Gutenberg School of Management and Economics

& Research Unit “Interdisciplinary Public Policy”

Discussion Paper Series

Belief Elicitation with Multiple Point Predictions

Markus Eyting, Patrick Schmidt

November 16, 2020

Discussion paper number 1818

Johannes Gutenberg University Mainz
Gutenberg School of Management and Economics
Jakob-Welder-Weg 9
55128 Mainz
Germany
https://wiwi.uni-mainz.de/

Contact details

Markus Eyting
Chair of Digital Economics
University of Mainz
Jakob-Welder-Weg 9
55128 Mainz
Goethe University Frankfurt
60323 Frankfurt am Main
Germany

meyting@uni-mainz.de

Patrick Schmidt
Goethe University Frankfurt
60323 Frankfurt am Main
Heidelberg Institute for Theoretical Studies
69118 Heidelberg
HITS gGmbH
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg
Germany

Patrick.Schmidt@h-its.org.

Belief Elicitation with Multiple Point Predictions

Markus Eyting
GSEFM Frankfurt and Johannes Gutenberg University Mainz
and
Patrick Schmidt*
University of Zurich

This version: November 16, 2020

Abstract

We propose a simple, incentive compatible procedure based on binarized linear scoring rules to elicit beliefs about real-valued outcomes - *multiple point predictions*. Simultaneously eliciting multiple point predictions with linear incentives reveals the subjective probability distribution without pre-defined intervals or probabilistic statements. We show that the approach is theoretically as robust as existing methods, while adapting flexibly to different beliefs. In a laboratory experiment, we compare our procedure to the standard approach of eliciting discrete probabilities on pre-defined intervals. We find that elicitation with multiple point predictions is faster, perceived as less difficult and more consistent with a subsequent decision. We further find that multiple point predictions are more accurate if beliefs vary between participants. Finally, we provide experimental evidence that pre-defined intervals anchor reports.

Keywords— elicitation of subjective expectations; partial identification; quantiles; experiment

*Address for correspondance: Patrick Schmidt, HITS gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany. E-mail: Patrick.Schmidt@h-its.org. Phone: 0049 69 798-34837.

1 Introduction

Economic modelling and decision making under uncertainty often rely on subjective beliefs and the elicitation thereof. We consider beliefs about real-valued variables (e.g., income, profit, inflation, growth rates, exchange rates, survival rates, infection rates, second-order probabilities, or the timing of an event), which take the form of continuous probability distributions. The economic literature provides numerous applications for eliciting beliefs about real-valued variables in practice.[1] While mechanisms to elicit these beliefs can be theoretically equivalent, they often have different psychological implications. Experimental evidence for the applicability of different methods is context dependent (for reviews about belief elicitation in the lab, see Schlag *et al.*, 2015; Schotter and Trevino, 2014; Trautmann and van de Kuilen, 2015).

In this paper, we propose to elicit subjective probability distributions framed as multiple point predictions (MPP). We avoid probabilistic language and use a simple incentive scheme based on aggregating multiple linear scoring rules (Schlaifer and Raiffa, 1961). We provide a theoretical framework for the incentivized elicitation of subjective probability distributions and compare our approach in a laboratory experiment to the commonly applied method of eliciting beliefs using interval probabilities (IP). We find that elicitation with MPP is faster, perceived as less difficult and more predictive of subsequent behavior. We further find that elicited distributions of MPP are more accurate if participants received different information signals and show that pre-defined intervals anchor reports.

1.1 Elicitation of Interval Probabilities

A common procedure for eliciting subjective probability distributions is to divide the real line into intervals and elicit the discrete distribution on those intervals. This elicitation of IP can be incentivized with the widely applied quadratic scoring rule (QSR) (Costa-Gomes *et al.*, 2014; Harrison *et al.*, 2014; Huck and Weizsäcker, 2002; McKelvey and Page, 1990; Nyarko and Schotter, 2002; Rutström and Wilcox, 2009). By rewarding subjects with the probability of winning a fixed

[1] For reviews on belief elicitation in macro and development economics in general see Manski (2018) and Delavande *et al.* (2011). Armantier *et al.* (2013) review elicitation of inflation expectations. For examples eliciting beliefs about real-valued variables see Dominitz and Manski (1997) for income, Vargas Hill (2009) for coffee prices, and De Mel *et al.* (2008) for small business profits. Other examples arise for the onset of an event, as the timing can be considered as real-valued outcome: Delavande and Kohler (2009) elicit beliefs on mortality over time. Similar data is collected in the Health and Retirement Study; See Wang (2014) for an application. Carman and Kooreman (2014) elicit beliefs about contracted influenza and heart disease over time.

payoff, binarized scoring rules are incentive compatible for risk averse or risk seeking preferences.[2]

We generalize this procedure and show in a unified framework, using binarized scoring rules, that any bounded density can be elicited without pre-defined intervals.

1.2 Elicitation of Multiple Point Predictions

For events, no simple linear scoring rule truthfully elicits the event probability. However, for subjective probability distributions on the real line, linear incentives identify pre-defined CDF levels. We propose to elicit MPP simultaneously by aggregating asymmetric linear incentives. This procedure identifies quantiles of the subjective probability distribution.

We show that MPP can be used to elicit points of the subjective CDF without assuming risk neutrality. While IP allow to choose at which outcome levels the CDF is revealed, MPP allow to choose at which probability levels the CDF is revealed. Without anchors or explicit probabilistic statements[3], MPP provide similar information about the subjective probability distribution as IP.

The intuition behind point predictions is rather natural. On a regular basis we encounter uncertainty such as "*How many days will I need to finish the project?*". We commonly express our beliefs in point estimates (e.g., "*I need* 10 *days.*") instead of probabilities ("*There is a* 50% *chance that I need less than* 10 *days.*"). Moreover, the consequences of over- or underestimation might be rather different. If finishing one day too late is more costly than one day too early, we would express a higher estimate ("12 *days.*"). If instead finishing a day early is more costly than a day late, we would express a lower estimate ("8 *days.*").

Applying MPP, we rely on the same intuition. Each point estimate influences the payout in a simple linear relationship. By varying the asymmetry between under- and overestimation, we can elicit different quantiles of the underlying subjective probability distribution. We argue that point predictions allow to construct a simple and intuitive elicitation mechanism.

Our procedure overcomes several potential caveats of currently used methods. Many methods require individuals to communicate their beliefs in probabilistic form. Whereas expert forecasters may have little difficulties communicating in probabilistic form, most populations (e.g., high school students) may struggle when asked for IP. Outside of the lab, individuals rarely communicate their beliefs in that way. Simple point predictions, however, were criticized for being uninformative

[2]This approach is also referred to as binary lottery procedure. The idea goes back to Smith (1961). Binarized scoring rules are analyzed in Hossain and Okui (2013) for generic properties and the mean, in Schlag and van der Weele (2013) for single quantiles, and in Harrison *et al.* (2015) for discrete distributions. Other methods that account for risk-aversion are based on the mechanism introduced in Karni (2009) which involves two layers of randomization. Demuynck (2013) proposes to elicit single quantiles and Qu (2012) IP.

[3]We call reports "probabilistic" if they are in the form of a probability distribution.

about the uncertainty (Engelberg *et al.*, 2009). MPP allow convenient communication and reveal uncertainty.

Moreover, as many existing methods depend on pre-defined intervals they suffer from anchoring and bin effects. Benjamin *et al.* (2017) provide evidence of bin effects in incentivized experiments, whereby the belief reports systematically depend on the intervals used to elicit beliefs and Tversky and Kahneman (1975) introduce individuals' tendency to insufficiently adjust their estimates from anchors. This effect has later been confirmed in experimental work by Wright and Anderson (1989) and quantitatively assessed by Jacowitz and Kahneman (1995).

Finally, incentivized reports are often based on complex payoff functions (e.g., *proper scoring rules* first introduced in Brier, 1950; Winkler, 1967). Some procedures show payoffs contingent on outcomes, which allows respondents to explore the incentive structure (e.g., Harrison *et al.*, 2014; Holt and Smith, 2016). Other procedures explicitly tell participants that it is optimal to report their "true beliefs". Offerman *et al.* (2009) argue that this recommendation is debatable, as it depends on decision theoretic assumptions. The simple linear scoring rule in MPP facilitates rather than complicates the elicitation of subjective beliefs.

1.3 Theoretical Contribution

To provide the theoretical foundations for MPP, we build on the seminal work in Hossain and Okui (2013) and extend binarized scores to the elicitation of the entire density (see Proposition 1) and *simultaneous* elicitation of multiple quantiles (see Theorem 1). We aggregate multiple linear scoring rules and elicit *a set of quantiles simultaneously* without assuming bounded support or limiting tail behavior. Previous work established similar results for the mean property (Hossain and Okui, 2013) and single quantiles (Schlag and van der Weele, 2013) under more restrictive assumptions on the belief distribution. Under risk-neutral preferences, related scoring procedures have been considered in Jose and Winkler (2009) and Fissler and Ziegel (2016). We consider the more general class of probabilistically sophisticated preferences as introduced in Machina and Schmeidler (1992). Thus, our results extend to risk-averse and risk-seeking expected utility preferences, but do not generally hold for uncertainty (or ambiguity) averse preferences.

We further show that extremely asymmetric linear incentives can reveal the minimum and maximum of the distribution (see Proposition 3). Bellini and Bignozzi (2015) show that the minimum and maximum are properties that are non-elicitable with standard procedures. To the best of our knowledge, we provide the first approach that reveals a non-elicitable property.

