Distributional Approach to Risk Preferences

Nir Chemaya
UCSB
Charles Johnson
UCSB
Brian Jabarian
Brian.Jabarian@chicagobooth.edu
University of Chicago https://orcid.org/0000-0001-8707-8596
Enoch Yeung
UCSB
Gary Charness
University of California, Santa Barbara

Article

Keywords:

Posted Date: June 11th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6787323/v1

Additional Declarations: There is **NO** Competing Interest.

Distributional Approach to Risk Preferences*

Nir Chemaya†
Charles Johnson‡
Brian Jabarian§
Enoch Yeung¶
Gary Charness||

Latest Version: May 27, 2025
Do not circulate this preliminary version without permission

Abstract

We propose a distributional framework for eliciting risk preferences that treats an individual's attitude toward risk as a full probability distribution rather than a point estimate. By parameterising preferences with the flexible beta family, our approach encompasses the entire spectrum from extreme risk aversion to risk neutrality and even risk-seeking behaviour, while simultaneously allowing for heterogeneous stability of those attitudes across contexts. Our agent-based simulations show that (i) the true underlying preference distribution is recoverable with negligible bias and (ii) the precision of recovery is a systematic function of elicitation design richness, providing clear guidance for experimental design. Benchmarking on the comprehensive laboratory dataset of Holzmeister Schmidt (2021) confirms two central results: (1) out-of-sample predictive accuracy is at least on par with canonical point-estimate methods, and (2) our method delivers a second, policy-relevant moment—the subject-specific variance of risk taking—without sacrificing parsimony.

*We thank Dan Friedman, Ryan Oprea, Daniel Martin and Kirby Nielsen, for their helpful comments and suggestions. The authors thank Helen Nesterenko, Laurence Liao and Dingning Yang for their excellent assistance with this research. Gary Charness passed away unexpectedly in May 2024. We dedicate this article to his memory.

†University of California, Santa-Barbara, nir@umail.ucsb.edu
‡University of California, Santa-Barbara, cajohnson@ucsb.edu
§University of Chicago Booth Business School, brian.jabarian@chicagobooth.edu
¶University of California, Santa-Barbara, eyeung@ucsb.edu
||University of California, Santa-Barbara

1 Introduction

Economists have long sought to elicit subjects' risk preferences using the revealed preferences approach (Beshears et al. (2008); Samuelson (1948)), which assumes that individuals' decisions reveal their true preferences. However, substantial evidence indicates that this process is fraught with complexity, often resulting in noisy and inaccurate measurements (Csermely and Rabas (2016); Friedman et al. (2014); Isaac and James (2000); Pedroni et al. (2017); Perez et al. (2021)). In this paper, we propose a new method for eliciting risk preferences. Departing from traditional point estimates, we argue that modeling risk preferences as distributions enhances the prediction of future decisions. Recognizing that individuals may commit errors or have preferences that are inadequately captured by point estimation, our approach integrates repeated measurements across tasks to capture risk preferences robustly.

Previous literature has attempted to distill individuals' risk preferences into single-point estimates, typically framed within constant relative risk aversion (CRRA) utility functions (Camerer and Ho (1994)). Under CRRA, the theory posits that each individual has a single risk preference parameter, denoted as r. Common methods for estimating r include multiple price lists (Holt and Laury (2002)), single choice lists (Eckel and Grossman (2002)), the Investment Game (Gneezy and Potters (1997)), the Certainty Equivalent Method (Cohen et al. (1987)), and the Bomb Risk Elicitation Task (Crosetto and Filippin (2013)). However, these methods frequently yield low correlations between tasks[1], a phenomenon known as the risk elicitation puzzle (Pedroni et al. (2017)), extensively documented in prior research (Crosetto and Filippin (2016); Holzmeister and Stefan (2021)).[2] It remains puzzling why risk elicitation methods—designed to measure consistent individual preferences—generate considerable inconsistencies. An explanation is task-dependent adjustments in subjects' risk preferences (Holzmeister and Stefan (2021)). However, these discrepancies, combined with evidence of low correlations between risk elicitation tasks and real-life decisions (Charness et al. (2020)), underscore the persistent challenges of accurately assessing individual risk preferences.

Several strategies have been proposed to improve predictive accuracy. First, ensuring task comprehensibility and implementing comprehension checks can enhance data reliability (Garagnani (2023)). Reliable elicitation is crucially dependent on a clear understanding of the tasks of the subjects (Charness et al. (2023)). However, when subjects repeat the same task several times and gain a better understanding, some agents still have variability in their decisions within the same task (Charness and Chemaya (2023)), which our distributional approach takes into account. Our approach allows agents to have risk preferences represented as a distribution rather than a single fixed value. Therefore, the variability in risky decisions can be captured within the distribution.

Second, response time recording has shown promise in improving the accuracy of out-of-sample prediction (Alós-Ferrer and Garagnani (2024)). This "time will tell" (TWT) method uses response time as an indicator of deviation from true risk preferences, where longer times suggest greater deviations. Although TWT demonstrates superior predictive performance, it is significantly dependent on repeated choices and reference options, constraining its broader applicability. Our distributional approach offers greater flexibility in estimating and predicting risky behavior while also providing insight into an agent's tendency to stick with or change

[1] Low correlations are observed both in raw decision data and in estimated r values.

[2] See also: Anderson and Mellor (2009), Berg et al. (2005), Bruner (2009), Charness et al. (2023), Dave et al. (2010), Deck et al. (2013), Deck et al. (2014), Drichoutis and Lusk (2016), Dulleck et al. (2015), Fausti and Gillespie (2000), Harbaugh et al. (2010), He et al. (2016), Ihli et al. (2016), Isaac and James (2000), Loomes and Pogrebna (2014), Menkhoff and Sakha (2017), Nielsen et al. (2013), Reynaud and Couture (2012), and Szrek et al. (2012).

their decisions.

The search for identifying "true" r remains elusive, lacking consensus on the appropriate metric to predict future decisions. The correlations between tasks weakly predict both other tasks and real-world behavior, highlighting the limitations of existing methods. A recent approach, the ORIV (obviously related instrumental variables) method (Gillen et al. (2019)), addresses measurement error, often overlooked as a key contributor to the puzzle, by requiring repeated task completions. ORIV reduces measurement error and produces adjusted r values with higher correlations between tasks. However, ORIV's effectiveness diminishes with substantially different tasks (Friedman et al. (2022)), and repeated measures may not always effectively separate signal from noise (Holden and Tilahun (2022)). In addition, ORIV is efficient at improving the correlation between pairs of risk elicitation tasks, but cannot be applied to larger datasets with many risky decision observations that could be used to better estimate and analyze such behavior.

Our approach provides an alternative perspective, suggesting that risk preferences are better estimated as distributions than as single-point estimates. Agents may exhibit variability in risk preferences across tasks due to the inherent flexibility in their decision-making, a phenomenon that point estimates often fail to capture. We offer a method to estimate risk preferences as a distribution, which is the main novelty of this paper; we do not answer the origin of the variability in risky decisions. Many economic models, such as stochastic choice or random-utility models, can justify this variability due to incomplete preferences. Although the origin of the agent's variability in risky decisions remains unclear, our method has a clear benefit, allowing us to conceptualize this variability through distributions of risk preferences and recognizing that some agents exhibit broad distributions indicative of fluctuating preferences. In contrast, others have narrow distributions similar to fixed-point estimates.

Our distributional model effectively reflects the risks and variability of agents, potentially providing a closer representation of their "true" preferences. This approach yields additional behavioral insights that are not available through traditional point estimation, notably by capturing the degree to which agents consistently adhere to or vary their risk decisions. Furthermore, unlike point estimation methods, our model provides a fully probabilistic distribution over an agent's entire decision set.

To operationalize this idea, we propose a model that captures agents' risk preferences as distributions using beta distributions[3]. Our paper presents a distinct and innovative approach that differs from the existing literature on using beta distributions in financial decision-making. Specifically, Johnson (1997), Libby and Novick (1982), and Parker and James (2024) employed the beta distribution. However, they primarily focus on utility modeling, portfolio management, and risk analysis. The authors used the beta distribution to model uncertainty in economic outcomes or asset returns rather than directly addressing individual risk preferences elicited through experiments.

Section 2 details the mathematical foundations of our model, introducing novel accuracy and confidence metrics for evaluating predictive power. Although accuracy metrics capture prediction deviations, confidence metrics uniquely assess how well the elicited distributions represent the variability inherent in risk preferences. Section 3 uses simulation analysis to demonstrate the model's ability to capture individuals' risk preferences with low error rates. However, its accuracy depends crucially on task design, specifically aligning task options with agents' variance in risk preferences. For agents exhibiting low-variance preferences, tasks

[3]Beta distributions have many applications across statistics, risk management, consumer behavior, and portfolio selection. For an introduction to these applications, see Gupta and Nadarajah (2004).

that provide numerous decision options (approaching continuity) optimize prediction accuracy with fewer task repetitions. In contrast, fewer decision options significantly enhance the accuracy of agents with high decision variance. In addition, increasing the number of task repetitions can improve accuracy, but it depends on the agent's risk distribution and whether repeating the task gives the model more useful information.

Finally, Section 4 evaluates our model using empirical data from Holzmeister and Stefan (2021) and benchmarks its performance against traditional point estimation techniques. The results show that our beta distribution framework maintains or exceeds the predictive power of point estimation methods, particularly in terms of confidence metrics. Crucially, this novel methodology captures agents' tendencies to consistently or variably engage in risk-taking behaviors, a dimension not observable through conventional methods. Incorporating additional data from repeated tasks further enhances our model's predictive performance, presenting a valuable and cost-effective tool for researchers and policymakers.

2 Risk elicitation methodology

Most of the research on risk preferences uses the expected utility model of Schoemaker (1982). We consider an agent whose underlying or "true" risk preference is represented by the parameter r. When faced with a decision involving uncertain outcomes (for example, lottery or risky investments), the agent maximizes expected utility under a Constant Relative Risk Aversion (CRRA) utility function:[4]

$$u(x) = \begin{cases} \ln(x) & \text{if } r = 1, \\ \frac{x^{1-r}}{1-r} & \text{otherwise,} \end{cases}$$

where x denotes the payoff. The parameter r indicates the individual's risk attitude as follows:

$$r = \begin{cases} \text{Risk averse,} & \text{if } r > 0, \\ \text{Risk neutral,} & \text{if } r = 0, \\ \text{Risk seeking,} & \text{if } r < 0. \end{cases}$$

In a more general setting, with fewer constraints, the risk preference parameter r can be drawn from a probability distribution. This representation allows for flexible preferences, ranging from constant preferences (points) (where $r = c$, a fixed constant) to maximally variable preferences (a uniform distribution).

2.1 Point Estimation of Risk Preferences

One prevalent method for quantifying risk preferences is point estimation, where elicitation tasks assign each individual a specific numeric risk value that represents their risk preference. Ideally, to achieve precise noise-free elicitation, individuals should select from a continuum of risk options, allowing direct identification of their exact r. Under ideal conditions, an individual's chosen option directly reveals their true r precisely.

However, in practice, most of the elicitation tasks offer only discrete and finite decision sets, implying that the risk preferences elicited are typically intervals rather than precise values. For

[4] We adopt the CRRA utility function as our baseline due to its widespread use and suitability for experimental economics (Camerer and Ho (1994); Holzmeister and Stefan (2021); Wakker (2008)).

example, in the single choice list (SCL) method, selecting the second lottery corresponds to risk preferences within the interval $r_1 \triangleq 1.16 \leq r \leq 3.46 \triangleq r_2$ (Charness et al. (2023)).

Thus, the risk elicitation task can be understood as a mapping function from an individual's decision (the interval $[r_1, r_2]$) the lottery chosen in the risk elicitation task, to a specific interval. Frequently, researchers assign the midpoint of this interval to the subject as a point estimate of their risk preference.[5]

In this case, if the agent indeed has risk preferences that can be represented as a point estimate, then two sources can generate an error in the elicitation. The first source is when the mapping function from the decision assigns, by structure, a different numeric value than the agent's true preference. The second source of error could be any potential error term $e_{i,t}$, which captures noise or task-specific measurement error.

