitation

Consistency

Figure 14 shows the proportion of subjects consistent in their choices with their stated BC game answer by task for the BB elicitation. The blue dashed line displays the average proportion of consistent choices, weighted for the number of observations per task. Two binary choice tasks span the entire state space and thus include all subjects: the choice between the events $[0, 50)$ and $[50, 100]$ and the choice between $[25, 75)$ and its complement $[25, 75)^C$. In these tasks, 75 and 83 participants, respectively, chose the consistent option, providing insight into overall behavior. On average, participants chose the lottery more often in the tasks where their declared guesses fell into the event that pays the prize in the corresponding lottery across all the tasks.

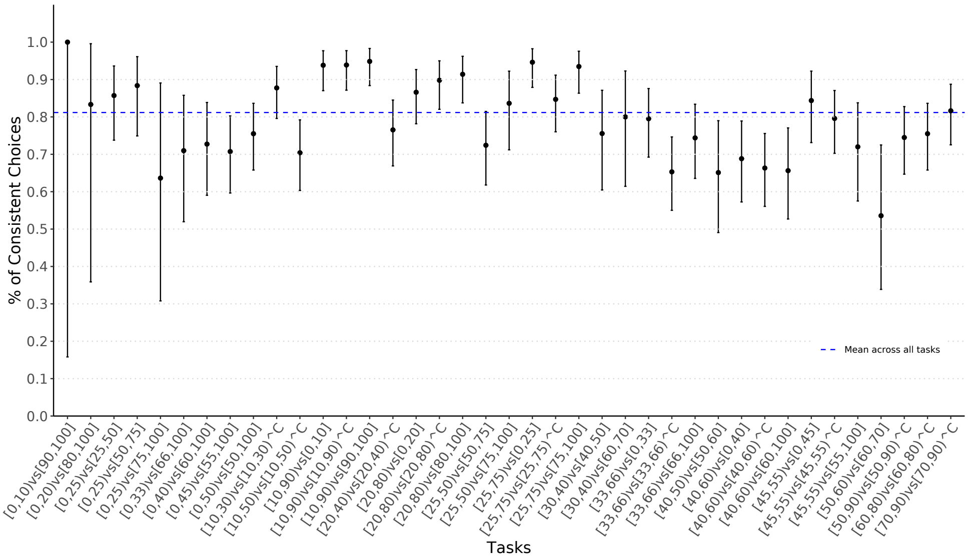


Figure 14: Proportion of Consistent Subjects in Experiment 1 - Bet-Based Method

Monotonicity

The Figure 15 shows monotonicity in choices for Experiment 1. In the right panel, lotteries with equal event sizes are merged. Table 10 reports the proportion of lottery choices for each set of tasks separately.

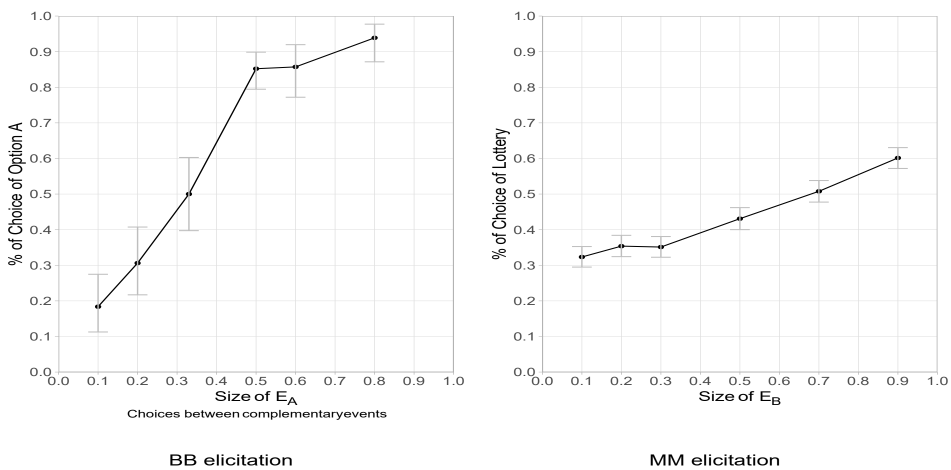


Figure 15: Monotonicity in choices

Parametric Results

Forest plots

Below are reported the forest plots about the subjects not shown in the main text for the BB, MM under SEu, and Introspective Method in Treatment I (BB Method). Forest Plots under further specifications are available upon request.