Compared to the QSR on IP, our linear incentives are unbounded for unbounded subjective belief distributions. Binarized scores require bounded scores to be incentive compatible. We show that the arising biases can be bounded conveniently. In numerical simulations we further illustrate

reference	framing	incentives	binarized	outcome
Manski and Neri (2013)	IP	QSR	No	elicited event probability
Neri (2015)	IP	QSR	No	bid and offer in auction experiment
Harrison *et al.* (2017)	IP	QSR	No	number of coloured balls
Budescu and Du (2007)	IP,Q	unspecific	No	stock prices
Palley and Bansal (2019)	IP,Q	QSR,LSR	Yes	multiple
Harrison *et al.* (2015)	IP	QSR	Yes	number of coloured balls
Hossain and Okui (2013)	SPP	QSR	Yes	hypothetical stock prices
Costa-Gomes *et al.* (2014)	SPP	QSR	No	transfer in trust game
Dufwenberg and Gneezy (2000)	SPP	LSR	No	return in lost wallet game
Sapienza *et al.* (2013)	SPP	PCSR	No	return in trust game
Charness and Dufwenberg (2006)	SPP	PCSR	No	ratio of cooperation in trust game
Kirchkamp and Reiß (2011)	SPP	LSR	No	bid in auction game
Survey of Professional Forecasters[5]	IP	-	-	inflation, GDP, etc.
Survey of Consumer Expectations[6]	IP	-	-	inflation, household income, etc.
Altig *et al.* (2020)	IP,MPP	-	-	sales, investments, etc.

Table 1: **Examples of belief elicitation on real-valued outcomes.** Framing of elicitation is denoted by interval probabilities (IP), single point predictions (SPP), quantiles (Q), and multiple point predictions (MPP). Incentives are denoted as quadratic scoring rule (QSR), linear scoring rule (LSR), and piecewise-constant scoring rule (PCSR).

adequate design choices to prevent those biases. In the simulations, we also consider linear incentives without binarizing and show that risk-aversion would lead to overreporting of uncertainty.

Finally, we discuss why MPP often provide sharper bounds on the CDF than IP. Taking an example from the Survey of Consumer Expectations (Armantier *et al.*, 2017) by the Federal Reserve Bank of New York, we illustrate how IP often provide little information on heterogeneous belief distributions, whereas MPP are informative irrespective of the heterogeneity of beliefs[4].

1.4 Experimental Evidence

In the second part of the paper, we present an experimental application of belief elicitation by MPP, and compare it to the elicitation of IP.

Previous laboratory experiments focused mainly on event probabilities (see Schlag *et al.*, 2015; Schotter and Trevino, 2014; Trautmann and van de Kuilen, 2015, for reviews). Belief elicitation on real-valued outcomes dominantly relied on IP (Table 1) or single point predictions.

The statistical literature focuses more on unincentivized expert elicitation (see O'Hagan *et al.*,

[4]We differentiate between "heterogeneous" and "homogeneous" beliefs. While the former describes varying beliefs across participants, the latter describes similar beliefs across participants. In our experiment, we induce belief heterogeneity by varying the level of information that participants receive about the outcome.

[5]See Croushore (1993).

[6]See Armantier *et al.* (2017).

2006, for a review). Similarly, economic surveys mostly rely on IP (often called histograms or bins) without incentives. We conjecture that applied work relies mostly on IP instead of quantile elicitation, as quantiles are arguably more complicated to explain to participants than IP.

We add to the sparse experimental literature on real-valued outcomes and provide first evidence on the merits and drawbacks of eliciting MPP compared to the widely used elicitation of IP.

In a concurrent paper, Palley and Bansal (2019) provide a related experimental comparison by explicitly eliciting quantiles. We rely on a more intuitive framing with MPP. Further, we analyse calibration (unbiasedness) and accuracy in a more general framework based on the realized value instead of empirical marginal distributions.

For the sake of comparability, we incentivize the IP reports with the QSR and apply binarized scores for both methods. In a laboratory experiment, we elicit subjective probabilities over five different real-valued outcomes. The domains differ in the level of complexity and cover symmetric and skewed distributions, ambiguity and skill-based assessments. We exogenously vary the strength of the available information and argue that this affects the degree of belief heterogeneity across participants. This allows us to compare performance under homogeneous beliefs (previous to an information update) as well as varying levels of heterogeneous beliefs (after an information update). We find that neither approach dominates across all evaluation metrics and applications. Despite the fact that the two approaches predict identical responses under probabilistically sophisticated preferences, we find strong evidence for differing response behaviors within evaluation metrics and applications.

Throughout, our results indicate that the interval thresholds serve as anchors. Increasing the information set on which beliefs are elicited, improves the accuracy of MPP reports more than the accuracy of IP reports. We conjecture, that insufficient adjustments from external anchors prevent participants reporting IP from incorporating the full information set. We further find that the distributions elicited by MPP are more consistent with a subsequent decision. This effect is especially prominent after providing strong information signals. This is an important property, when elicitation is used in economic modelling, instead of forecasting.

We confirm our hypothesis of the interval thresholds functioning as anchors in a separate experiment and show that the position and length of the intervals heavily influence the mean and standard deviation of elicited distributions. In applications, beliefs are unknown and vary across participants, rendering uniformly adequate intervals infeasible. Individual specific adaptation of intervals, on the other hand, can influence responses and complicate comparisons across individuals. Especially for non-expert respondents these effects may lead to distorted reports. As MPP operates without pre-defined outcome values, reports cannot be biased by externally given anchors.

With regard to applicability, participants that reported beliefs with MPP required less time

and were more likely to react positively to subjective perception questions after the experiment.

In the following section, we provide the theoretical background on property elicitation, show a numerical study on biases from risk-aversion and unbounded support, and discuss how to recover probability distributions after eliciting CDF points. Section 3 describes the experimental design. Results are provided in Section 4, followed by a discussion in Section 5. The appendix contains a more technical treatment, proofs, and additional results. An online supplementary document is available with additional details and a description of the experiment.

2 Theory of Property Elicitation

In this section we review the theoretical background of the elicitation of subjective probability distributions with binarized scoring rules. Consider the task of eliciting an agent's belief about a real-valued random variable y. A state of belief is represented by a subjective probability $\mathbb{P}$, which is denoted as a CDF –or where possible by its density– on the outcome space $\mathbb{R}$ of y.

We elicit specific properties $T(\mathbb{P})$ of the subjective probability (e.g., a quantile or the likelihood of an interval) based on the well-known procedure of binarized scoring rules. The agent chooses a report x from the report space $\mathcal{X}$. After observing the random variable y, the agent is remunerated based on the scoring rule s, which is a function of the outcome y and the issued report x. Specifically, the agent receives a prize if the score $s(x, y)$ exceeds a uniformly distributed random variable with suitably chosen support.

We assume that the agent has no other stakes concerning the random variable y and acts probabilistically sophisticated (Machina and Schmeidler, 1992). See Regularity Conditions 1 in the Appendix for details. Thus, our results hold for expected utility maximizing agents, irrespective of their risk attitude.[7] The remainder of this section tackles the question which properties of the distribution can be elicited by the approach above and how the reports (partially) identify the subjective probability distribution $\mathbb{P}$. We call an elicitation method *incentive compatible for a property* T if the optimal report of a probabilistically sophisticated agent with subjective probability $\mathbb{P}$ is $T(\mathbb{P})$. Further, a series of elicitation methods is called *essentially incentive compatible* if the optimal report of an agent converges to $T(\mathbb{P})$. Thus, essentially incentive compatible mechanisms allow to elicit a property with an arbitrary degree of accuracy.

Hossain and Okui (2013) provide an essentially incentive compatible mechanism for the mean under tail assumptions. They also construct a generic incentive compatible mechanism for every

[7] We note that there is mixed empirical evidence on binarized incentives inducing risk neutral behavior with Cox and Oaxaca (1995) and Selten *et al.* (1999) providing evidence against and Harrison *et al.* (2013, 2015) and Hossain and Okui (2013) providing evidence for the validity of the procedure.

property with bounded scoring rules. Schlag and van der Weele (2013) show examples for several properties, including the quantile assuming bounded support. In Section 2.2, we add a straight forward extension to reveal the entire density. In Section 2.3, we focus on MPP and show that our procedure is essentially incentive compatible for multiple quantiles simultaneously without assuming bounded support or restricting tail behaviour. Additionally, we show that extreme quantiles can be used to construct essentially incentive compatible mechanisms for the minimum and maximum of the distributions.

2.1 Eliciting Interval Probabilities

Let us consider the most prominent example for the elicitation of a property: interval probabilities. The common approach is to choose some thresholds $c_1, \ldots, c_{n-1}$ that define the respective property $T_c(\mathbb{P}) = (\mathbb{P}(y \leq c_1), \mathbb{P}(c_1 < y \leq c_2), \ldots, \mathbb{P}(y > c_{n-1}))$ and to apply the QSR for discrete probabilities. The eligible reports are probability vectors for n outcome values, $\mathcal{X} := \{x \in [0,1]^n | \sum_i x_i = 1\}$. After issuing the report $x = (x_1, \ldots, x_n)$ and observing the outcome y in the k^{th} interval, the agent wins the prize if the QSR for multiple events,

$$s(x, y) = 2x_k - \sum_i x_i^2 + 1, \tag{1}$$

exceeds a uniformly drawn random variable with support $[0, 2]$. Under probabilistic sophistication, this elicitation mechanism is incentive compatible for the discrete probability distribution T_c as the score is bounded (Hossain and Okui, 2013). That means, the agent is incentivized to report the true probability distribution. Note, however, that this procedure does not reveal the entire distribution $\mathbb{P}$. We refer to this method as elicitation of interval probabilities.