Point estimation simplifies the agent's decision to a single variable instead of providing an interval (lower and upper bounds). This method often leads to an overfitting problem, where estimating risk preferences based on a single task results in poor predictions on other tasks (the risk elicitation puzzle). Moving to a distributional method can reduce the overfitting problem by using more information, incorporating lower and upper bounds of the interval from each decision, and combining data from different tasks.

2.2 Risk Preferences as Distributions

Alternatively, we propose that individuals' risk preferences may be inherently distributional rather than fixed point estimates. Under this assumption, an agent may have a range of acceptable risk preferences or exhibit variability in their choices, which is better represented by a probability distribution. For example, we could assume that the risk preference of a subject is normally distributed with a mean of μ and a variance of σ^2:

$$r \sim \mathcal{N}(\mu, \sigma^2).$$

However, the specific form of the underlying distribution does not need to be normal; subjects may have various other distributions (e.g., skewed, uniform, multimodal). The task of eliciting risk under this framework aims at estimating an individual's risk preference distribution.

Several interpretations exist regarding the economic meaning of risk preference distributions. One perspective involves random utility models, where the agent randomly selects a utility function (from a set determined by the values of r) when making a decision. Another interpretation views the distribution as representing decision-making noise, whereby subjects deviate from their true optimal decision due to stochastic error.[6] Alternatively, subjects may have inherently incomplete or less structured preferences that the utility model cannot fully capture.

Our analysis does not advocate one interpretation over the others and the origin of the variability in risky decisions. Instead, it highlights the value added by eliciting risk preferences as distributions, particularly in terms of predictive accuracy and confidence in subsequent decisions.

[5] When subjects choose extreme options (e.g., first or last options), the interval is typically open-ended, and the lower or upper bound is used instead to prevent unrealistically extreme parameter values.

[6] In this case, an alternative perspective is that risk preferences can be captured as a point estimate combined with an individual error term with a specific distributional shape. In that case, our model can capture the individual error distribution.

2.3 Modeling Risk Preferences with the Beta Distribution

Modeling risk preferences as probability distributions naturally prompts a critical methodological question. Which distribution should we select? We propose that an ideal distribution for capturing risk preferences should satisfy four criteria: (1) finite support, (2) continuity, (3) limited parameter complexity, and (4) asymmetry.

Finite support ensures that the risk preferences elicited lie within realistic upper and lower limits determined by experimental designs. Although the r values in our model have no restrictions and can range from negative to positive infinity, there is clear evidence that agents' risk decisions typically fall within a specific range of r parameters across many different tasks and experiments (Crosetto and Filippin (2016)). During experimental design, researchers can set the upper and lower bounds accordingly based on historical data.

A continuous distribution is appropriate because risk preferences can plausibly vary smoothly. Choosing a low-parameter distribution allows for efficient estimation even from relatively few elicitation tasks.

To see why asymmetry is important, consider the following thought experiment. Imagine that a symmetric distribution was used to model the preferences of an extreme risk taker with moderate to low variance in their preferences. Assuming that, at the time of elicitation, the maximum level of risk postnormalization for which we can test is 1, then one of three things would need to happen. Either a significant portion of the distribution's probability would have to be located outside of the acceptable interval, a significant portion of the probability would have to be associated with low-risk behavior, or the distribution would need to have extremely low variance. In the first case, the symmetric distribution attributes a significant probability to levels of riskiness that are beyond maximal. In the second case, the distribution fails to classify the subject as risk-loving on average. In the third case, the distribution fails to classify the subject as a variable in their preferences. An asymmetric distribution, in such a scenario, could successfully model the individual as risk-loving on average, variable in their preferences, and all the model probabilities would be within the established boundaries.

The Beta distribution, characterized by two shape parameters α and β, is a natural candidate that meets all three criteria. Defining the interval $[0, 1]$, the Beta distribution captures a wide range of preference behaviors and is widely used to model uncertain outcomes with limited experimental data. It also conveniently covers three important extremes of preference variability.

1. Complete consistency across tasks (minimal entropy), resembling a point estimate.

2. Normally distributed risk preference, because normal distributions can approximate beta distributions whenever the beta distribution is not "U" or "J" shaped Peizer and Pratt (1968). This implies that normal distributions that have been scaled so that their probability lies almost entirely in the interval $[0, 1]$ can be well approximated by beta distributions.

3. Uniform random choice (maximal entropy), reflecting maximal uncertainty.

Given the advantages of the beta distribution, we assume that the scaled individual risk preference r is Beta distributed:

$$r \sim \text{Beta}(\alpha, \beta),$$

where the parameters α and β govern the shape of the distribution, including its mean and variance. Thus, given a particular interval $[r_1, r_2] \subseteq [0, 1]$, the probability that the individual's

risk behavior falls within this interval can be computed directly from the Beta distribution's cumulative distribution function (c.d.f.), $F_r(\alpha, \beta, r)$:

$$P(r_1 \leq r \leq r_2) = F_r(\alpha, \beta, r_2) - F_r(\alpha, \beta, r_1).$$

This flexible modeling approach provides richer behavioral insights, allowing for more accurate and nuanced predictions than traditional point estimation methods.

2.4 Using risk elicitation tasks to specify a distribution

We now address the question of how to use the results of a risk elicitation task to specify a beta distribution that models an individual's risk preferences. First, the risk intervals for the task must be scaled to fall between values of 0 and 1. The experiment designer must determine the upper and lower boundaries and the scale of the data accordingly. [7]

Then, since each scaled risk elicitation task results in an interval of possible risk preference scores $[r_1, r_2]$, each elicitation task returns two usable pieces of information. This information can be expressed in several ways. Typically, the results of a single task are represented as the mean of their risk preference score

$$\frac{r_1 + r_2}{2}.$$

However, expressing the interval as a single real number wastes information. Our solution is to interpret the interval returned by each elicitation task as a 2-parameter distribution over possible risk scores.

Given a risk preference interval, $[r_0, r_1], r_0 < r_1$, we interpret the results of the task to mean that any risk preference in the return interval, $r^* \in [r_0, r_1]$, is equally likely. This means that we interpret the result of our risk elicitation task as a uniform distribution with support at the interval $[r_0, r_1]$[8]. The final beta distribution that we parameterize to model the individual's risk preference is then computed as the closest beta distribution (measured by maximum likelihood estimation) to the sum of all the uniform distributions returned by the elicitation tasks.

Now, there is a countably infinite set of discrete uniform distributions in the interval $[r_0, r_1]$, as that interval (assuming that $r_0 \neq r_1$) may be evenly divided into N subintervals for any positive integer N. Assuming that the midpoint of the interval, $\frac{r_0+r_1}{2}$, is the true risk preference of the individual that corresponds to choosing $N = 1$. Using the two extreme values of the interval corresponds to choosing $N = 2$. Representing the interval as endpoints and the midpoint (3 points total) corresponds to choosing $N = 3$. The number of subdivisions may be chosen to be arbitrarily high. In the limit of an infinite number of subdivisions, the resulting uniform distribution approaches the continuous uniform distribution with support on the interval $[r_0, r_1]$.

The interval length, $r_1 - r_0$, may not be the same between tasks or even within tasks. Since the intervals returned from the elicitation tasks are not uniform in length, the modeler needs to

[7] This can be done in a large dataset by using the minimum and maximum (r_i) values and scaling the data accordingly, as we do in our dataset.

[8] This interpretation assumes that agents' risk preferences follow a uniform distribution over the interval. We adopted this methodological approach due to its simplicity and the lack of precise knowledge about the true distribution of agents' risk preferences. As an alternative, we also implemented a maximum likelihood estimation (MLE) approach by interpreting the interval as interval-censored data. This alternative yielded significantly worse results across all our simulation analyses in Section 3.2 except for a subset of extremely high and low risk agents. Additionally, this alternative yielded similar results for the human subject data in Section 4. Therefore, we chose to retain the current methodology. However, as discussed in Section 5, we believe that future research could develop more effective ways to interpret interval data and improve the elicitation process.

make some choices. If all the intervals returned from our elicitation tasks were of uniform size, the only problem would be to choose the number of subintervals to split the resulting interval into. But, the interval lengths differ, so we need to decide if it is more appropriate to:

1. choose one number of subintervals, N, for all tasks or
2. choose a fixed subinterval length, l, and then give each interval a number of subintervals so that their length is as close to l as possible.

The consequence of choosing a fixed subinterval length, l, and letting the number of subintervals vary according to the interval length is that larger intervals would contribute more data points to the fitting of the beta distribution, which will capture the summative effect of the uniform distributions returned by each elicitation task. The consequence of choosing a number of subintervals, N, for all tasks would be that smaller intervals will have a relatively heavier weight in learning the mean of the final beta distribution. If we assume that each risk elicitation task accurately captures true risk preferences, then smaller intervals are preferable to build an accurate model. This is because (in expectation) samples from a discrete uniform distribution over a small interval will be closer to each other (internally precise) than samples from a discrete uniform distribution over a large interval. So, we choose to have a number of subintervals, N, for each task to build our distribution models of individual risk preferences.

This naturally poses a final problem before building models from the results of risk elicitation tasks: how many subintervals should we use, what value do we choose for N ? As discussed above, $N = 1$ is too small as it results in a loss of information returned by the elicitation task. In theory, as $N \to \infty$, the resulting distribution will approach a continuous uniform distribution defined over the given interval, $[r_0, r_1]$. No matter the choice of N, the mean of the resulting uniform distribution will remain constant:

$$\mu(r_0, r_1) = \frac{r_0 + r_1}{2}. \tag{1}$$

However, the variance of the resulting distribution will monotonically decrease as a function of N:

$$\begin{aligned} Var_{Discrete}(r_0, r_1, N) &= \frac{N^2 - 1}{12(N-1)^2}(r_1 - r_0)^2 \\ Var_{Continuous}(r_0, r_1) &= \frac{1}{12}(r_1 - r_0)^2. \end{aligned} \tag{2}$$

So, when $N = 2$, the variance will be $\frac{1}{4}$ the square of the interval length, and as $N \to \infty$, the variance will approach $\frac{1}{12}$ the square of the interval length. Which value for N should be chosen? From running experimental results with different values of N and choosing various numbers of samples from the continuous uniform distribution, we found that the choice of N made little difference in the resulting distribution performance (see the section below for definitions of performance). Given that similar results came from all tested values of $N > 1$, we use $N = 2$ as it takes the least computing power.

2.5 Measuring the success of a model of risk preference

How can we score model success after fitting a beta distribution to a subject's risk preference? Our methodology uses the resulting distribution as a predictive model. So, to score the model, we predict the outcomes of other risk elicitation tasks. If one understands an individual's risk

preferences, then one should be able to predict how that individual will evaluate risk preferences.

To score the success of a beta distribution in modeling the risk preferences of an individual, we choose to evaluate two things.

1. how accurately the model predicts the outcome of a risk elicitation task and
2. how surprised the model is when presented with the outcome of a risk elicitation task.

If the model predicts that the task's risk preference interval, $[r_0, r_1]$, was the most likely interval of its size, then the model has perfectly predicted the outcome with zero error and zero surprises. However, if the model assigns zero probability to the task's risk preference interval, regardless of its accuracy, we would understand that the model would be totally surprised.

So, how do we compare the model predictions with the true intervals learned from the risk elicitation task, especially when different tasks yield risk preference intervals of different sizes? If the true interval is $[r_0, r_1]$, then it has a length of $r_1 - r_0 \leq 1$. With a few exceptions, each beta distribution will have a single, most probable, interval of length $r_1 - r_0$, call it $[r_0^*, r_1^*]$. The interval $[r_0^*, r_1^*]$ can be considered as the model's prediction of the outcome of the risk elicitation task. In this case, accuracy is simply the distance between the predicted and the true interval. So we have

$$\text{Accuracy}(r_0, r_1, r_0^*, r_1^*) = r_0^* - r_0 = r_1^* - r_1. \tag{3}$$

On the other hand, computing the model's surprise or confidence in its predictions is slightly more nuanced.