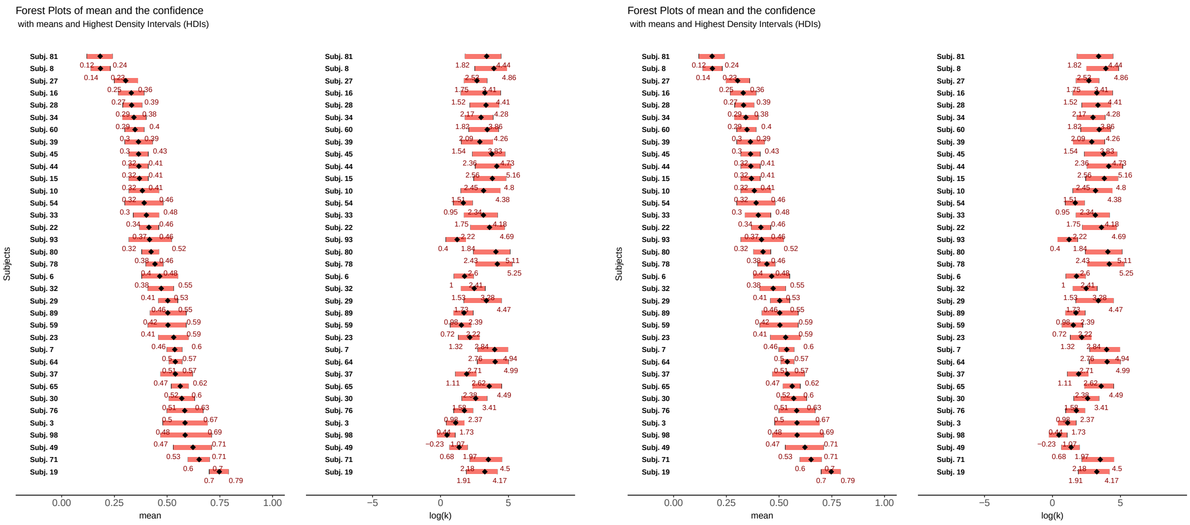


Figure 16: Forest plot: BC - Bet-Based Method

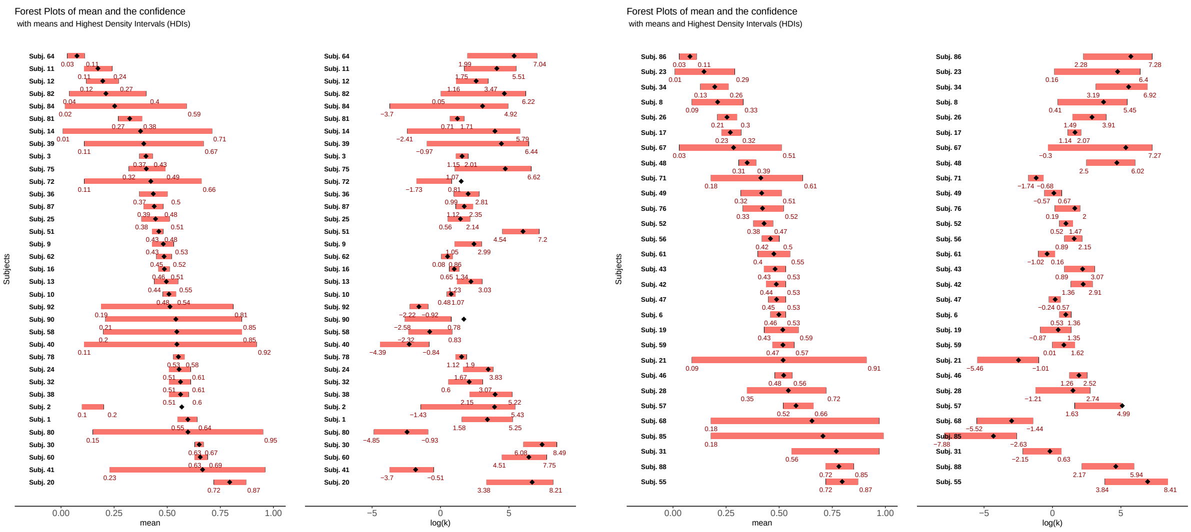


Figure 17: Forest plot: BC - Money Method under SEU

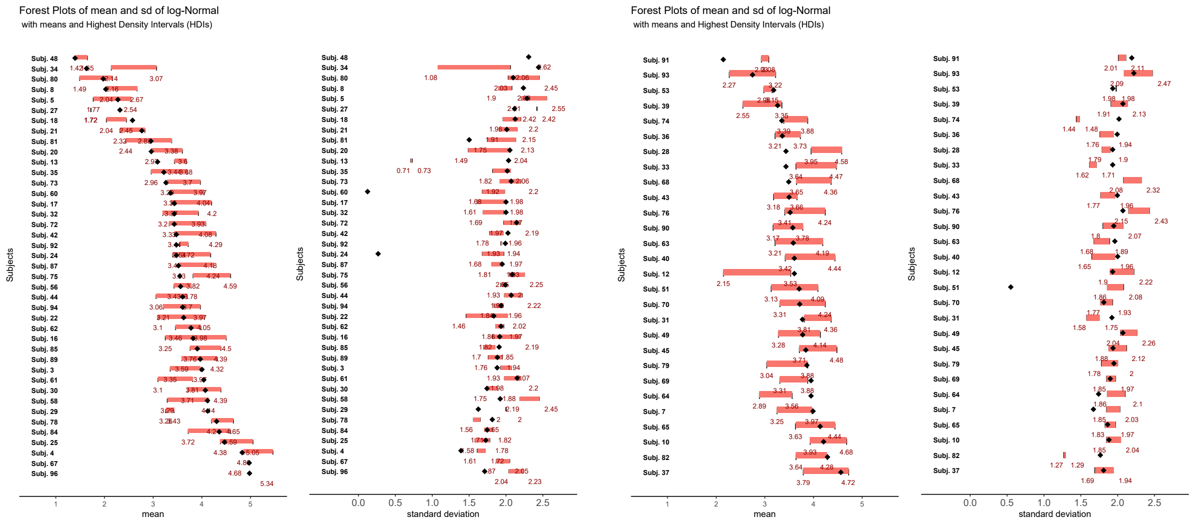


Figure 18: Forest plot: BC - Introspective Method Treatment I

Experiment 2

Binary Choice Tasks and Choice Proportion

In *Experiment 2*, subjects answered a total of 198 questions, with the events and binary choices varying between the elicitation of the belief about the acceptance probability of allocation 1 or allocation 2 in the UG. The sure amount in option B changes from 0 to 20 in 11 steps equal between the two experiments.

For allocation 1 ($19 to the proposer, $1 to the responder), participants are required to answer 11 binary choices for the following events: $E_1 = [0, 30]$, $E_2 = [0, 50]$, $E_3 = [0, 70]$ and their complementary events $E_1^c = [30 - 100]$, $E_2^c = [50 - 100]$, $E_3^c = [70 - 100]$, and E_4=[10,30], $E_5 = [10 - 100]$, $E_6 = [0 - 25]$. These events are also asked in Experiment 1, except for E_6. In total, I thus asked 99 questions.

For allocation 2 ($16 to the proposer, $4 to the responder), subjects answered 11 binary choices for the following events: $E_1 = [0, 70]$, E_2=[10,30], $E_3 = [30 - 100]$, $E_4 = [70 - 100]$. Furthermore, to identify the utility function, a total of 40 tasks were employed for event $E_5 = [0, 50]$ with the payoffs equal to the ones to elicit utility in Experiment 1, but additional sure amounts to improve elicitation. Finally, 4 questions were asked again to check for consistency, for a total of 99 questions. In line with Treatment I, in Experiment 2, I asked the same number of questions for each allocation.

Consistency

Figure 19 and Figure 20 shows the proportion of subjects consistent in their choices in the BB elicitation tasks with their stated mean belief in the *introspective* answer, for allocation 1 and 2, respectively.