2.2 Eliciting the Entire Probability Distribution

We show how to elicit the entire probability distribution using a continuous generalization of the QSR (Matheson and Winkler, 1976). The report space $\mathcal{X}$ contains probability density functions. We assume that the eligible distributions have bounded densities with some bound B. Given a reported density function $p(\cdot)$, we compute the score as

$$s(p, y) = 2p(y) - \int_{\mathbb{R}} p(w)^2 dw + B.$$

Subsequently, we draw a uniformly distributed random variable on $[0, 3B]$. The agent receives a fixed payoff if the score exceeds the random draw.

Proposition 1 (density)**.** *Under probabilistic sophistication the mechanism described above is incentive compatible for the probability density function.*

Up to technicalities, the proposition follows from the properness (Gneiting and Raftery, 2007) and boundedness of the score (Hossain and Okui, 2013, Theorem 1). An elementary proof is given in Appendix A.

With Proposition 1 any property could be elicited indirectly by eliciting the density and subsequently calculating the respective property. However, the communication of a whole distribution can be burdensome, or impossible, without parametric assumptions and the involved scoring rule is complex.

2.3 Elicitation of Multiple Point Predictions

Instead of probabilistic reports, we propose to elicit multiple point predictions. Each prediction is incentivized by a different linear scoring rule, which allows to infer quantiles of the underlying distribution. While the quantile is a rather complex concept, there exist simple linear proper scoring rules. In contrast, interval probabilities are simple concepts, that can only be incentivized by complex scoring rules (like the QSR). A key element of our approach is to focus the participant's attention on the payoff function. The probabilities are subsequently inferred by the researcher. In doing so, MPP do not require the participant to understand formal probability concepts, nor to communicate in probabilistic statements.

For each point prediction x the payoff depends on the distance between x and the true outcome y. In particular, this difference is multiplied by a positive factor a or b, depending on whether the point forecast underestimates or overestimates, and then deducted from an initial endowment e.

$$s_{a,b}(x,y) = \begin{cases} e - a \cdot |x-y| & \text{if } \; x \leq y \text{ (underestimation)}, \\ e - b \cdot |x-y| & \text{if } \; x > y \text{ (overestimation)}. \end{cases} \tag{2}$$

The expected score maximizing strategy is to report the quantile of P with level $\alpha = \dfrac{a}{a+b}$ (Schlaifer and Raiffa, 1961). With the results in Hossain and Okui (2013) it follows that the binarized version of the score in Equation (2) is robust beyond risk neutrality if the support of the distribution is bounded. A single point prediction, however, does not allow to do meaningful inference on the subjective probability distribution.

The following theorem shows how to elicit multiple quantiles simultaneously without assuming bounded support. Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be a vector of n different quantile levels on the unit interval $(0,1)$. We choose appropriate positive numbers a_i, b_i such that $\alpha_i = \frac{a_i}{a_i+b_i}$. For each level the agent

issues a point estimate x_i. Subsequently, the final score is computed by summing up the positive values of each single score, i.e.,

$$s_\alpha(x, y) = \sum_{i=1}^{n} max(s_{a_i,b_i}(x_i, y), 0). \tag{3}$$

The agent receives a fixed payoff if the score exceeds a uniformly distributed random draw on $[0, ne]$.

Theorem 1 (multiple quantiles)**.** *Assume a probabilistically sophisticated agent with a subjective probability with strictly positive density.*

(i) For sufficiently large endowments e the mechanism above is essentially incentive compatible for the quantiles with levels $\alpha_1, \ldots, \alpha_n$.

(ii) Consider a given point prediction x_i^*. *If the mass of the tail intervals can be bounded by* $\mathbb{P}_0(y < x_i^* - e/b_i) < c_1$ *and* $\mathbb{P}_0(y > x_i^* + e/a_i) < c_2$, *the mechanism is incentive compatible for the quantile with level* α_i^* *such that the error in terms of the quantile level can be bounded by*

$$-\alpha_i c_2 < \alpha_i^* - \alpha_i < (1 - \alpha_i)c_1.$$

See Appendix A for technical details. The intuition is as follows: The agent avoids large prediction errors, thus the location of the point predictions follows the belief distribution. The more asymmetric a_i and b_i are, the more the agent is incentivized to report a point prediction in the tail of the distribution. More uncertainty is reflected in a wider spread of point predictions.

As the agent cannot be punished for extreme forecast errors beyond having probability zero of receiving the reward, small initial endowments incentivize neglecting the tails of the distribution. To avoid eliciting distorted quantiles, the initial endowment e has to be chosen large enough. Point (i) of Theorem 1 shows that the arising distortion vanishes with increasing initial endowment.

Point (ii) of Theorem 1 allows to bound the error in terms of the actually reported quantile level. Given a point forecast x_i^*, one can form an assumption about the subjective probability of extreme forecast errors, and obtain bounds on the actually reported quantile level α_i^*. The lower bound $-\alpha_i c_2$ arises if the left tail is neglected, and the reported point forecast is too low and corresponds to a lower quantile level α_i^*. The upper bound $(1 - \alpha_i)c_1$ arises if the right tail is neglected, and the reported point forecast corresponds to a higher quantile level α_i^*. Those are worst case bounds. In applications, distributions are often unbounded in both tails, and the arising biases cancel out partly. See the numerical simulations in Section 2.4 for an illustration.

Here, we propose to sum up the positive individual scores of each point prediction. Similar

results can be obtained by summing up the unrestricted scores of each prediction at the cost of small endowments in one prediction possibly distorting other predictions.

We assume the existence of a strictly positive density for convenience only. For partially constant CDFs and discrete measures the quantile is set-valued and the respective point prediction is an element of the set. For an example with a discrete distribution, consider an agent who is sure that the outcome will be 0. In this case, all quantiles of the belief distribution are 0 and the agent would indeed issue 0 for any point prediction.

Under bounded support, Theorem 1 implies a multiple quantile version of the well-known result that asymmetric linear loss functions with binarized scores are incentive compatible for a quantile (e.g., Schlag and van der Weele, 2013).

Proposition 2 (bounded support)**.** *If the agent has a subjective probability with bounded support of length B and $e > B\max(a_1, b_1, \ldots, a_k, b_k)$,the mechanism described above is incentive compatible for the quantiles with levels $\alpha_1, \ldots, \alpha_n$.*

The minimum and maximum of a distribution are generally not elicitable (Bellini and Bignozzi, 2015). However, the following proposition shows that they are essentially elicitable in the sense that they can be approximated by extreme quantiles. We propose to elicit a set of extreme quantile levels (e.g., $\alpha = (0.1, 0.01, 0.001)$).

Proposition 3 (minimum)**.** *For large b_i and large $\frac{e}{b_i}$, the mechanism above is essentially incentive compatible for the minimum of the distribution.*

Note that for subjective probabilities with infinite support the minimum may be $-\infty$ in which case the best responses also diverges. Analogously the maximum can be approximated with levels close to one (e.g., $\alpha = (0.9, 0.99, 0.999)$). See Appendix A for details.

2.4 Numerical Study

In this subsection we investigate two issues. First, we analyse how risk aversion would distort MPP in the absence of binarized incentives. Second, we investigate the adequate choice of the initial endowment e for MPP with binarized incentives.

Throughout, we numerically compute the MPP reports based on three point predictions that elicit the 0.25, 0.5, and 0.75-quantile. The results can be seen in Figure 1. We consider agents with three different subjective probability distributions, as illustrated in the first row through their densities. The agents hold a uni-modal, a multi-modal, and an asymmetric belief distribution, respectively. The second and third row depict the optimal reports for the three point predictions, where the dashed line shows the true values of the target quantiles.

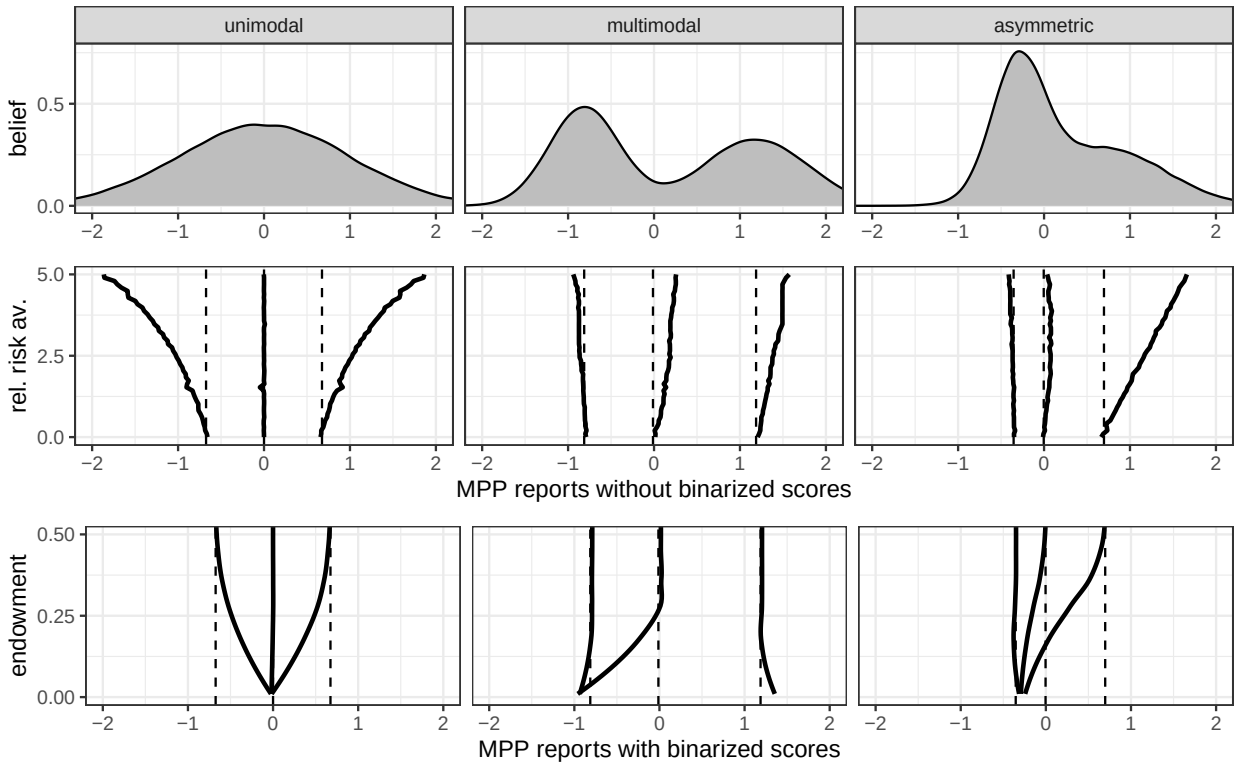


Figure 1: **Numerical solutions to MPP.** Each column assumes a different belief distribution. The belief's density is illustrated in the first row. The second row shows the optimal response of a risk averse agent to MPP without binarized scores for different levels of relative risk aversion. The third row shows the best response of a risk averse agent to MPP with binarized scores for different initial endowments. The dashed lines show the true quantile values.