An ideal metric for model surprise/confidence would be

1. scale-free,
2. agnostic to the prediction task, and
3. easily interpretable.

Our solution is to take a ratio of two probabilities: the probability that the model assigns to the most likely outcome (its predicted interval) and the probability that the model assigns to what actually was measured.

Each risk elicitation task divides the spectrum of possible risk preferences into intervals. The size of these intervals differs from task to task. However, given any interval, $[r_0, r_1]$, the CDF of the learned Beta distribution, $F_r(\alpha, \beta, r_i)$, can be used to calculate the probability that the true risk preference score will be in that interval. This means that given an interval width (that is, given a risk elicitation task), there is a unique interval of that width, $[r_0^*, r_1^*]$, with the highest probability. So, to score our Beta distribution's model success we can take the ratio of our highest probability interval to the other interval of the same length, specifically:

$$\text{Confidence}(\alpha, \beta, r_0, r_1, r_0^*, r_1^*) = \frac{F_r(\alpha, \beta, r_1) - F_r(\alpha, \beta, r_0)}{F_r(\alpha, \beta, r_1^*) - F_r(\alpha, \beta, r_0^*)}. \tag{4}$$

This means that if $[r_0, r_1] = [r_0^*, r_1^*]$, then $\text{Confidence}(\alpha, \beta, r_0, r_1) = 1$, on the other hand if the model considers the true risk preference value to have zero probability to be in the interval $[r_0, r_1]$, then $\text{Confidence}(\alpha, \beta, r_0, r_1) = 0$. This confidence or surprise score can be interpreted as a percentage of the model's prediction's optimality, where a value of 1 is 100% optimal and a

value of 0 is 0% optimal. The plots in Figure 1 are schematics designed to visually represent and interpret accuracy and confidence scores, as well as to show how intervals from risk elicitation tasks inform a beta distribution model of risk preference.

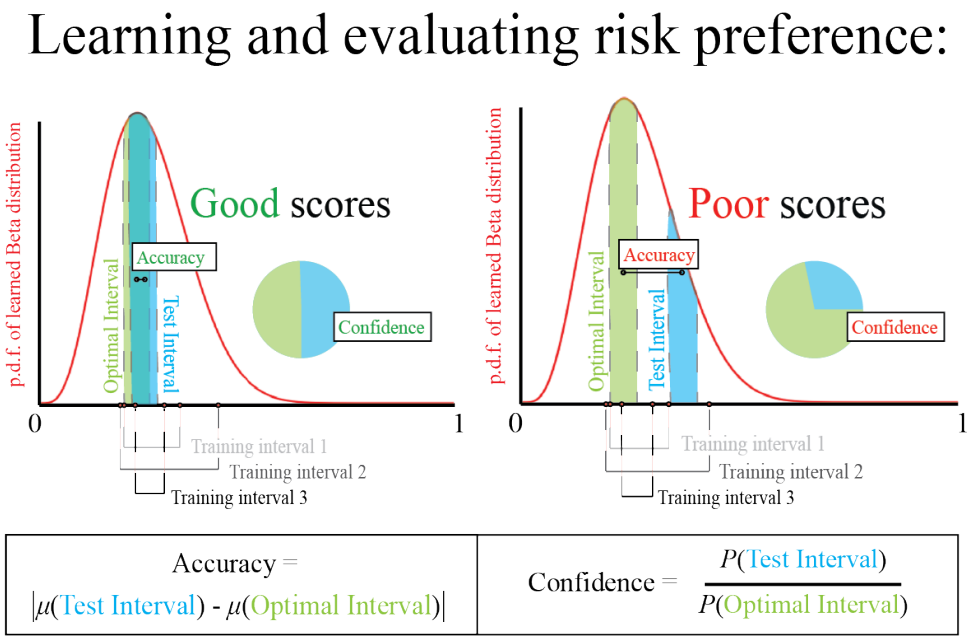

Figure 1: An illustration of the Beta distribution model trained on the risk preference intervals from three tasks and scored for prediction optimality on the fourth task. The result on the left would be a nearly perfect score, and the result on the right would be much lower.

3 Results from simulated agents

To examine the impact of the quantity and precision of risk elicitation tasks on learning the true risk preference distribution, we conduct a series of agent-based simulations. In these simulations, we know the risk preference distribution of each simulated agent. This allows us to effectively test our model from Section 2.4.

We focus on two types of simulations. The first type, in Section 3.1, examines four prototypical agents to determine which tasks are more (or less) effective in capturing the true distribution. In Section 3.2, we expand our analysis to the four risk elicitation tasks of Holzmeister and Stefan (2021), testing our model on an array of agents with different parameters to better understand the relationship between model accuracy, agent distributions, and task design.

3.1 Four prototypical agents

In this section, we simulate the modeling process on four prototypical agents. Each agent has an internal beta distribution with specified parameters. A sample of the agent distribution is taken for n repetitions of a risk elicitation task. Each sample will naturally fall into one of the m evenly spaced subintervals of $[0, 1]$, the support of the beta distribution. The collection of all n of these returned subintervals of size $\frac{1}{m}$ is then used to fit a model of the true distribution of the agent. The average pointwise error between the true and predicted distributions is used to compute the percentage accuracy of the learned model parameters. The results highlight the

efficacy of our modeling process and provide insights into best practices for building accurate models of the preference of the subject for risk.

The four agents align with the four data groups shown in Table 3, which were calculated based on human decisions from Holzmeister and Stefan (2021). Each sample their risk preferences from a beta distribution with distinct parameters. These parameters are as follows:

1. Agent 1: Uniform risk preferences has ($\alpha = 1$, $\beta = 1$)
2. Agent 2: Moderate variance in risk preferences (symmetric) has ($\alpha = 25$,$\beta = 25$)
3. Agent 3: Extreme low variance in risk preferences has ($\alpha = 625$,$\beta = 625$)
4. Agent 4: The moderate variance in the risk preferences (asymmetric) has ($\alpha = 5$,$\beta = 25$)

Figure 2 contains a visual comparison of all four agent probability density functions. The risk preference of Agent 1 is sampled from a uniform distribution. This is possible because the beta distribution generalizes both the uniform distribution and (in the limit of infinite parameters) the Dirac delta distribution. This aligns with the fourth row in Table 3. Agents 2 and 4 sample from beta distributions that qualitatively match the second and third rows in Table 3. The difference between the two is that Agent 2 is centered at 0.5, the midpoint of the interval, and Agent 4 is off-center. Agent 3 samples from a low-variance distribution and aligns with the subjects in the first row of Table 3.

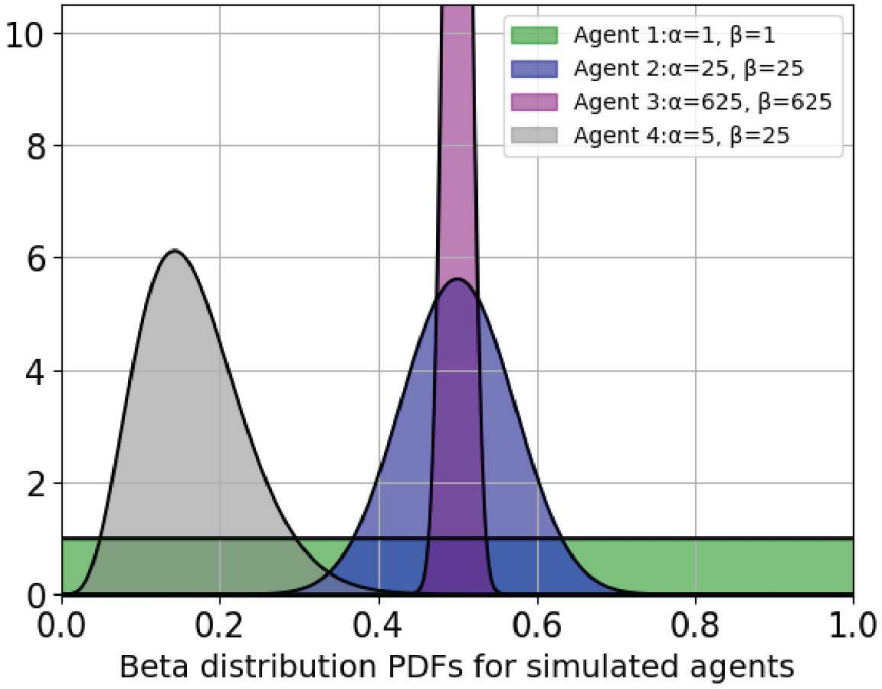

Figure 2: Plots of the probability distribution functions for all four simulated agents.

We considered 1 to 20 repetitions of a risk elicitation task as well as a non-exhaustive selection of 21 to 1280 repetitions in our agent simulations. We also considered 1 to 20 interval subdivisions as well as 24 and 99 subdivisions. We performed 1,000 modeling simulations for each agent, for each of the 400 combinations of risk preference elicitation tasks and intervals.

The results of the 1000 modeling simulations were averaged to draw inferences on the learning problem.

To analyze these results, we considered cross sections of the results where the number of repetitions of each task or the number of intervals (precision level) of each task was held constant. Each of these cross sections relates the model error as a dependent variable to either the number of intervals or the number of task repetitions as the independent variable. We fit a three-parameter exponential curve to the mean model error across all 1,000 simulations of each combination of the number of repetitions and the number of variables. The model is as follows:

$$y = ae^{-dx} + c, \tag{5}$$

where y is the mean model error and x is either the number of intervals or task repetitions (depending on which we hold constant and which we allow to vary). The parameter d is the decay rate of the exponential curve. The parameter c sets the theoretical floor for the model error because one of the independent variables is kept constant. For example, when a task has only four intervals, the model error will never reach zero due to the task's lack of precision; this will even occur in the limit of an infinite number of repetitions of the risk elicitation task. Finally, the parameter a scales the exponential function to fit the error metric.[9]

3.1.1 Agent 1: Uniform risk preferences

When an agent makes a decision, our model already assumes that this decision corresponds to an interval of r, $[r_0, r_1]$, where the value r is uniformly distributed within that interval. Therefore, for Agent 1, the optimal way to model the agent's distribution is through a task with fewer options. For example, in the extreme case where the task has only one option covering the entire range $[0, 1]$, the model structurally assumes that the agent has a uniform distribution. This intuition helps to explain the results shown in Figure 8, where the model's ability to capture random preferences accurately decreases as the number of options (intervals) in the risk-elicitation task increases. Finally, repetition can help capture random preferences; however, it is highly sensitive to the task design, specifically the number of intervals. Tasks with four intervals can achieve the same expected accuracy with an order of magnitude fewer repetitions than tasks with 99 intervals, as shown in Table 2.

3.1.2 Agent 2: Moderate variance in risk preferences, symmetric

Agent 2 can be seen as a middle ground between a random preference with high variability (Agent 1) and a nearly point-estimated preference with very low variability (Agent 3). For Agent 2 with an increase in both the number of task repetitions and the number of task intervals (up to 20) in each task, the model error decreases exponentially; see Figure 9.

However, repetitions can only improve the model accuracy if the number of intervals in the task is large enough to capture the agent's low variability. For example, Table 2 shows that with a small number of intervals, such as four options, the model cannot accurately recover the true distribution of the agent, even with thousands of repetitions. The main idea here is that by design, the Beta model interprets each chosen interval as a uniform distribution over its range.

[9] These curves were fit using the Levenberg-Marquardt algorithm, implemented in Python using the `scipy.optimize.least_squares()` function from the scipy package. The following are some key takeaways from the results for each agent.

When intervals are wide, this structural assumption leads the model to infer a higher variance than the true distribution of the agent.

In contrast, using a very large number of intervals, such as 99, requires many more repetitions to accurately capture Agent 2's true preferences compared to a task with 9 intervals. This is because the true preferences, while still low in variability, are now distributed across a much finer-grained set of options, making each interval less informative on its own.