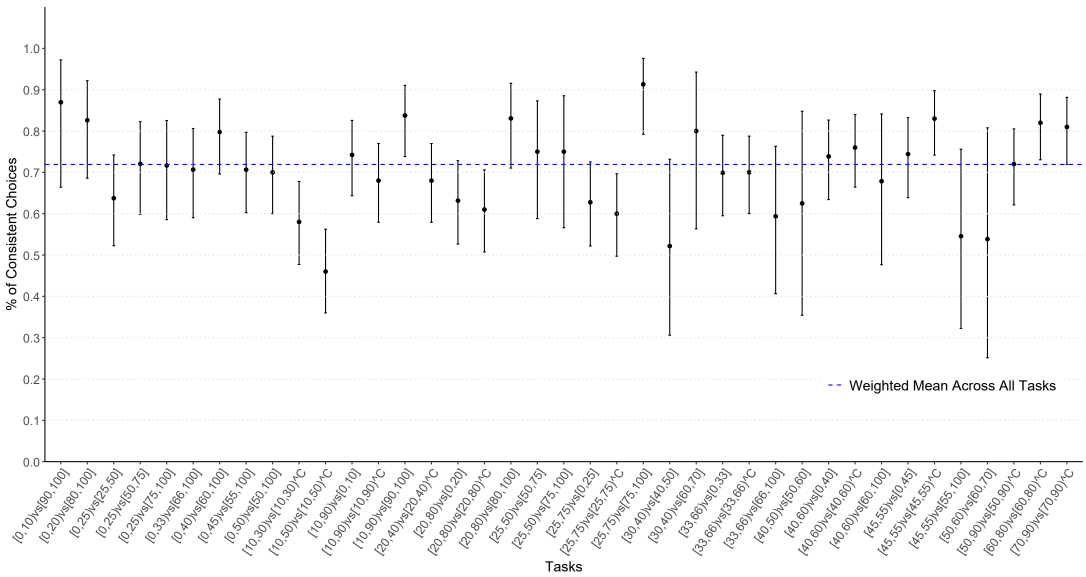


Figure 19: Proportion of Consistent Subjects in Experiment 2 for allocation 1 - Bet-Based Method

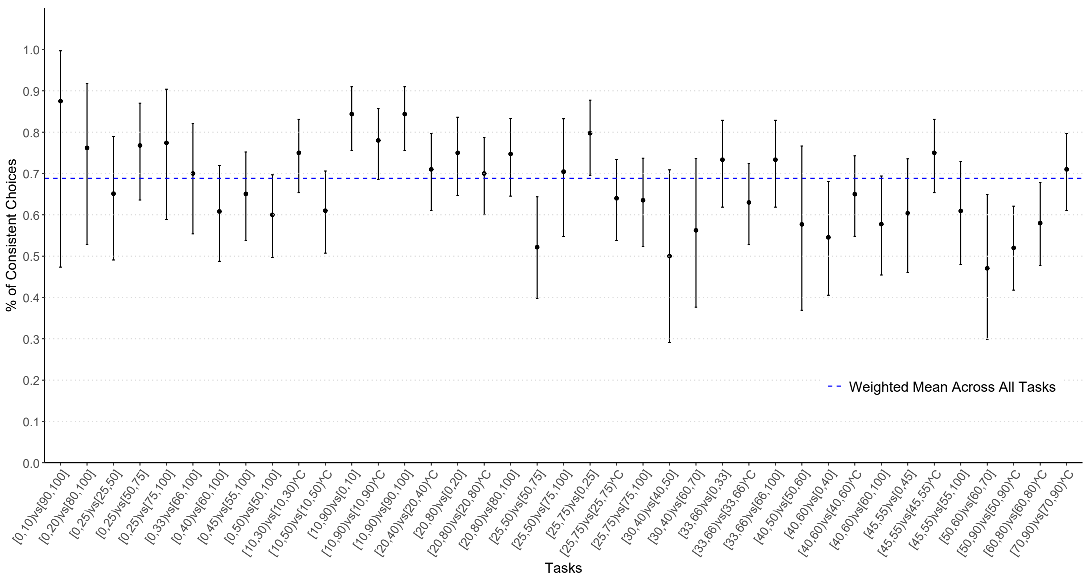


Figure 20: Proportion of Consistent Subjects in Experiment 2 for allocation 2 - Bet-Based Method

Parametric Results

Forest Plots

Forest plots for the remaining subjects are available upon request.