In particular, the second row shows the best response for preferences with constant relative risk aversion if MPP is administered without binarizing, paying out the score $s_{(0.25,0.5,0.75)}$ of Equation (3) in profits. In this case, a risk neutral agent reports the quantiles correctly. A risk averse agent, however, puts additional weight on extreme forecast errors and reports an overly large spread. The results suggest that the distortion can be considerable for moderate coefficients of relative risk aversion. Also note, that the median is distorted by risk-aversion under asymmetric distributions only. Finally, the asymmetric distribution illustrates that risk aversion leads to over-reporting of heavy tails. Harrison *et al.* (2017) provide a related discussion on the effects of risk-aversion on the elicitation of IP without binarized scores. They find that the distortions from risk attitudes are less severe than for eliciting event probabilities.

The third row considers MPP with binarized scores as proposed in Section 2.3 for a risk averse agent for different initial endowments e. The penalty terms a_i, b_i for each quantile are normalized

to one. The endowments e are given as the probability they cover. The value 0.5 means that the endowment is equal to the length of the 0.5 central confidence interval (or the difference between the 0.75- and 0.25-quantile). Proposition 2 states that the optimal response is identical to the elicited quantiles, if e covers the whole support. If the support is unbounded, as here, Theorem 1 states that the best response converges to the elicited quantiles for large e. An elicitor has to choose the initial endowment large enough. As illustrated in the last row of Figure 1, for small initial endowments the responses are drawn towards the modal interval, neglecting the tails of the distribution and therefore under-reporting uncertainty for unimodal distributions.

In applications, the choice of the initial endowment e is crucial, as small values induce biases in the optimal response as shown here. For practical implementation, the results suggest the following rule of thumb: The initial endowment e should be as least as large as the length of the 50% central confidence interval of the subjective belief distribution. For all three distributions considered here, this choice essentially delivers the targeted quantiles.

2.5 Partial Identification and Bounds

The quantile reports $x = (x_1, \ldots, x_n)$ from MPP for levels $\alpha = (\alpha_1, \ldots, \alpha_n)$ allow to infer about the subjective probability $\mathbb{P}$, that $\alpha_i = \mathbb{P}(y \leq x_i)$ for $i = 1, \ldots, n$. This coincides with the information obtained when eliciting IP with thresholds $c = x$ on the $n + 1$ intervals $(-\infty, x_1], \ldots, (x_n, \infty)$. By design, MPP allow to fix the probability levels α_i and IP allow to fix the thresholds c_i. Both essentially reveal the same amount of information about the subjective probability distribution.

The subjective probability distribution $\mathbb{P}$ is only partially identified. However, we obtain the set of distributions that is consistent with the elicited CDF points and bounds on properties of interest (compare Bissonnette and de Bresser, 2018; Engelberg *et al.*, 2009).

It is a core feature of the quantile reports that they are automatically distributed over the mass of the distribution as wished by the elicitor. In elicitation of IP, the support of the subjective distribution might be located outside of the elicited intervals, or the whole support might be located in one single interval.

Consider an example that is loosely based on the Survey of Consumer Expectations (Armantier *et al.*, 2017) by the Federal Reserve Bank of New York that elicits the expected percentage change of earnings in one year.[8] We consider three individuals with different beliefs in Figure 2. The first row depicts the belief density. The second and third row depict the true CDF in black. The dots mark CDF points identified by MPP and IP respectively. The grey area illustrates the set of

[8]The actual thresholds are $c = (-12, -8, -4, -2, 0, 2, 4, 8, 12)$. For another example of unincentivized elicitation on pre-defined intervals see the inflation and output growth forecasts in the Survey of Professional Forecasters (Croushore, 1993).

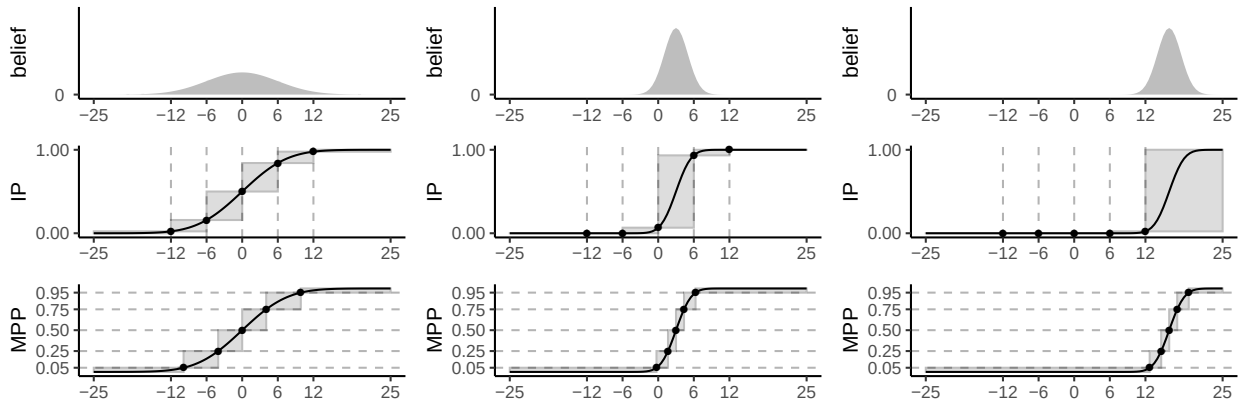


Figure 2: **Elicited CDF points and the space of consistent CDFs for three examples.** Each column contains the true belief as pdf in the first row, and the elicited CDF points and the space of consistent CDFs in the second and third row for IP with $c = (-12, -6, 0, 6, 12)$ and MPP with $\alpha = (0.05, 0.25, 0.5, 0.75, 0.95)$ respectively. The black line depicts the true CDF in the second and third row. The dashed lines illustrate the interval thresholds for IP and the probability levels for MPP.

consistent CDFs.[9] The individual in the first column perceives significant uncertainty about her earnings but expects on average no changes. The IP thresholds are well suited to identify this belief. The second individual is more optimistic and more certain about future earnings. With IP the elicitor obtains little information about the CDF shape in the interval $[0, 6]$. Asking for MPP the elicitor obtains no information about the tails of the CDF. The third individual expects more than 12% income rise. IP provide essentially no information about the expectations beyond 12%, as illustrated by the large grey area, while MPP adapts flexibly and remains informative about the belief. In summary, the examples show that MPP can adapt more flexibly to heterogeneous beliefs, but cannot bound the tails of the CDF beyond the elicited levels without additional assumptions. Proposition 3, however, guarantees that the extreme points of the distribution can be approximated with extreme quantiles.

Let us consider bounding other properties, e.g. the mean, median or interquartile range. Any property that is monotone with respect to stochastic dominance (e.g., the mean or the median), can be bounded easily by the respective property for the CDF that dominates and is dominated by all other consistent CDFs. In the following we analyze which method provides sharper bounds. We assume that the support of the distribution is bounded.[10] For the mean property, it follows

[9]For the ease of exposition we abstract from rounding or the bounds derived in Theorem 1. The more general framework would allow to infer that $\mathbb{P}(y \leq x_i) \in [\alpha_i - c_i, \alpha_i + C_i]$ for suitable bounds c_i and C_i.

[10]For unbounded support, the boarder parts of the CDF would be unbounded and so would be most properties (e.g., the mean).

from linearity arguments that the bounds after eliciting IP and MPP are equally sharp. The median property is uniquely identified by MPP, whereas IP can only identify the interval in which the median lies. The mode property cannot be bounded without further assumptions by either method.[11]

Generally, it is harder to find valid bounds on measures of dispersion.[12] Conveniently, elicitation of MPP identifies the interquantile range between the elicited quantiles. If, for example, the 0.25 and 0.75-quantile are elicited, MPP identifies the 0.5 confidence interval.

3 Experimental Design

In a laboratory experiment we compare belief elicitation with MPP to the standard procedure of eliciting beliefs via IP. The experiment was based on OTree (Chen *et al.*, 2016) and was conducted at the Frankfurt Laboratory for Experimental Economic Research (FLEX) between December 2017 and December 2019. In total we recruited 327 subjects through the online system ORSEE (Greiner *et al.*, 2004) divided into 16 sessions with $8 - 24$ participants in each session. Our subject pool consisted of German undergraduate and graduate students with a median age of 23 years, where 58% have taken at least one university level statistics course.