3.1.3 Agent 3: Extremely low variance in risk preferences

Agent 3 qualitatively approximates an extreme representative of the first group of subjects in Table 3. The variance of this distribution is 0.0002, compared to 0.0005, the variance of the average learned distribution in the first row of Table 3. Since Agent 3 is so consistent from task to task, increasing the number of task repetitions shows little to no effect on prediction error (see Figure 10). As Figure 10 shows, increasing the intervals from 9 to 19 will improve the accuracy of the model by 45%. This illustrates the trend that taking more fine-grained (precise) risk-elicitation tasks provides an exponential improvement in model accuracy for this agent. [10] Given the low variability in Agent 3's distribution, the optimal way to accurately capture preferences that are nearly point estimates is to design a risk elicitation task with many intervals. This enables the model to estimate the precise r value of the agent with minimal variation.

3.1.4 Agent 4: Moderate variance in risk preferences, asymmetric

Agent 4 and Agent 2 exhibit very similar variability in their distributions. The main difference between them is that Agent 2 is centered at 0.5 while Agent 4 is off center. The simulation results for Agent 4 closely resemble those of Agent 2, see Figure 11 and Table 2. This suggests that *when the risk elicitation task uses equal intervals*, the mean of the distribution has little impact on the model's ability to recover the underlying risk preference distribution. [11] Instead, the main challenge arises from the variance of the distribution and from finding the optimal task design to capture this variability.

An overall takeaway from the four agent simulations is that when the task intervals are too large or too small, the learning results are sub-optimal.

3.1.5 Optimal interval endpoints for predicting the true distribution of Agents 1, 2, 3 and 4

The simulation results suggest that for each distribution of risk preferences, there exists a task design that can optimally enable the model to learn the true distribution with a low number of repetitions. In this section, we mathematically approximate this optimal design.

[10] We additionally note a marked difference between having an even and odd number of evenly-spaced intervals in the elicitation task for this agent specifically, see Figure 10. Since the agent's mean is zero, having an even number of intervals resulted in a significantly better model, as it placed overlapping endpoints of the elicitation intervals exactly on the true mean. Since we used a discrete approximation of the uniform distribution to make our calculations, this overlap resulted in a final distribution with a more accurate mean. This effect is an artifact of the true distribution that aligns exactly with the boundary between two risk intervals, and it is only marked in rare subjects with extreme consistency in risk behavior. Therefore, this effect will likely not occur outside of a simulation.

[11] However, as Section 3.2 will show, when the intervals are unequal, this may change.

Agent	True parameters	Agent var.	Optimal int.	Int. length	Num. int.
1	$\alpha = 1, \beta = 1$	0.0833	$[0.2113, 0.7887]$	0.57735	1.732
2	$\alpha = 25, \beta = 25$	0.0049	$[0.4300, 0.5700]$	0.14002	7.141
3	$\alpha = 625, \beta = 625$	0.0002	$[0.4859, 0.5141]$	0.02827	35.37
4	$\alpha = 5, \beta = 25$	0.0045	$[0.0997, 0.2336]$	0.13387	7.470

Table 1: Optimal interval endpoints for predicting the true distribution of Agents 1, 2, 3 and 4 using a single interval and the method of moments to approximate the true parameter values. The final column, **Num. int.** gives the number of intervals of the specified length that fit in $[0, 1]$, the support of the beta distribution. Note that higher variance agents are optimally predicted in intervals divided into a larger number of evenly-spaced subintervals. Given the optimal interval, each agent's parameters are predicted with over 99.99% accuracy.

Using the method of moments, we can reverse-engineer a single risk-preference interval, which would result in the exact learning of the true distribution parameters for each of our agents. This can be done as follows.

Given a risk interval, $[r_1, r_2]$, where r_1 and r_2 are taken from the interval $[0, 1]$ and $r_1 < r_2$ we assume the mean of the sample and the unbiased variance of the sample,

$$\mu_{sample} = \frac{r_1 + r_2}{2};\ \sigma^2_{sample} = (r_1 - \mu_{sample})^2 + (r_2 - \mu_{sample})^2, \tag{6}$$

of the interval endpoints are the mean and variance of the true distribution. Equivalently, we assume that we have only one task interval to infer the true distribution. Then, via the method of moments, we approximate the true distribution parameters to be

$$\alpha_{approx} = \left(\frac{\mu_{sample}(1 - \mu_{sample})}{\sigma^2_{sample}} - 1 \right) (\mu_{sample}), \text{ and} \tag{7}$$

$$\beta_{approx} = \left(\frac{\mu_{sample}(1 - \mu_{sample})}{\sigma^2_{sample}} - 1 \right) (1 - \mu_{sample}). \tag{8}$$

This means that the percent error of the approximate parameters to the true parameters, α_{true} and β_{true} will be

$$\text{Percent error}(r_1, r_2) = 100 \left(1 - \sqrt{\frac{(\alpha_{approx}(r_1, r_2) - \alpha_{true})^2 + (\beta_{approx}(r_1, r_2) - \beta_{true})^2}{\alpha^2_{true} + \beta^2_{true}}} \right). \tag{9}$$

We then maximize the percent error[12] for each agent to determine the ideal, single risk-preference interval for discovering the agent's true distribution parameters. The results of this are summarized in Table 1. Ultimately, we found that the ideal risk interval width is smaller for agents with lower variance and larger for those with higher variance. This suggests that tasks with fewer intervals are better suited for learning the true parameters of high-variance subjects, while tasks with more intervals are best for high-variance subjects.

[12] Using the Limited-menory Broyden-Fletcher-Goldfarb-Shanno-B algorithm, implemented in Python using the `scipy.optimize.minimize()` function from the scipy package.

3.1.6 Main conclusions

We found that the process of representing intervals from tasks as uniform distributions and summing those distributions to fit a beta distribution using MLE enables an accurate reconstruction of the original beta distribution used to sample from the elicitation tasks. Table 1 shows that for each type of simulated agent, we can find the optimal task type (with different intervals, meaning different numbers of decisions) to achieve the highest accuracy in the model's learning of the true distribution with only one task.

The main takeaway is that prediction accuracy depends on how well the decision space aligns with the agent's risk distribution. Random agents (such as Agent 1) are more easily predicted when the task involves a small number of options. In contrast, agents with low-variance risk preference distributions (such as Agent 3) are better predicted for tasks with many choices. In the extreme case of constant (zero-variance) preferences, a continuous risk elicitation task is necessary. Agents with distributions between these two extremes (like Agents 2 and 4) are best captured by tasks with a moderate number of risky options that reflect their specific variability.

In addition, repeating the same task can improve the model's ability to capture an agent's risk distribution. However, this effect also depends on the nature of the agent's risk preferences and whether the specific task design enables the model to learn more from repetition for that agent.

However, in experimental settings, we typically do not know each subject's risk preferences or their variability in advance, which is actually what we seek to reveal from their decisions. Table 2 suggests that the conservative way to reveal preferences as a distribution in our model with high accuracy will be using a risk elicitation task with many options (intervals) and having many repetitions of the task to capture any agent-type preferences, including random ones. However, this could be very costly and, as Table 2 suggests, a task design with 99 options (intervals) requires repeating the task over 29 repetitions to have at least 80% accuracy for all four agent types. An alternative approach is to develop a combination of tasks with different intervals (that is, sets of risky options) that can approximate cover variability across all types of agents with a lower cost. We explore this approach in Section 3.2.

Agent	# intervals	80%	90%	95%	99%
Agent 1	4	9	15	26	69
(Maximal var)	9	17	42	101	612
$\alpha = 1$	24	25	55	164	98.6% at 1226
$\beta = 1$	99	27	67	181	98.4% at 1214
Agent 2	4	29.8% at 1	-	-	-
(Moderate var)	9	2	2	92.4% at 2	-
$\alpha = 25$	24	11	14	18	25
$\beta = 25$	99	18	28	55	152
Agent 3	4	1.2% at 1	-	-	-
(Extreme low var)	9	6.5% at 20	-	-	-
$\alpha = 625$	24	45.9% at 1	-	-	-
$\beta = 625$	99	8	12	14	16
Agent 4	4	23.1% at 1	-	-	-
(Moderate var)	9	2	2	94.4% at 2	-
$\alpha = 5$	24	9	12	14	18
$\beta = 25$	99	18	26	49	126

Table 2: The minimal number of task repetitions required to average the specified percent parameter accuracy over 1000 experiments given a specified number of intervals in the task. Entries in gray have failed to reach the specified accuracy and give the maximum average parameter percent accuracy as well as the lowest number of task repetitions needed to reach it. Note that we tested a maximum of 1280 task repetitions, and our coverage of the search space was not exhaustive. Additionally, these values are not monotonic; higher task repetitions can lead to lower average accuracy in parameter prediction.

3.2 Continuous Set of Agents

In this Section, we want to test how well a combination of different risk elicitation tasks can capture risk preference as a distribution in cases where the sample includes agents with various distributions. We specifically focus on the risk elicitation tasks from Holzmeister and Stefan (2021) which are well established in the literature and for which we also have empirical data to later test our model.

To approximate the range of accurate distribution learning for the specific combination of the BRET, CEM, MPL, and SCL tasks from Holzmeister and Stefan (2021)[13], we also simulate 3,481 different agents with unique internal beta distributions. The intervals from which these agents sample match those of the four tasks from Holzmeister and Stefan (2021). We run and average the learned parameters for 1000 experiments for each agent. The agents' parameters α

[13]In these tasks, unlike in Section 3.1, the task intervals are not split equally over the support of the beta distribution. [0,1].

and β range from 1 to 650, so the variance and mean range throughout nearly the entire allowed spectrum for a beta distribution (see Figure 3).

When we map the percent accuracy of the average results for each agent versus the choice of parameters, the results indicate a nuanced relationship between the two parameters and the percent accuracy of the average prediction (see Figure 12). After converting the results to instead map the mean and variance (see Figure 3) of the resulting distribution to the percent accuracy, we learn two things:

1. The results depend primarily on the variance of the agent distribution. It is harder for the model to capture extreme low-variance distributions with just these 4 tasks.

2. For distributions with a mean around 0.5, the model has higher predictive power even when the variance is low. This is mainly due to the fact that the intervals are not evenly distributed across the support of the beta distribution in these tasks. As a result, the model performs better on distributions whose means are covered by more lottery options around the center (0.5) compared to extreme values that are covered by only a few task intervals.

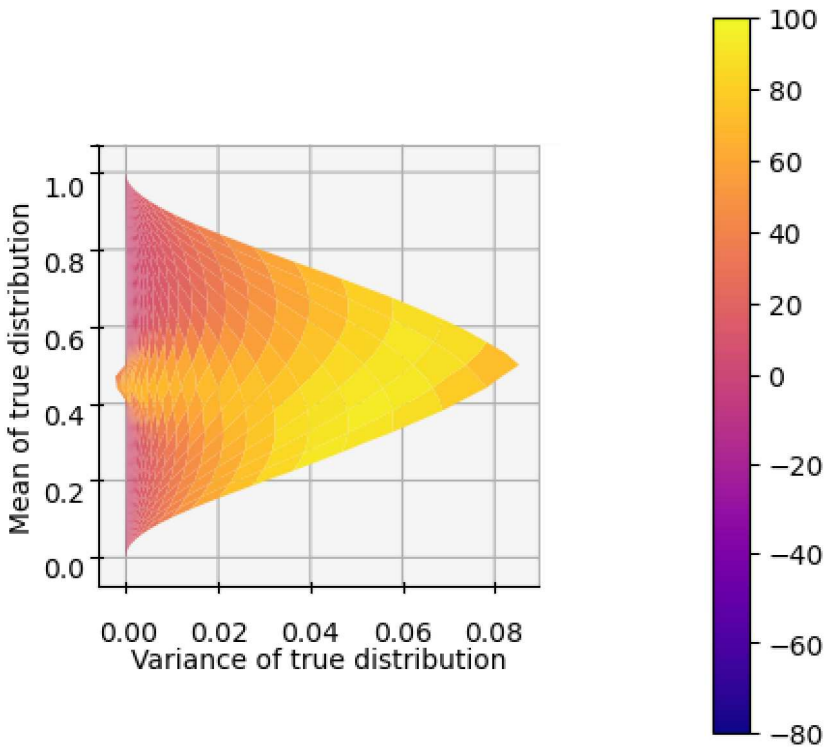

Figure 3: Percent accuracy of the beta distribution model as a function of the true distribution moments.