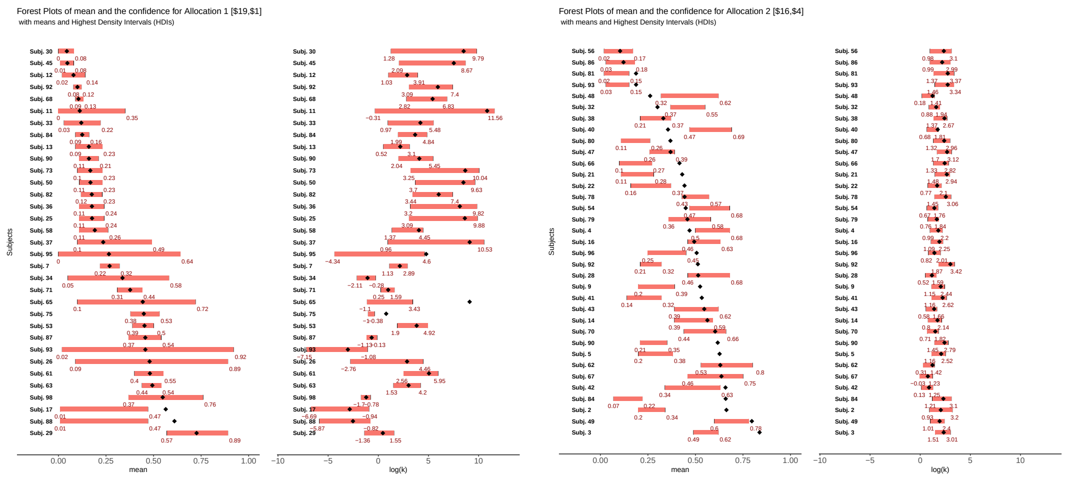


Figure 21: Forest plot: UG - Bet-Based Method

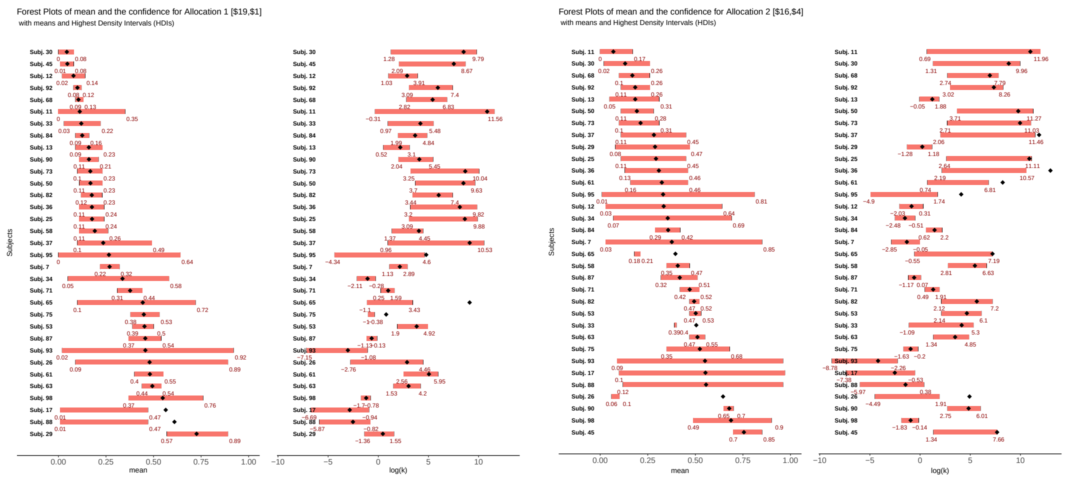


Figure 22: Forest plot: UG - Money Method under SEU

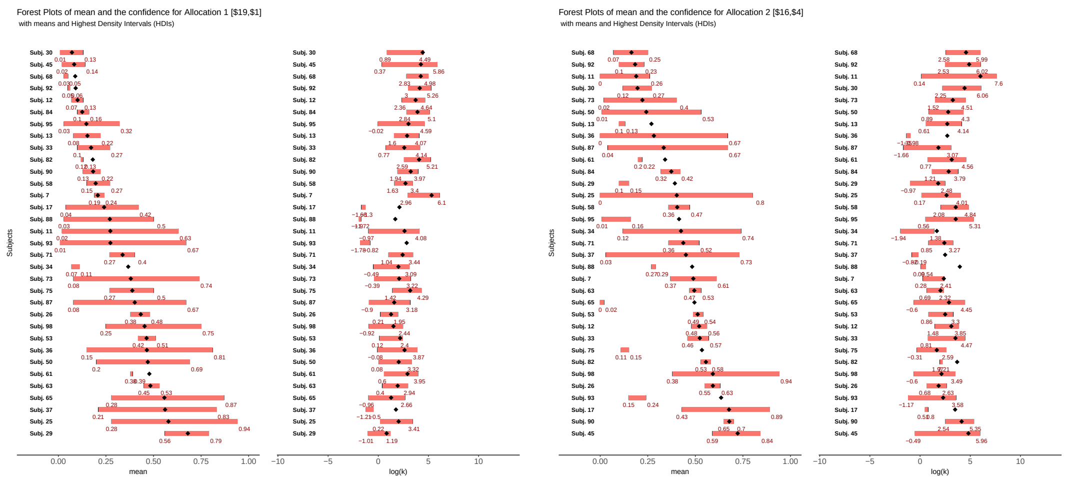


Figure 23: Forest plot: UG - Money Method under PW