3.1 Main Experimental Treatments

The main experiment comprised four (2×2) different treatments as illustrated in Figure 3a. Each subject was randomly assigned to one of the four treatments. In two of the treatments we elicited IP using the binarized QSR as described in Section 2.1. In the remaining two treatments we elicited MPP using binarized linear incentives as described in Section 2.3.

We analyze how the heterogeneity of beliefs influences the comparison. The second treatment variation was the information updates that were presented to the subjects during the experiment: Participants were either exposed to a weak or strong information update.

3.2 Domains

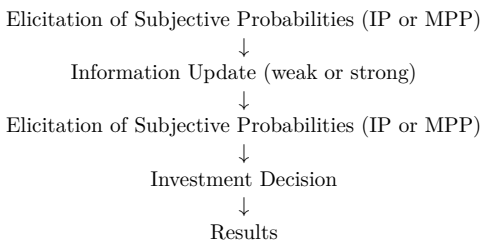
We elicited subjective probabilities for five different fields of application. The order in which these domains were presented varied randomly between subjects to avoid order effects. Figure

[11] Note that Engelberg *et al.* (2009) require the additional assumption that the mode lies in the interval with the highest probability to provide partial identification.

[12] Dillon (2016) provides results for bounding the mean and variance simultaneously allowing for imprecisely reported interval probabilities.

IPweak (n=74)	**IPstrong** (n=64)
IP with QRS	IP with QRS
&	&
Weak Information Update	Strong Information Update
MPPweak (n=70)	**MPPstrong** (n=74)
MPP with linear scores	MPP with linear scores
&	&
Weak Information Update	Strong Information Update

(a) Experimental treatments



(b) Elicitation procedure within each domain

Figure 3: **Experimental Design.** This table summarizes the experimental treatments and elicitation procedure within each of the five domains

3b illustrates the timeline within each domain. An explanation screen was shown first. Then, the subjective probabilities were elicited. Next, subjects were given an information update, which provided either weak or strong information. Afterwards, we elicited the subjects' beliefs again. After the second round of belief elicitation, we elicited if participants were willing to pay a given amount of credits[13] in exchange for receiving the uncertain outcome y in credits. Participants could either accept or reject the offer. This one-shot decision gives insights into the location of the mean within a participant's subjective distribution. We frame this measure as a common investment decision and use it as a consistency check of elicited beliefs with a subsequent decision.

For a summary of the domains and the used information updates see Table 2. In the *dice* domain, the uncertain outcome was the sum of ten virtual dice rolls. In the *dots* domain it was the amount of dots shown on the computer screen. The true value was randomly chosen between 150 and 250. The *number* domain exhibited a randomly chosen number between 0 and 99. In the more complicated *ball* domain, we virtually presented an urn with 60 balls, numbered from $1-60$. As a second step, we blindly drew 10 out of these 60 balls without replacement and put them into another urn, without showing the result to the participants. Next, we drew three balls with replacement from the second urn and showed them to the participants before putting them back into the urn. Finally, the participants were asked to estimate the number on the fourth ball that was drawn from that second urn. In the *temperature* domain subjects were asked to report on the highest temperature in Frankfurt of a particular day in 2016.

Whereas the *ball* domain should induce asymmetric beliefs, the *dice* domain and *number* domain should induce symmetric beliefs. These domains are inherently random. In comparison, the *dots* domain is skill/effort-based. Finally, the *temperature* domain is closest to a real life forecasting task. We chose domains that vary with respect to their level of complexity in order to provide

[13]Offers were uniformly distributed around the unconditional mean of the uncertain outcome y, i.e. on the interval $[0.9\mathbb{E}[y], 1.1\mathbb{E}[y]]$.

Domain	Weak Update	Strong Update
Dice: Sum of ten dice rolls	Two dices	Six dices
Dots: Number of dots	Comparison with small rectangle	Comparison with similar sized rectangle
Number: Random number (0-99)	Second digit	First digit
Ball: Urn draw of numbered balls (1-60)	One additional draw	Six additional draws
Temperature: Past temperature in Frankfurt	Temperature one week before	Temperature one week & one day before

Table 2: **Information Updates.** This table shows the information updates that were given to participants before the second round of belief elicitation.

insights on the performance of our method in a wide array of potential applications.

3.3 Belief Elicitation in Detail

With both methods, we elicited three CDF points. In the *IP treatments*, we elicited the four probabilities simultaneously:

> *What do you think is the percentage chance that y is smaller than* c_1*?*
> *What do you think is the percentage chance that y is between* c_1 *and* c_2*?*
> *What do you think is the percentage chance that y is between* c_2 *and* c_3*?*
> *What do you think is the percentage chance that y is larger than* c_3*?*

The thresholds (c_1 to c_3) were fixed within each domain and were chosen to divide the attainable values into four equally sized segments. The amount of credits earned was calculated using a rescaled version of Equation (1) such that each credit represented a 0.5 percentage chance of winning an extra 10€ if that round of belief elicitation was drawn for payoff in the end.

In the *MPP treatments*, we elicited three point forecasts simultaneously:

> *What do you say is y,. . .*
> *. . . if underestimation is three times less costly than overestimation?*
> *. . . if overestimation and underestimation are equally costly?*
> *. . . if overestimation is three times less costly than understimation?*

For all three questions, the credits earned were calculated using the linear scoring rule from Equation (2). Varying costs for over- and underestimation is equivalent to choosing different parameters for a and b. Here, we chose the parameters for a and b such that the 0.25-quantile, the 0.5-quantile, and the 0.75-quantile were elicited. We tried to provide equal stakes in all treatments. Resulting average earnings were 10.40€ and 11.10€ in the MPP, and 11.40€ and 11.70€ in the IP treatment, for the weak and strong updates, respectively. For details on the experiment and instructions, see the online supplement.

4 Results

There are two major applications of subjective probabilities in economics. First, they can be used as forecasts or prior distributions in Bayesian analysis (Garthwaite *et al.*, 2005; O'Hagan *et al.*, 2006). Second, they are used as input in decision-theoretic models to explain behavior under uncertainty (Manski, 2004). We denote the two different applications as *forecasting* and *economic modelling*. Forecasting strives to accurately describe the actual distribution of the unknown outcome, whereas in economic modelling the subjective probabilities should accurately represent the belief of an agent – which may differ from the actual distribution of the outcome. Hence, depending on the application at hand, different evaluation criteria are important.

In Section 4.1, we evaluate potential biases of the elicited subjective probability distributions compared to the Bayes distribution and the actual outcome. In Section 4.2, we focus on the accuracy of the elicited subjective probability distributions. We use average linear scores/loss functions to judge the predictive value of the elicited quantiles for the true realization. In Section 4.3, we assess the consistency of the elicited beliefs with a subsequent decision, which is an important property for economic modelling. Subsequently, in Section 4.4, we zoom into the issue of anchoring and provide evidence for a possible channel that drives our results. Finally, in Section 4.5, we report results on the applicability of the two methods.

4.1 Forecasting: Biases

We begin with arguably the most common approach (Schlag *et al.*, 2015, Table 1) for the evaluation of potential biases in subjective probabilities: The comparison with Bayesian probabilities. Note that this kind of analysis is not possible for the temperature or dots domain, where no natural Bayesian probabilities are available.

Subsequently, we compare the predictive distributions to the actual outcome y. Since we do not rely on Bayesian predictions for this, we are able to include the dots and temperature domain in the analysis. We understand the outcome y and the predictive distribution $\mathbb{P}$ as realizations in a joint probability space[14], which allows us to analyze forecasting performance without unduly strong assumptions on the data generating process. This more widely applicable approach comes at the cost of additional noise, as the comparison substitutes the unknown distribution with a draw from this distribution.

[14]The idea goes back to DeGroot and Fienberg (1983); Murphy and Winkler (1987). See Gneiting and Katzfuss (2014) for a recent review.

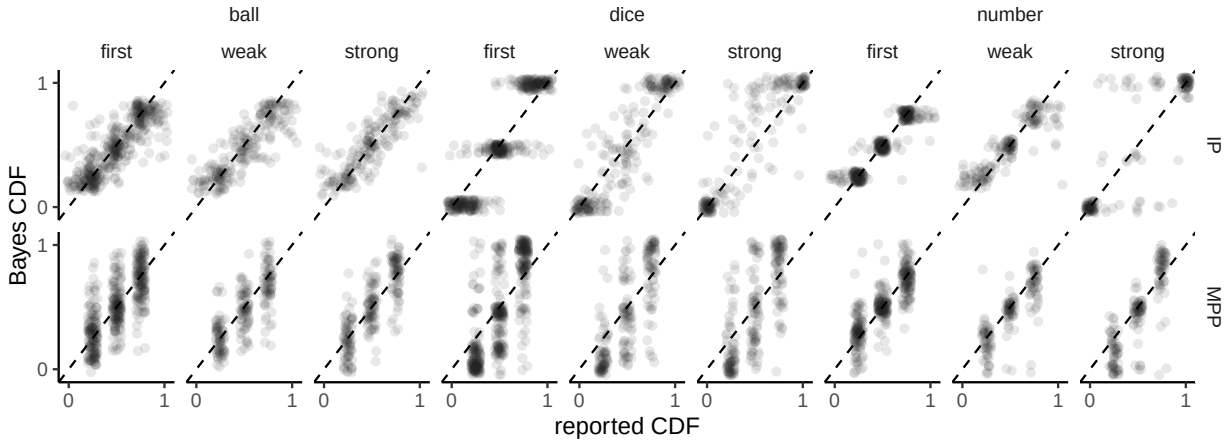


Figure 4: **Comparison between reports and Bayesian distributions.** For the IP treatment, the reported probabilities are plotted against the Bayes probabilities at the elicited threshold levels. For the MPP treatment, the elicited quantile levels are plotted against the reported quantile levels of the Bayes distribution. The Bayes distribution is computed based on the individual and time specific information. Dashed lines depict perfect Bayesian reports. Additional noise in form of a uniform distribution with support $[-0.05, 0.05]$ was added.