The simulation suggests that combining four different risk elicitation tasks can help our model learn risk preferences as distributions for a larger group of agent types. The model is especially good in learning agents distributions with high variance or centered around a mean of 0.45. However, there is room for improvement in enabling the model to accurately learn a broader range of distribution types.

Section 3.1's results suggest the optimal way for the model to learn agents' risk preferences is by tailoring the task (i.e., lottery intervals) to their individual risk variability. This section's results suggest that different *combinations* of risk elicitation tasks may identify differently shaped risk preference distributions. A promising direction for future research is to identify optimal combinations of elicitation methods. Another is to develop adaptive risk elicitation methods that respond in-time to agent decisions.

4 Results from real data

In this part, we will analyze data from Holzmeister and Stefan (2021)[14], which elicited subjects' risk preferences in different risk-elicitation tasks. Subjects (N=198) completed four risk elicitation tasks: BRET (T1), CEM (T2), MPL (T3) and SCL (T4). We will use these data and our model to test the out-of-sample prediction power. Basically, we will use risk elicitation data decisions to predict "out-of-sample".[15] For example, using decisions from T1, T2, and T3 to train our model to predict subject decisions on T4.

The beta distribution is defined over values between 0 and 1. To fit our data to this requirement, we scaled the data values to a range of 0 to 1.[16] We focus primarily on the range of values r for each decision, which captures the interval of values r that can support each risk decision. Using this set of lower and upper bounds for r for each task (T1, T2, T3, and T4), we analyzed and elicited the r distribution using our model.[17]

4.1 Challenges to modeling risk preference as a point-estimate

In this section, we argue that our data suggest that it is more sensible to assume that an individual's risk preference is sampled at the time of elicitation from an underlying distribution than to assume that said individual has a fixed risk preference score. To clarify, we are not assuming that the risk score is an unknown point estimate and using a distribution to model our uncertainty about its true value. Instead, we propose that the risk preference of an individual be based on samples from an internal distribution at the time of elicitation.

Why claim that risk preference is a distribution rather than a point estimate? We use the data to falsify model risk preferences as point estimates. If we assume that

1. each individual has a single risk preference score and
2. each elicitation task returns an interval containing the individual's true risk preference score,

then all of the intervals returned by the risk elicitation tasks should overlap. Under these assumptions, the true risk preference of an individual should be at the intersection of all intervals returned by different risk elicitation tasks.

[14]Link to the data:
https://osf.io/5sn2v/

[15]In this paper, the subjects needed to decide on four different risk elicitation tasks when the order of the elicitation tasks was randomly selected. In that spirit, we can test the predictive power of some decisions in the dataset for other decisions.

[16]Here, 0 represents the minimum risk parameter in our dataset (r = -6.96), and 1 represents the maximum (r = 9.14). Risk neutral preferences in our normalized dataset correspond to a normalized value of r = 0.43.

[17]Some subjects in the dataset had empty lower (upper) bound values, indicating that any r value lower (higher) could support their decision. In these cases, we assign the minimum (maximum) value of r to these subjects.

Given assumption 2 (each elicitation task returns an interval containing the individual's true risk preference score), the data from our 198 subjects disprove assumption 1 (each individual has a single risk preference score. Only 8 of the 198 subjects have overlapped risk preference intervals across the four elicitation tasks. Even when we relax assumption 2 by allowing each task to apply an affine transformation to an interval containing the true score, we see the same result: the same 8 of the 198 subjects have overlapping intervals.

We tested the relaxed version of assumption two by choosing constants a_i, b_i to satisfy:

$$\min_{a_i, b_i} \left(\sum_{j=1}^{198} \sum_{i=1}^{4} (H_{ji} + a_i + b_i - L_{ji} + a_i - b_i)^2 \right)^{\frac{1}{2}}, \tag{10}$$

where $H, L \in \mathbb{R}^{198 \times 4}$ are the upper and lower bounds of each risk elicitation task for each subject, respectively. This optimization postulates that the interval returned for the i^{th} individual on the j^{th} task, (L_{ij}, H_{ij}) indicates that the i^{th} individual's true risk preference lies on the interval $(L_{ij} + a_j - b_j, H_{ij} + a_j + b_j)$ and seeks to find the a_js and b_js across all subjects that make the true risk preference intervals as close to each other as possible in the 2-norm. The optimization was carried out using nonlinear least squares with the Levenberg-Marquardt algorithm from 1000 random initial conditions. The final result was chosen based on the transformed intervals that were closest, to test for overlapping intervals. Given these results, we argue that it is more reasonable to assume that the true risk preference of any individual is a distribution rather than a point estimate.

4.2 Subject Estimated Risk Distributions

In this section, we provide summary statistics of agents' estimated risk preferences as distributions, including the distribution of their mean values shown in Figure 4, and the distribution of their variance values shown in Figure 5.

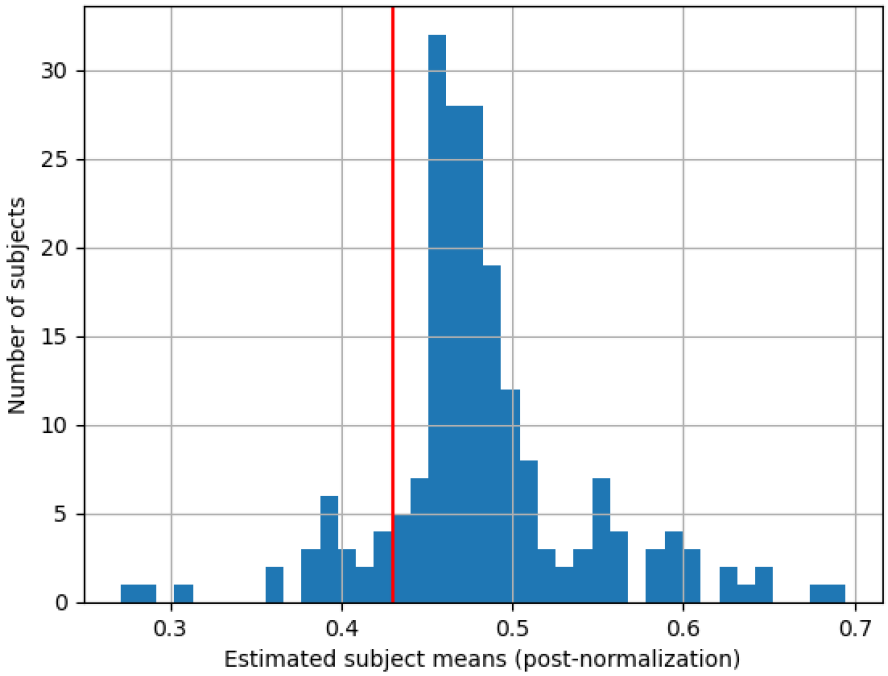

Figure 4: Histogram of subject means post normalization, the red line represents the normalized r value of a risk-neutral agent, which is 0.43.

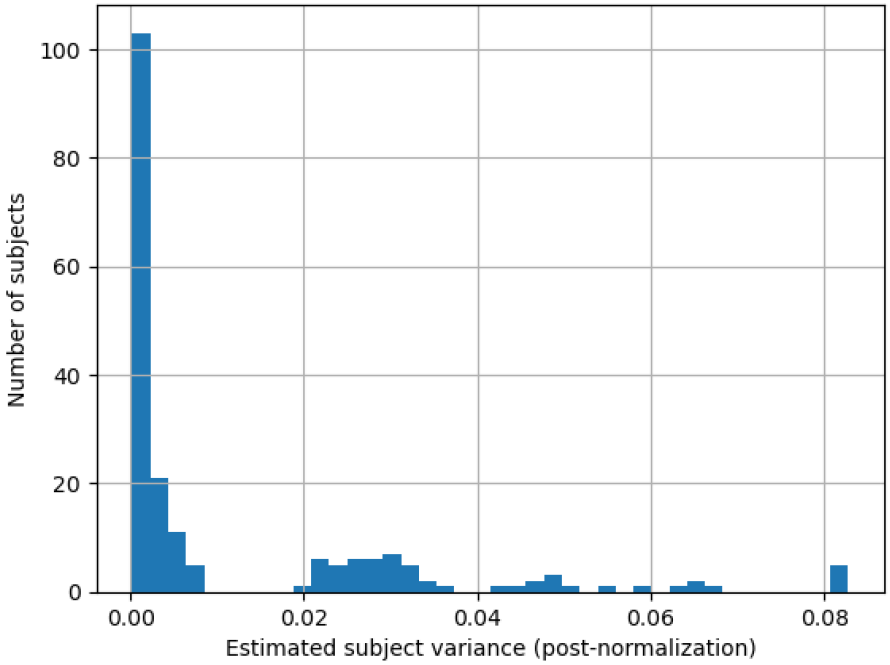

Figure 5: Histogram of subject variance post normalization

The summary information for the beta distribution models collected from subjects in four **data groups** can be found in Table 3. Group one consists of subjects whose model distribution variance is within half a standard deviation of the mean. This group includes all eight subjects

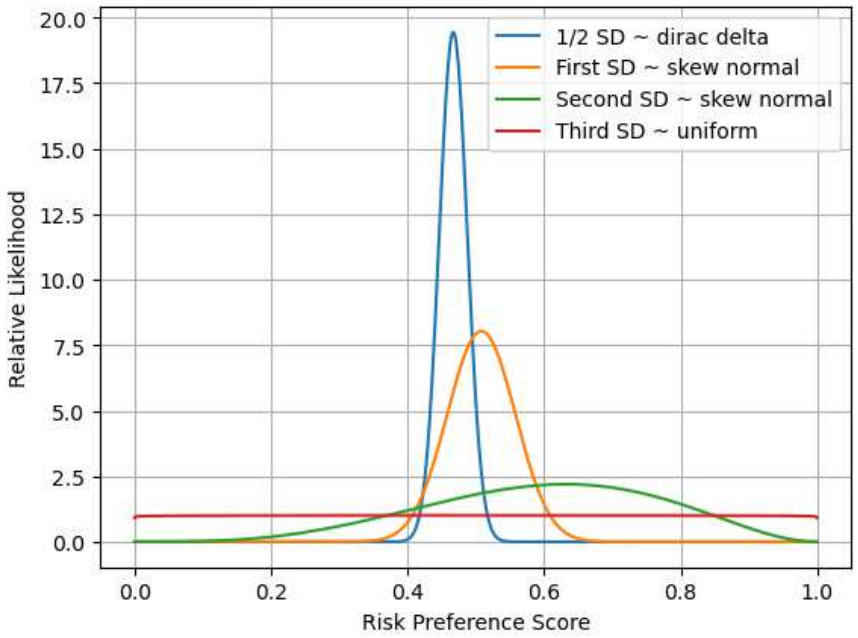

Figure 6: Median representatives of subjects grouped by variance of risk preference within and across tasks. Beta distribution models collected for median representatives of the subjects within half standard deviation, between half and one standard deviation, between one and two standard deviations, and greater than two standard deviations of the mean subject risk-preference distribution variance. See Table 3 for a summary of the four groups.

with consistent r values in Section 4.1.[18] Group two includes subjects whose model distribution variance is from one-half to one standard deviation of the mean; group three includes those with variances between one and two standard deviations; and group four includes subjects with variances greater than two standard deviations. Within each group, the model distributions share qualitative characteristics with the respective distributions listed under **Shape**. The average variance of each group is listed under **Group Var**. See Figure 6 for the representative median model of the four groups.

Data Group	Shape	Num. Subjects	% Total Subjects	Group Var
Half SD	Dirac Delta (Point Estimations)	75	37.88%	0.00047
First SD	Normal (Low Var)	60	30.30%	0.00247
Second SD	Skew Normal	53	26.77%	0.02897
Third SD	Uniform	10	5.05%	0.08593

Table 3: Summary information for beta distribution models across four **data groups**. See Figure 6 for median representatives from each group.