4.1.1 Comparison of reports with Bayesian distributions

Figure 4 depicts a comparison between the reported subjective CDF values and Bayesian CDF values. For both, MPP and IP, we observe that reports deviate from the Bayesian distributions. Table 6 in Appendix B shows that average biases range from 0% to at most 12.8% and are partly significant. Overall, the visual evidence suggests a strong correlation between the reported subjective and the Bayesian CDF points.

4.1.2 Comparison of predictive distributions with Bayesian distributions

For a more in-depth comparison between MPP and IP, we consider the distributions obtained by fitting parametric distributions to the elicited CDF values. We call the resulting distribution a *predictive distribution*. We consider four commonly applied procedures that are illustrated in Figure 5. The *atoms* distribution assumes a discrete distribution, where the mass is located at the midpoint of each elicited interval[15] (compare Hill, 2010; Lahiri and Teigland, 1987; Lahiri *et al.*, 1988). The *pl* distribution assumes a *piecewise linear* CDF (compare Diebold *et al.*, 1999; Guiso *et al.*, 2002; Zarnowitz and Lambros, 1987), which is equivalent to a piecewise constant density.

[15]The outer intervals are assumed to have the same length as the neighboring interval.

[htp]

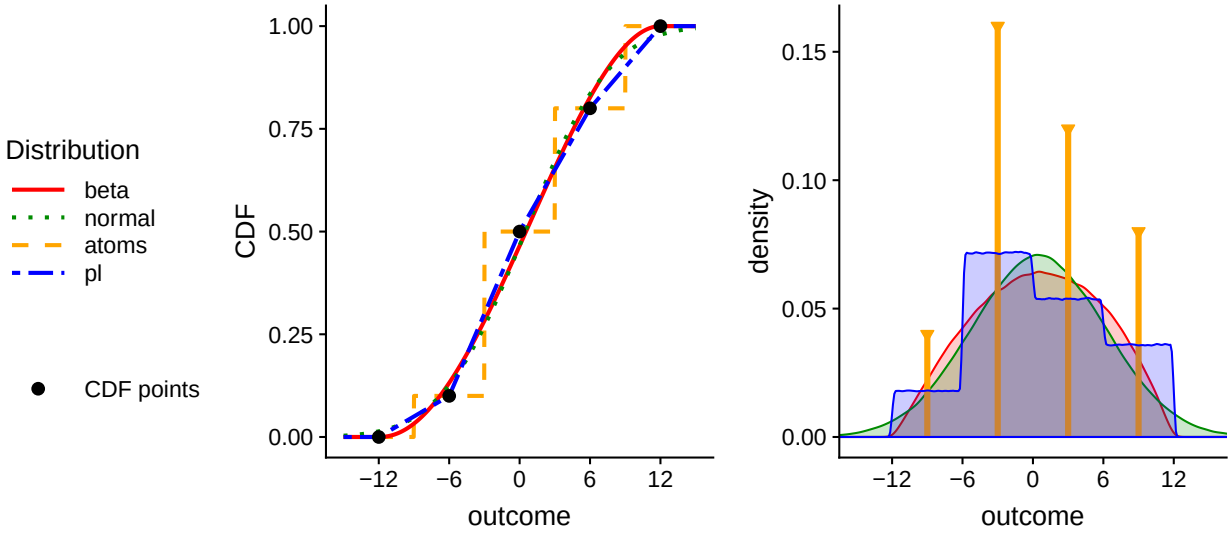


Figure 5: **Elicited CDF points and parametric distributions.** The fitted distributions are illustrated as CDFs (left plot) and densities (right plot).

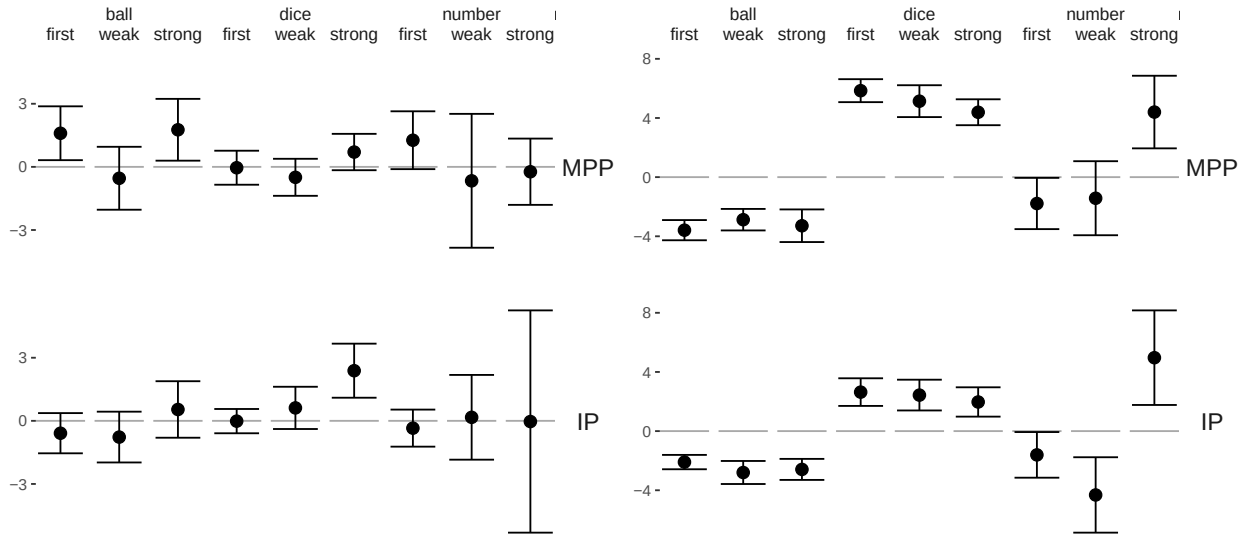
Further, we fit predictive distributions by minimizing

$$\inf_{\theta} \sum_{i} (F(x_i; \theta) - \alpha_i)^2,$$

where x_i and α_i are obtained by the reports and $F(\cdot, \theta)$ is the CDF of the distribution for the parameter θ. In particular, we fit a *normal* distribution (compare Boero *et al.*, 2015; Clements, 2014; Dominitz and Manski, 2011; Giordani and Söderlind, 2003; Gouret and Hollard, 2011; Hurd *et al.*, 2011) and a *beta* distribution (compare Delavande, 2008; Engelberg *et al.*, 2009; Manski and Neri, 2013; Neri, 2015).[16]

Figure 5 illustrates several stylized facts about the parametric assumptions: Measures of central tendency, like the mean and median, are relatively robust to the choice of distributional assumptions. Measures of dispersion, like interquantile ranges or the variance, depend heavily on the chosen fit, where the *atoms* distribution always has a lower variance then the *pl* distribution. This simple observation challenges the common approach to construct variance estimates of the subjective distribution with only a small number of elicited CDF points. In particular, the assumption invoked for the *atoms* distribution, putting all mass at the midpoints of the intervals, potentially

[16]Other distributional fits in the literature that we do not consider are the log-normal distribution (used for income expectations in Dominitz, 2001; Dominitz and Manski, 1997; McKenzie *et al.*, 2013), triangular densities (Kaufmann and Pistaferri, 2009) and cubic-splines (Bellemare *et al.*, 2012).



(a) **Mean.** (b) **Standard deviation.**

Figure 6: **Comparison between fitted distributions and Bayesian distributions.** For the mean plot, the target variable is $Z := mean(\mathbb{P}) - mean(\mathbb{P}_{bayes})$, where $\mathbb{P}$ denotes the fitted predictive distribution and $\mathbb{P}_{bayes}$ the Bayesian distribution. For the standard deviation plot the target variable is $Z := sd(\mathbb{P}) - sd(\mathbb{P}_{bayes})$. Throughout, the error bars show 95% confidence intervals. Participants are pooled in the first round of elicitation (*first*), and distinguished after receiving the information update (*weak* and *strong*).

underestimates the uncertainty, if the true subjective belief distribution is not atomic.

We apply different fits for each domain. In particular, we fix the beta distribution for the ball and dots domains as they are bounded (and the ball domain often exhibits asymmetric distributions), and the normal distribution for the remaining domains as it does not require that we externally choose the support of the distribution. In Appendix B, we provide results of all fitting methods described here. All results are based on simple averages or differences in means.

Figure 6 depicts the average difference between the predictive distributions and the Bayesian distributions in terms of the mean and standard deviation. The left panel (a) shows that the position of the predictive distributions are largely unbiased. The right panel (b) provides evidence that the uncertainty is consistently misjudged. While we find large uncertainty estimates for the dice domain, both elicitation procedures provide overly confident predictive distributions for the ball domain[17].

[17]A fact referred to as overconfidence, or more specifically as overprecision (compare Alpert and Raiffa, 1982; Brenner *et al.*, 1996; Fox and Clemen, 2005; Haran *et al.*, 2010; Lichtenstein and Fischhoff, 1977).

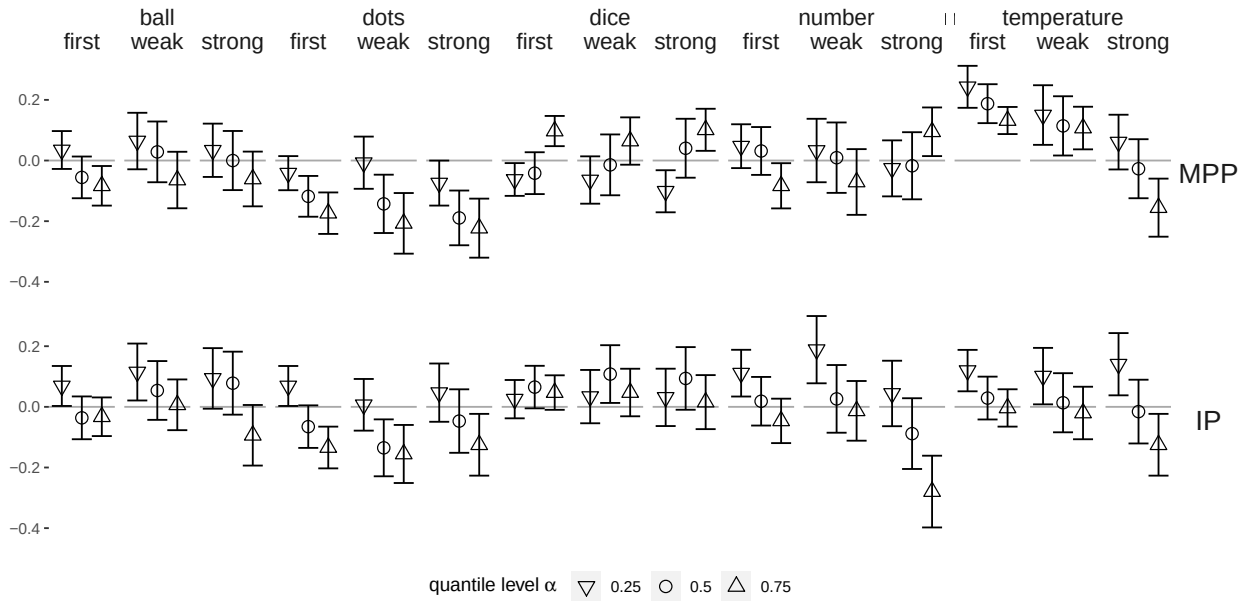


Figure 7: **Ratio of over-predictions compared to quantile level.** The target variable is $Z := \mathbb{1}(y < q_\alpha(\mathbb{P})) - \alpha$, where $q_\alpha(\mathbb{P})$ denotes the α-quantile of the fitted predictive distribution. The plot depicts average ratio of overpredictions compared to expected ratio. Values over zero indicate a tendency to overpredict the outcome, values below a tendency to underpredict the outcome. Throughout, the error bars show 95% confidence intervals. Participants are pooled in the first round of elicitation (*first*), and distinguished after receiving the information update (*weak* and *strong*).

4.1.3 Comparison of predictive distributions with realized outcomes

We now turn to the analysis of the bias of the predictive distributions compared to the realized outcome. If the predictive distributions are unbiased, the α quantile exceeds the realized outcome with the probability α. Figure 7 depicts the ratio of overpredictions compared to the expected ratio. A coefficient of zero indicates an unbiased or calibrated forecast. We can reject unbiased distributions in the dots and temperature domain for the MPP treatment. The strong information update erases a large part of the bias in the temperature domain. The IP treatment shows some evidence against an unbiased distribution in the dots domain, where the 0.75-quantile tends to be underpredicted. The observed pattern is consistent with the intervals providing helpful anchors in the first round that alleviate biases in perception. We will have a closer look at this explanation in section 4.4. The results after the information update suggest that the bias in the MPP treatment can be reduced by additional information.

At this point it is worth noting that the analysis above considers different elicitation and fit methods for a random participant and a specific domain. As each participant encountered each

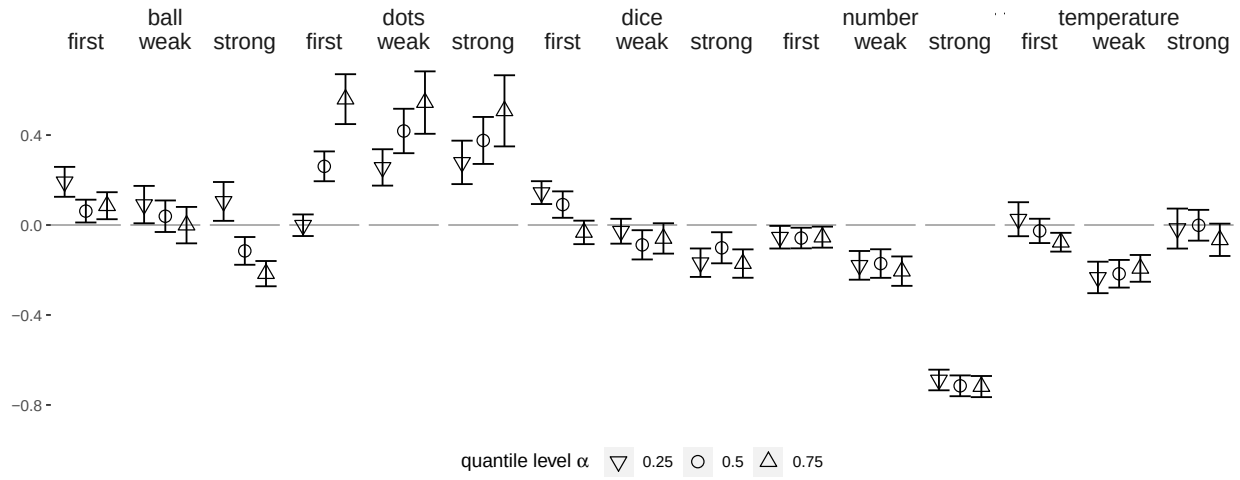


Figure 8: **Difference in accuracy.** The target variable is denoted as $Z := l_\alpha(q_\alpha(\mathbb{P}), y)$ with the linear asymmetric loss function $l_\alpha(x, y) = |x - y||\mathbb{1}(x > y) - \alpha|$, where $q_\alpha(\mathbb{P})$ denotes the α-quantile of the fitted predictive distribution. The plot depicts differences in average linear loss between the two procedures divided by the average in the IP treatment, so that negative values indicate superior accuracy of MPP.

domain only once, we cannot evaluate the calibration of one specific participant in one specific domain, nor can we make a statement for real economic decisions without additional assumptions. The observed biases are, for example, perfectly consistent with participants' beliefs following some rule of thumb that is unbiased across the relevant decisions in daily life, while being biased at the specific tasks encountered in our experiment.

4.2 Forecasting: Accuracy

Unbiasedness and accuracy are different concepts, which do not necessarily agree as to which elicitation method performs best (Gneiting and Katzfuss, 2014). An elicitation method can provide perfectly unbiased distributions that provide little useful information because they lack sharpness. We now consider the accuracy of the predictive distributions.

The accuracy of a property of the predictive distribution can be measured with proper scoring rules (Gneiting, 2011). The average linear asymmetric absolute error between the 0.25-quantile, the median, and the 0.75-quantile of the predictive distribution and the realized outcome is depicted in Figure 8. Before the information update both methods seem to provide equally accurate predictive distributions across almost all quantiles, except for the dots domain, where the bias analyzed above translates into higher errors for MPP. After the information update, the evidence suggests that MPP provide at least equally accurate predictive distributions for all but the dots domain across

	(1) 0.25-quantile	(2) Median	(3) 0.75-quantile	(4) Consistency
MPP * Weak info	−0.462	7.317	7.130	0.017
	(11.858)	(12.575)	(12.221)	(0.036)
MPP * Strong info	13.012	22.691*	45.289***	0.119***
	(13.131)	(13.304)	(14.318)	(0.041)
MPP	−10.698	−8.856	−11.484	−0.028
	(8.778)	(8.477)	(8.652)	(0.026)
Weak information	18.018**	8.230	5.897	0.020
	(8.188)	(8.352)	(8.288)	(0.025)
Strong information	55.619***	41.991***	31.134***	0.009
	(9.339)	(9.177)	(10.177)	(0.030)
Constant	257.826***	260.489***	262.141***	0.648***
	(5.908)	(5.574)	(6.100)	(0.019)

Table 3: **Accuracy and consistency on information heterogeneity.** Dependent variables in columns (1-3) are denoted as $Z := l_\alpha(q_\alpha(\mathbb{P}), y)$ with the linear asymmetric loss function $l_\alpha(x, y) = |x - y||\mathbb{1}(x > y) - \alpha|$, where $q_\alpha(\mathbb{P})$ denotes the α-quantile of the fitted predictive distribution. Realized losses are ranked within each domain. Higher ranks indicate more accurate predictions. The dependent variable in column (4) is the consistency between fitted predictive distribution and investment decision, which is given if the mean of the fitted predictive distribution is above the offer and the offer was accepted or if the mean is below and the offer was rejected. Otherwise, the responses are inconsistent. Standard errors are clustered at the individual level. *p<0.1; **p<0.05; ***p<0.01

all quantiles. In most domains, the difference in accuracy is more favorable for MPP under more heterogeneous information (after the strong information update).

We test whether the strong information update indeed improved the accuracy of MPP more than the accuracy of IP. Columns $1-3$ of Table 3 show that both methods benefit from the information update, but the improvement is stronger for MPP reports. A weak information update, as opposed to no information, improves the relative performance of MPP for the 0.25-quantile and the 0.75-quantile, albeit these effects are not significant. A strong information update benefits the accuracy of distributions based on MPP reports significantly more for the median and the 0.75-quantile. The relative performance increase for the 0.25-quantile is not statistically significant. Arguably, IP could potentially benefit here if thresholds adapted to the current information.

To summarize, the bias in the dots domain led to less accurate beliefs with MPP. In all other domains, MPP provided at least as accurate beliefs. Averaged over all five applications, additional

information improved the accuracy of MPP more than the accuracy of IP. This is consistent with the argument that pre-defined intervals act as anchors, which improves accuracy for uninformed participants. Under increasing information those anchors are less helpful (or even distorting). Consequently, eliciting MPP should be considered whenever it is challenging to construct intervals that are uniformly adequate across all participants.