[18] In our model, since all (r) values are scaled by the full range of (r) values possible across all tasks, the variance of the distribution can still be relatively low. This is even if a subject's (r) intervals, which correspond to their decisions, do not overlap from task to task. As a result, this group includes many more subjects than in Section 4.1.

4.3 Testing naive distribution models

In this section, we assume that individual risk preference is sampled from an internal risk preference distribution. We further assume that the risk elicitation methods return an interval containing a sample score from the true underlying distribution. To account for this, we postulate that any score within the interval is equally likely, and so we represent the interval with a discrete uniform distribution; see Section 2.4. We combine the resulting discrete uniform distributions and summarize their net outcome with a single beta distribution to represent the subject's internal risk preference. From that distribution, we can predict the outcome of other risk elicitation tasks.

We compare the learned beta distribution with a special case of the beta distribution, a point estimate, or the Dirac delta distribution centered on the mean of the intervals returned by the elicitation tasks. We compare the accuracy and confidence scores[19] between the point estimate model and the beta model across different amounts of training data, as shown in Figure 7.

When we perform this comparison, we find that the point estimate model is typically very surprised[20] by the outcomes of the risk elicitation task it is set to predict. In other words, the point estimate model has a low confidence score of around 20% regardless of the number of tasks it has to be trained on, while the naive beta distribution model gives a much higher confidence score. Our model confidence score increases when the model is trained by more tasks (3) and receives a higher confidence score of around 70%. An alternative perspective is that the point estimation model presupposes an absence of variability in risk preferences. However, since subjects do exhibit variability in their preferences, the point estimation model fails to capture this and therefore receives a low confidence score (i.e., the model is surprised by the agent's decisions). In contrast, the beta model accounts for variability and achieves a higher confidence score.

Regarding the accuracy score (prediction error), both models have similar error rates when trained with 1 or 3 tasks, without significant difference. When trained on the results of two risk elicitation tasks, the point estimation model consistently achieves a higher accuracy score than the beta distribution model, highlighting a potential trade-off between model accuracy and model confidence in some circumstances. Notably, when both models are trained with three tasks (the highest level of information available), the beta model matches the accuracy of the point estimation model while achieving a much higher confidence score, making it superior to point estimation overall. This is consistent with Section 3, which suggests that to accurately recover risk preferences as a distribution, a richer dataset is required for the beta model to perform effectively.

Since we have confidence in the suitability of the risk preference distribution model (see Section 4.1), we should be able to minimize the accuracy trade-off of our beta distribution while keeping the confidence score above that of the point estimate model. To do so, we challenge the assumption that risk elicitation methods return an interval containing a sample score from the true underlying distribution. In other words, we allow for the impact of framing effects in risk elicitation tasks, which can violate this assumption.

[19] Defined in equations (3) and (4).

[20] In the sense given by the confidence score in Equation (4).

4.4 Testing distribution models with basic accounting for a framing effect in elicitation tasks

In this section, we postulate that there is a form of framing effect in risk-elicitation tasks. We account for this effect by assuming that, rather than returning an interval containing a sample from the individual's underlying risk preference distribution, the task returns a *transformed* interval that contains a sample from the individual's underlying risk preference distribution. We also assume that each elicitation task has a different degree of trustworthiness[21]. The combination of the task's trustworthiness and the transformation it applies before returning intervals is carried out as follows. We take the results of a risk elicitation task used to train our final beta distribution model, the interval $[r_0, r_1]$, and we:

1. stretch or compress the interval about its mean by a multiplicative factor,
2. shift the interval by a bias term and
3. weight its importance (trustworthiness) in computing the MLE of the final beta distribution.

This means that each training task has three associated task interpretation parameters that are used to build beta distributions of predictive risk preferences. We train each of these parameters on two thirds of the data and then build and test beta distribution models of risk preference on the remaining one-third of the subjects. In the end, we achieve an accuracy comparable to that of the point estimation model while maintaining a higher confidence score (see Figure 7).

When we compare the naive model to the one that accounts for the framing effect, the trade-off between accuracy and confidence score remains open, given that the naive model still has a significantly higher confidence score than the model that accounts for the framing effect.

4.5 Possible explanation for the risk elicitation puzzle

The results and simulation analysis of our work can shed light and open a discussion on another potential explanation for the risk elicitation puzzle. Although researchers have tried to solve this puzzle by identifying the source of inconsistencies across risk elicitation methods, our work suggests that part of this inconsistency may be explained by inherent variability in risk preferences, a variability that can be captured by eliciting risk preferences as distributions.

Figure 7 shows that when we have more data on agents' risky decisions, the naive beta model outperforms the point estimation model across all scores (accuracy and confidence). Our simulation analysis suggests that this gap may even increase as more risky decision data become available, improving the accuracy of our beta model.

However, when we control for framing effects in the beta model, we observe a trade-off between accuracy and confidence scores across the two beta models with and without framing control. This trade-off suggests that multiple factors may drive the risk elicitation puzzle: one is the well-known influence of framing, and the other, as our data suggest, is the inherent variability in risk preferences, which we identify as an additional potential explanation.

[21] Trustworthiness can be seen as the ability of the task to capture the true risk preference with the lowest noise, which can also be interpreted as reliability.

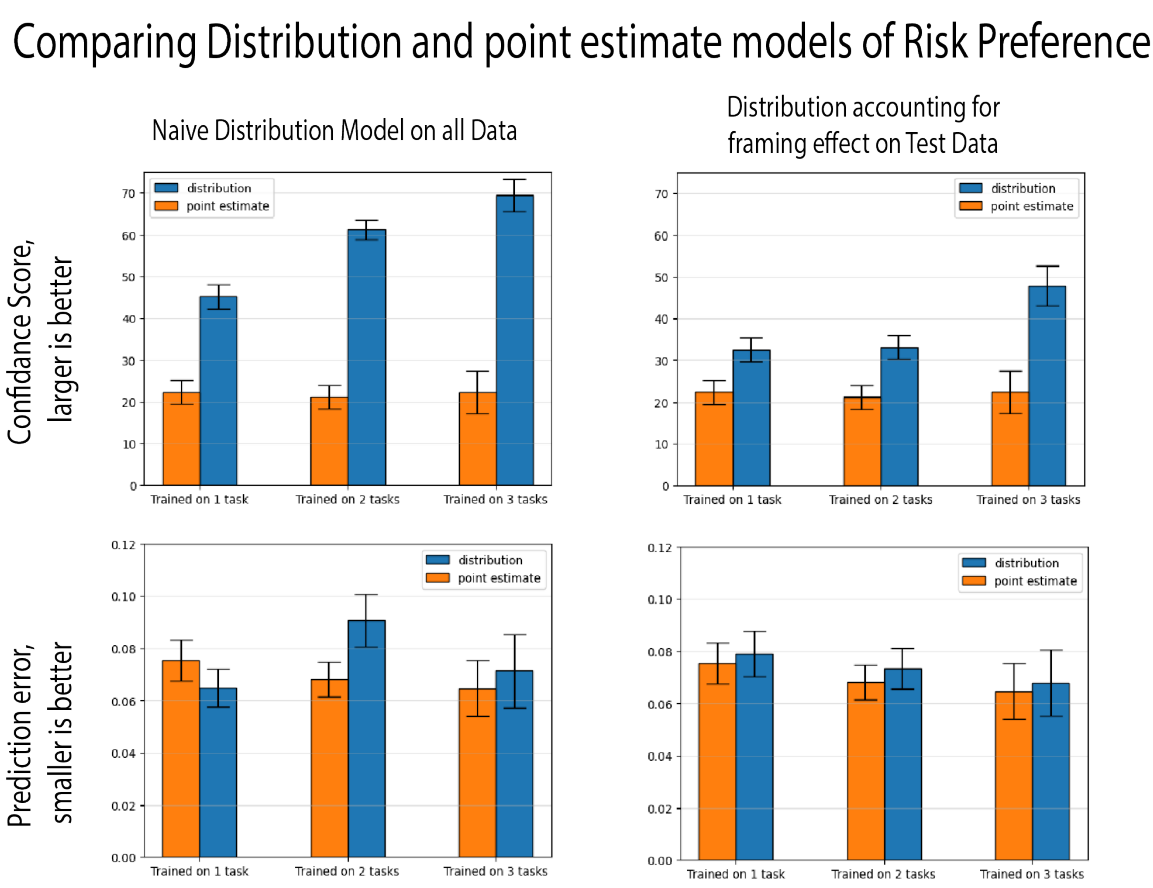

Figure 7: Summary results of how modeling risk preferences with a distribution compares to modeling with a point estimate. Each model uses the results of one, two, or three elicitation tasks to predict the outcome for a separate elicitation task. The naive distribution model is discussed in Section 4.3, and the distribution model accounting for framing effects is discussed in Section 4.4. Error bars show 95% confidence intervals computed using the z-statistic. Note that changing the interpretation of elicitation tasks to account for a framing effect results in a model with statistically indistinguishable prediction error at the cost of a lower average confidence score. However, in any case, the confidence score of the distribution model is significantly higher than that of the point estimate.

5 Discussion

Eliciting risk preferences is a highly complex process, and economists attempt to do so in various ways, struggling to find an efficient method, especially in experimental economics. Strong findings often reveal inconsistencies across elicitation methods and their correlation with risky decisions outside the lab. This work suggests that one direction to improve elicitation methods and address part of this puzzle is to elicit risk preferences as a distribution rather than a point estimate (e.g., a single constant number).

We propose a methodology that uses the Beta distribution model to elicit subjects' risk preferences. We demonstrate that this methodology accurately predicts agents' risk preferences (model parameters) under simulation tests. The model's ability to capture risk preferences as a distribution is strong, but further research can enhance this approach by exploring several untapped venues, such as (1) examining the applicability of this methodology using different probability distributions. These could include multi-modal distributions of higher parameters. (2) Reconsider the assumptions that an individual's risk preference distribution is time-invariant and that samples from this distribution are taken independently. There are experimental setups and data collection methods that enable a modeler to consider a dynamic model of risk behavior. Finally, (3) considering additional sources of data on each subject beyond their risk preference elicitation results. Currently, it is unclear which types of metadata could best improve a distribution-based model of risk preferences.

Our work also sheds light on the importance of tailoring risk elicitation tasks to efficiently capture variability in agent decisions, allowing us to predict risk preferences as distributions using our model accurately. As discussed in Section 3, achieving this with low cost requires carefully designed tasks. One promising direction is to innovate new, more dynamic risk elicitation tasks, such as the approch in Chapman et al. (2024). Going in this direction can allow us to effectively capture agents' risk preferences as distributions with fewer tasks and repetitions, by using a dynamic task that adapts to each agent based on the variability in their risky decisions.

Additionally, we show that empirical data from human subjects support the idea that agents' decisions are better represented as a distribution rather than a single-point estimate (scalar value). Therefore, it could be beneficial to adopt and explore this new method across various datasets and environments. A key advantage of our approach is that it provides predictive power that is better than or similar to traditional point estimation methods, while offering new insights into agents' decision-making. Specifically, it reveals their tendency to risk preferences (r-value) and captures the variance in their decisions, which can be learned from the distribution. The main critique of distributional methods for eliciting risk preferences is that they require much more data collection from subjects to accurately elicit their preferences. We hope that more researchers will explore this distributional approach and that novel methods will be innovated to minimize those costs. More research on distributional elicitation will improve our ability to elicit risk preferences and understand user behavior in risky environments.

References

Abdellaoui, M., Driouchi, A., and l'Haridon, O. (2011). Risk aversion elicitation: reconciling tractability and bias minimization. *Theory and Decision*, 71:63–80.

Alós-Ferrer, C. and Garagnani, M. (2024). Improving risky-choice predictions using response times. *Journal of Political Economy Microeconomics*, 2(2):335–354.

Anderson, L. R. and Mellor, J. M. (2009). Are risk preferences stable? Comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty*, 39:137–160.

Berg, J., Dickhaut, J., and McCabe, K. (2005). Risk preference instability across institutions: A dilemma. *Proceedings of the National Academy of Sciences*, 102(11):4209–4214.

Beshears, J., Choi, J. J., Laibson, D., and Madrian, B. C. (2008). How are preferences revealed? *Journal of Public Economics*, 92(8-9):1787–1794.

Bruner, D. M. (2009). Changing the probability versus changing the reward. *Experimental Economics*, 12(4):367–385.

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

Chapman, J., Snowberg, E., Wang, S. W., and Camerer, C. (2024). Dynamically optimized sequential experimentation (dose) for estimating economic preference parameters. Technical report, National Bureau of Economic Research.

Charness, G. and Chemaya, N. (2023). Repeated experience and consistent risk preferences. *Economics Letters*, 233:111375.

Charness, G., Chemaya, N., and Trujano-Ochoa, D. (2023). Learning your own risk preferences. *Journal of Risk and Uncertainty*, 67(1):1–19.

Charness, G., Garcia, T., Offerman, T., and Villeval, M. C. (2020). Do measures of risk attitude in the laboratory predict behavior under risk in and outside of the laboratory? *Journal of Risk and Uncertainty*, 60:99–123.

Cohen, M., Jaffray, J.-Y., and Said, T. (1987). Experimental comparison of individual behavior under risk and under uncertainty for gains and for losses. *Organizational Behavior and Human Decision Processes*, 39(1):1–22.

Crosetto, P. and Filippin, A. (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, 47:31–65.

Crosetto, P. and Filippin, A. (2016). A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, 19:613–641.

Csermely, T. and Rabas, A. (2016). How to reveal people's preferences: Comparing time consistency and predictive power of multiple price list risk elicitation methods. *Journal of Risk and Uncertainty*, 53:107–136.

Dave, C., Eckel, C. C., Johnson, C. A., and Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, 41:219–243.

Deck, C., Lee, J., and Reyes, J. (2014). Investing versus gambling: experimental evidence of multi-domain risk attitudes. *Applied Economics Letters*, 21(1):19–23.

Deck, C., Lee, J., Reyes, J. A., and Rosen, C. C. (2013). A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes. *Journal of Economic Behavior & Organization*, 87:1–24.

Drichoutis, A. C. and Lusk, J. L. (2016). What can multiple price lists really tell us about risk preferences? *Journal of Risk and Uncertainty*, 53:89–106.

Dulleck, U., Fooken, J., and Fell, J. (2015). Within-subject intra-and inter-method consistency of two experimental risk attitude elicitation methods. *German Economic Review*, 16(1):104–121.

Eckel, C. C. and Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, 23(4):281–295.

Eckel, C. C. and Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, 68(1):1–17.

Fausti, S. W. and Gillespie, J. M. (2000). A comparative analysis of risk preference elicitation procedures using mail survey results. 2000 Annual Meeting, June 29-July 1, 2000, Vancouver, British Columbia 36469, Western Agricultural Economics Association.

Friedman, D., Habib, S., James, D., and Williams, B. (2022). Varieties of risk preference elicitation. *Games and Economic Behavior*, 133:58–76.

Friedman, D., Isaac, R. M., James, D., and Sunder, S. (2014). *Risky Curves: On the Empirical Failure of Expected Utility*. Routledge, London.

Garagnani, M. (2023). The predictive power of risk elicitation tasks. *Journal of Risk and Uncertainty*, 67:165–192.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the Caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, 112(2):631–645.

Gupta, A. K. and Nadarajah, S. (2004). *Handbook of beta distribution and its applications*. CRC Press.

Harbaugh, W. T., Krause, K., and Vesterlund, L. (2010). The fourfold pattern of risk attitudes in choice and pricing tasks. *The Economic Journal*, 120(545):595–611.

He, P., Veronesi, M., Engel, S., et al. (2016). Consistency of risk preference measures and the role of ambiguity: An artefactual field experiment from China. Technical report, University of Verona.

Holden, S. T. and Tilahun, M. (2022). Can the risky investment game predict real world investments? Technical report, Centre for Land Tenure Studies Working Paper.

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

Holzmeister, F. and Stefan, M. (2021). The risk elicitation puzzle revisited: Across-methods (in) consistency? *Experimental Economics*, 24:593–616.

Ihli, H. J., Chiputwa, B., and Musshoff, O. (2016). Do changing probabilities or payoffs in lottery-choice experiments affect risk preference outcomes? Evidence from rural Uganda. *Journal of Agricultural and Resource Economics*, 41(2):324–345.

Isaac, R. M. and James, D. (2000). Just who are you calling risk averse? *Journal of Risk and Uncertainty*, 20:177–187.

Johnson, D. (1997). The triangular distribution as a proxy for the beta distribution in risk analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(3):387–398.

Libby, D. L. and Novick, M. R. (1982). Multivariate generalized beta distributions with applications to utility assessment. *Journal of Educational and Behavioral Statistics*, 7(4):271–294.

Loomes, G. and Pogrebna, G. (2014). Measuring individual risk attitudes when preferences are imprecise. *The Economic Journal*, 124(576):569–593.

Menkhoff, L. and Sakha, S. (2017). Estimating risky behavior with multiple-item risk measures. *Journal of Economic Psychology*, 59:59–86.

Nielsen, T., Keil, A., and Zeller, M. (2013). Assessing farmers' risk preferences and their determinants in a marginal upland area of Vietnam: a comparison of multiple elicitation techniques. *Agricultural Economics*, 44(3):255–273.

Parker, C. and James, F. (2024). Portfolio modeling and selection when asset returns are drawn from any distribution. *Available at SSRN 4706714*.

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., and Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, 1(11):803–809.

Peizer, D. B. and Pratt, J. W. (1968). A normal approximation for binomial, f, beta, and other common, related tail probabilities, i. *Journal of the American Statistical Association*, 63(324):1416–1456.

Perez, F., Hollard, G., and Vranceanu, R. (2021). How serious is the measurement-error problem in risk-aversion tasks? *Journal of Risk and Uncertainty*, 63(3):319–342.

Reynaud, A. and Couture, S. (2012). Stability of risk preference measures: results from a field experiment on French farmers. *Theory and Decision*, 73:203–221.

Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica*, 15(60):243–253.

Schoemaker, P. J. (1982). The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature*, 20(2):529–563.

Szrek, H., Chao, L.-W., Ramlagan, S., and Peltzer, K. (2012). Predicting (un) healthy behavior: A comparison of risk-taking propensity measures. *Judgment and Decision Making*, 7(6):716–727.

Wakker, P. P. (2008). Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, 17(12):1329–1344.

A Elicitation methods

This section briefly summarizes four risk elicitation tasks, which are among the most common in the experimental economics literature and form the basis of our empirical data. Each task is adapted from the experiment conducted by Holzmeister and Stefan (2021), which includes a modified version of these four tasks, and we used Holzmeister and Stefan (2021) empirical data to test our estimation method.

First, Single Choice Lists (SCL) developed by Eckel and Grossman (2002) (using five gambles) and further in Eckel and Grossman (2008) (using six gambles) presents an individual with a list of gambles in which they choose between a risky and a safe gamble. The individual's range of r is determined by which gamble they choose from the list.

Second, Multiple Price Lists (MPL) developed by Holt and Laury (2002) presents an individual with ten lotteries in which they must choose between two gambles, risky gamble A or safe gamble B. The list is designed so that each individual chooses gamble B in the first decision, and the individual's range of r is determined by which lottery they deviate from choosing gamble B to choosing gamble A. Also referred to as HL in the literature.

Third, the Certainty Equivalent Method (CEM) developed by Cohen et al. (1987) determines the point where an individual is indifferent between a fixed risky lottery and a series of sure payoffs. Similarly to the MPL, an individual must choose between risky option A and safe option B. With parametrization as in Abdellaoui et al. (2011), the participant has a set of 9 decisions.

Fourth, the Bomb Risk Elicitation Task (BRET) developed by Crosetto and Filippin (2013) determines how long an individual is willing to continue collecting earnings in the face of potentially losing what they have collected. The participant chooses when to stop the task and can collect up to 100 boxes. The number of boxes that are collected determines the participant's r value.

B Figures and tables

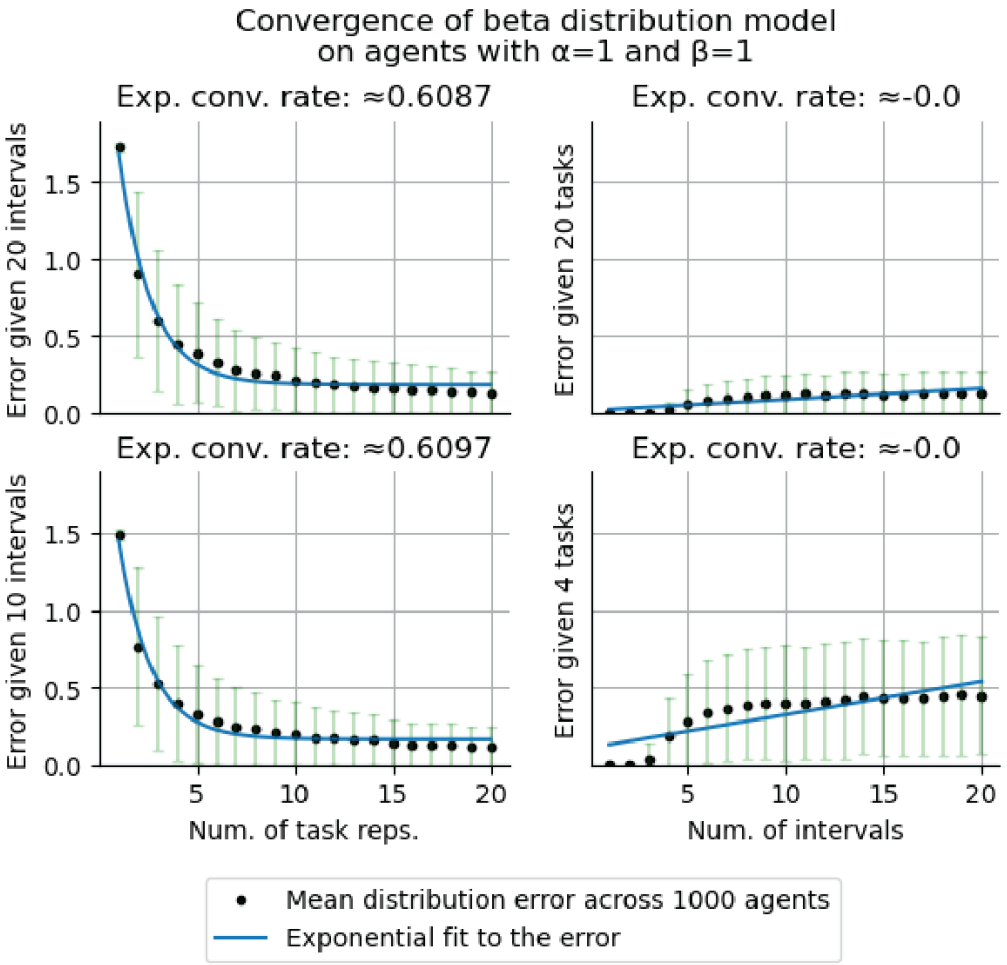

Figure 8: Sample plots of the results from fitting a beta distribution to Agent 1. Lower granularity in elicitation tasks leads to overfitting the data as it is generated by an effectively lower parameter distribution than the model. Error bars in green give the standard deviation of the model error across 1000 trials.

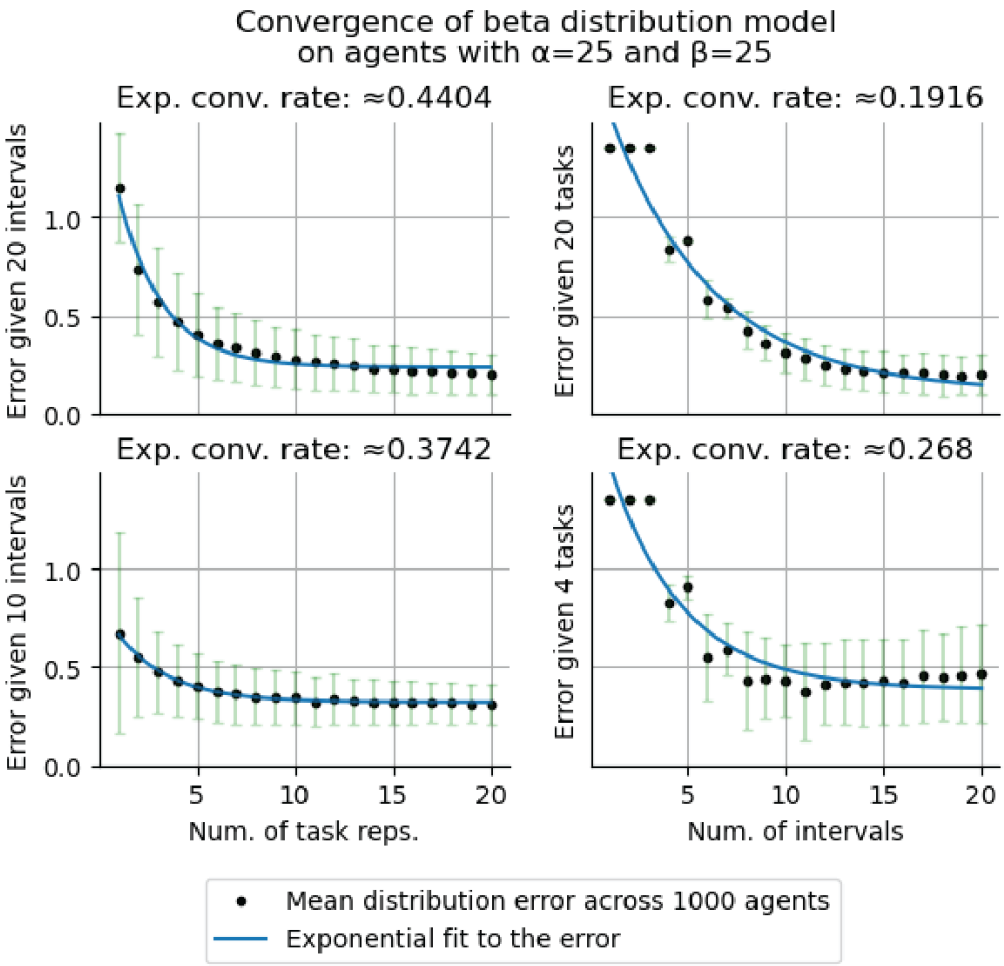

Figure 9: Sample plots of the results from fitting a beta distribution to Agent 2. An increase in a number of task repetitions and/or interval granularity reduced model error exponentially. Error bars in green give the standard deviation of the model error across 1000 trials.

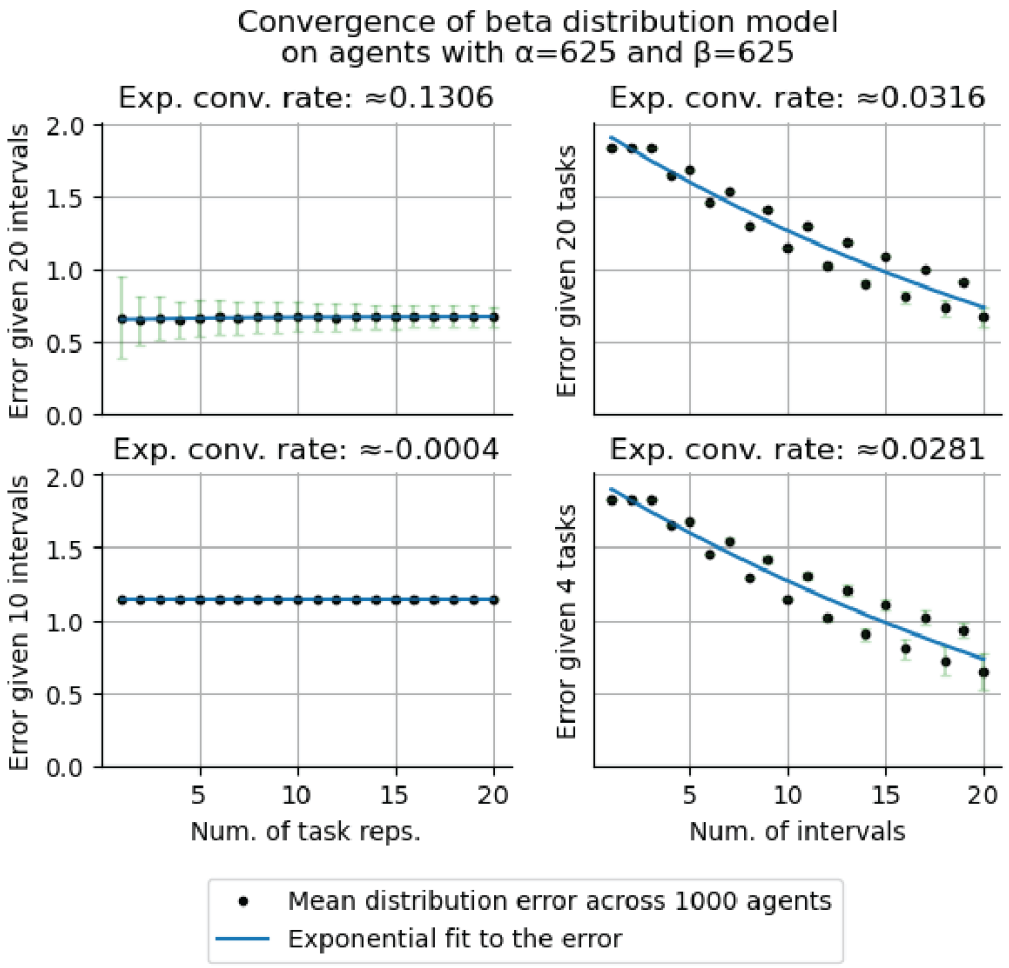

Figure 10: Sample plots of the results from fitting a beta distribution to Agent 3. Since task-to-task responses are so consistent, increasing the number of task repetitions has little effect. The parity of the interval number has a strong effect on the final model error. This is unlikely to occur outside of simulation; see Section 3.1.3. Error bars in green give the standard deviation of the model error across 1000 trials.

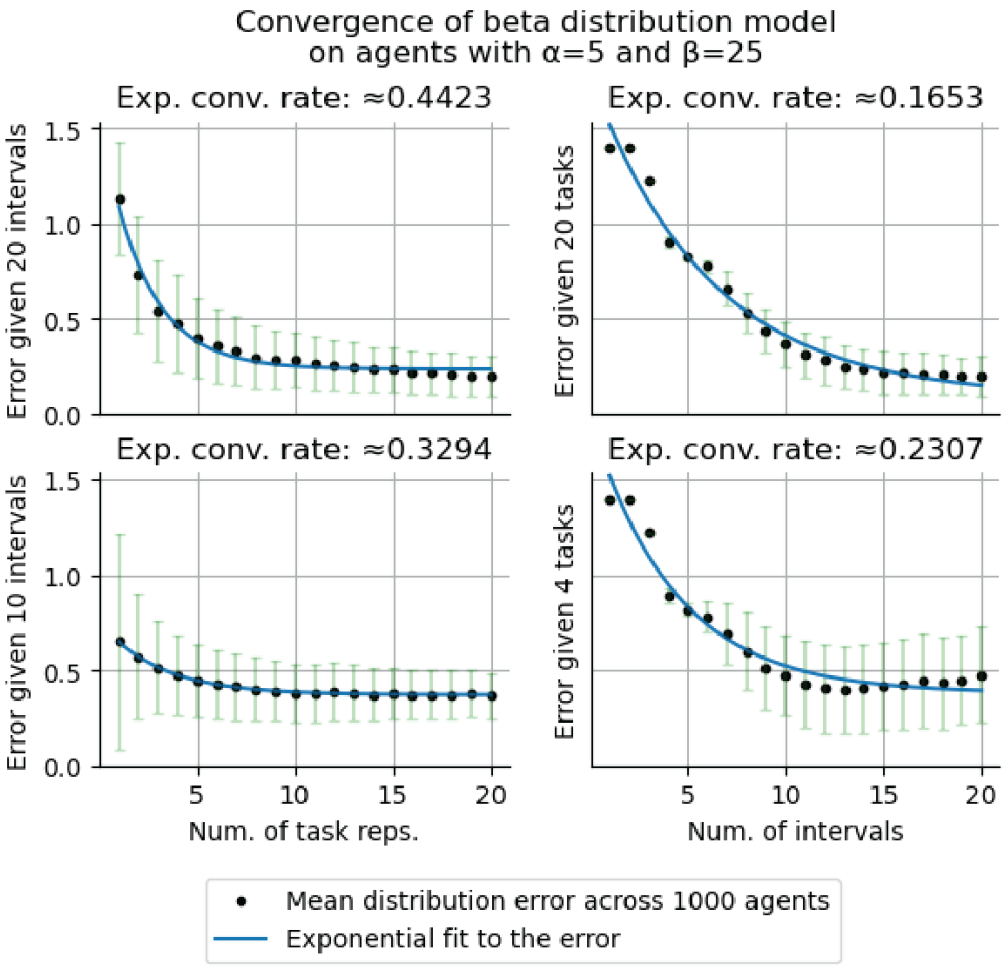

Figure 11: Sample plots of the results from fitting a beta distribution to Agent 4. An increase in a number of task repetitions and/or interval granularity reduced model error exponentially. Error bars in green give the standard deviation of the model error across 1000 trials.

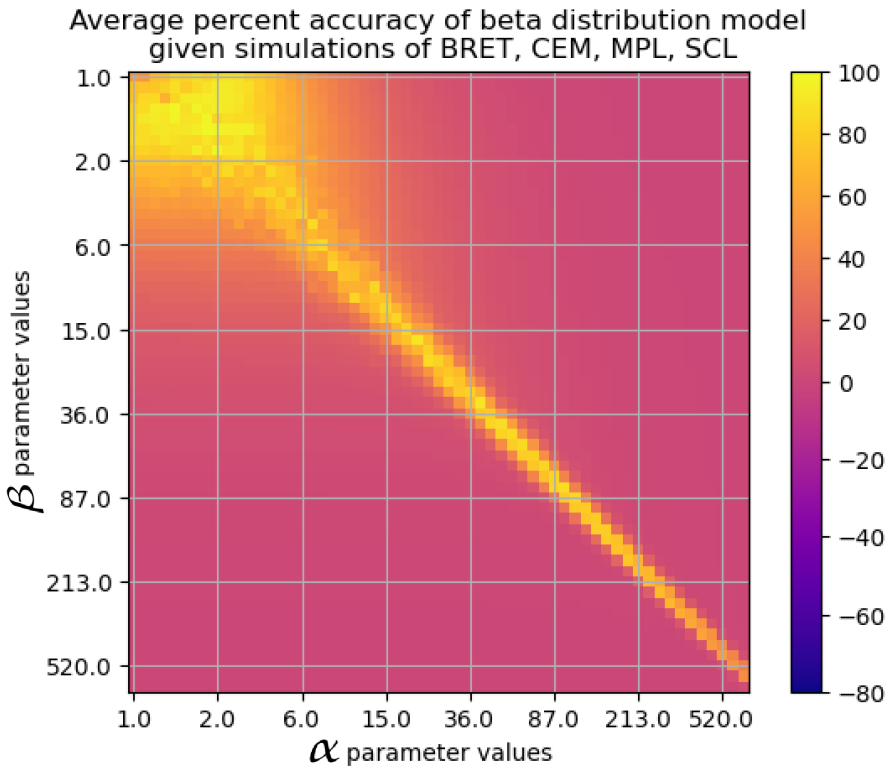

Figure 12: Average percent accuracy of the beta distribution model given simulations with intervals matching those of BRET, CEM, MPL, and SCL in Holzmeister and Stefan (2021).