4.3 Consistency with Investment Decision

For economic modelling an elicitation mechanism should provide accurate evidence on the subjective belief of an agent, not necessarily on the actual distribution of the outcome. A natural benchmark arises if we use the subjective probabilities to predict behavior based on a decision-theoretic model. A well-behaved elicitation mechanism provides *accurate predictions of individual behaviour in economic decisions.*

As described in Section 3.2, we confronted participants with an investment decision by offering them an amount of credits in exchange for receiving the uncertain outcome y in credits after the second round of belief elicitation. We analyze if the investment decision is consistent with the mean of the predictive distribution. Elicited distributions and subsequent action are denoted as consistent if the mean of the predictive distribution is above (below) the offer and the offer was accepted (rejected). Otherwise, the observed actions are denoted as inconsistent.[18]

Figure 9 depicts the difference in average consistency. On average both elicitation methods provide a high consistency rate of about 50% to 85% (compare Figure 14 in Appendix B). No consistent pattern arises after the weak information update. After the strong information update, the beliefs elicited with MPP were 5% to 20% more often consistent with the investment decision. Column (4) of Table 3 confirms the positive effect of the strong information update for MPP elicitation.

Again, anchoring is a possible channel for MPP outperforming IP as IP reports may be distorted towards the anchors provided. Those anchors have a larger impact in the strong information treatment, where beliefs are heterogeneous and therefore the provided bins less suitable on average. If the anchors distorted the reports more than the subsequent investment decision, the beliefs are more likely to be inconsistent, as observed in the data.

[18]Several studies find that belief elicitation can influence subsequent action (Croson, 1999, 2000; Erev *et al.*, 1993; Gächter and Renner, 2010; Rutström and Wilcox, 2009), whereas others could not detect this effect (Costa-Gomes and Weizsäcker, 2008; Nyarko and Schotter, 2002; Wilcox, 2006). Our setting does not allow to test if elicited beliefs are consistent with behavior that would have occurred in the absence of the elicitation procedure. We find, however, no significant deviations in the investment decision between the two different elicitation procedures.

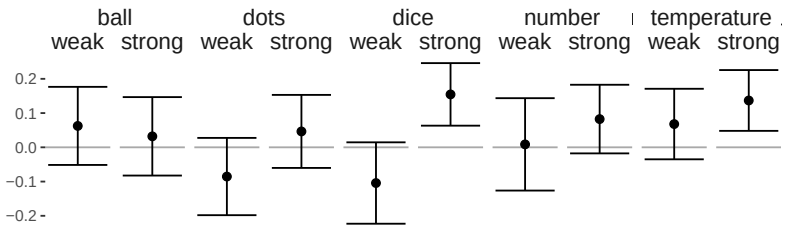


Figure 9: **Difference in consistency with investment decision.** The plot depicts differences in the ratio of consistent choices, where positive values indicate superior consistency of MPP. The distribution and the investment decision are consistent, if the mean of the fitted predictive distribution is above the offer and the offer was accepted or if the mean is below and the offer was rejected. Otherwise, the responses are inconsistent.

We find that MPP is more predictive of subsequent behavior in an experimental game. Naturally, it remains an open question if similar results hold for other potentially more consequential economic decisions. The results presented here can only be regarded as first evidence that MPP might be a suitable tool for economic modelling.

4.4 Anchoring in IP

Based on the previous findings, we conducted an additional experiment, in which we took a deeper look at the influence of the pre-defined thresholds for IP. We confronted 45 participants, who did not participate in the main experiment, with an instance from the dot domain containing 225 dots. In this version, we elicited IP under varying thresholds (c_1 to c_3) between subjects. We manipulated the intervals in two dimensions, as illustrated in Figure 10: For participants of group *Right*, we shifted all original thresholds by 50 points to the right. In a cross-variation, we reduced the original length of each interval from 50 points to 25 points (*Narrow*) while keeping them centered at 200 and 250 respectively. Otherwise, the description and incentives were identical to the main experiment.

We compute the mean and standard deviation of the reported distributions of this additional experiment, assuming that the mass of the distributions is located at the midpoint of the intervals, and subsequently regress mean and standard deviation on the treatment indicators in Table 4.

The results show that the pre-defined intervals strongly anchor the elicited IP. Shifting the intervals by 50 points to the right, causes a 40 points (80%) increase in the mean of the predictive distributions. Similarly, narrowing the intervals by half, decreases the standard deviation of the elicited distributions by 11 points (70%). We conclude that the intervals provide anchors that influence participants. Consequently, if beliefs are heterogeneous across participants, elicitation of IP suffers from the fact that the fixed intervals influence some of the reports. Generally, an elicitor

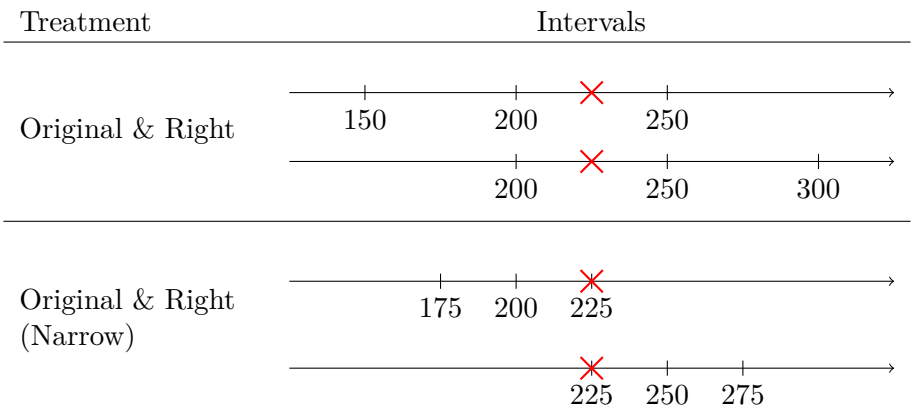


Figure 10: **Intervals in anchoring experiment.** The table depicts the interval thresholds in the four different treatments relative to the true number of dots, 225.

should be aware that the choice of interval thresholds may have a severe impact on the elicited beliefs.

4.5 Applicability: Time and Participant's Perception

Finally, we compare the applicability of the elicitation methods. The participants spent on average 28 minutes on the whole survey, where MPP participants were 9% faster (p-value < 0.01). For the first probability elicitation, IP subjects took on average 96 seconds and MPP subjects were 25% faster (p-value < 0.01). For the last probability elicitation IP took on average 47 seconds and MPP was 11% faster (p-value $= 0.12$). Thus, MPP was faster for unexperienced participants and the advantage reduced when applying the same mechanism multiple times[19].

Table 5 shows the treatment's impact on subjective perception questions elicited after the experiment. On average MPP participants reported that the experiment was less difficult, and stated to have felt less insecure and stressed and more content during the experiment. Overall, the experience seems to be more comfortable with MPP elicitation, an important property for preventing drop-outs and ensuring high take-up rates in panel data sets.

[19]Note that participants in the IP treatment had to report four probabilities whereas participants in the MPP treatment only had to report three point estimates. While we cannot rule out that this explains a large part of the observed difference, we consider the fact that learning reduces the difference as first suggestive evidence against this explanation.

	Mean	Standard deviation
Right	39.6***	2.3
	(7.3)	(2.7)
Narrow	19.0**	-10.6***
	(7.3)	(2.7)
Constant	204.6***	14.4***
	(6.6)	(2.5)
Observations	45	45
R^2	0.5	0.3

Table 4: **Anchoring.** *Right* reports effect estimates of increasing the pre-specified intervals by 50 points on the mean and standard deviation of the elicited distributions. *Narrow* reports the effect estimates of reducing the length of the intervals by half. *Note: *p<0.1; **p<0.05; ***p<0.01*

	MPP	IP	Difference	p-value
How mentally exhausting did you find the experiment?	**0.09**	0.31	-0.22	0.13
How difficult did you find your tasks in the experiment?	**0.02**	0.28	-0.26	0.06
How insecure did you feel during the experiment?	**-0.07**	0.36	-0.43	0.01
How stressed did you feel during the experiment?	**-0.77**	-0.47	-0.29	0.07
How bored did you feel during the experiment?	-0.97	**-1.09**	0.12	0.38
How relaxed did you feel during the experiment?	**0.51**	0.44	0.08	0.58
How content did you feel during the experiment?	**0.51**	0.17	0.35	0.01

Table 5: **Perception of experiment.** The responses are encoded by *Not at all* (-2), *Almost not at all* (-1), *Not sure/Neutral* (0), *Quite a bit* (1), and *Very much* (2). The MPP and IP columns indicate average values across all participants ($n = 282$). For each question, the better performing method is marked in bold letters. The p-values are derived from an unpaired two-sample t-test with two-sided alternative.

5 Discussion

Cognitive abilities to assess beliefs in form of subjective probability distributions were found to be limited (Hogarth, 1975). This renders elicitation burdensome and threatens the validity of probabilistic statements in economic modelling and expert forecasting altogether. Probabilistically sophisticated preferences like subjective expected utility (Savage, 1954) postulate that agents act "as if" they hold a belief in form of a probability distribution. However, choices can be consistent with a probability distribution, while the agent is unable to express this distribution explicitly.

In a general framework for the elicitation of real valued beliefs, we show how to elicit the entire subjective probability distribution, which provides more information than eliciting a discrete distribution on pre-defined intervals.[20] Acknowledging the complexity of reporting and scoring probability distributions, we propose to use MPP with linear incentives instead. Similar to IP, this procedure reveals a finite set of CDF points of the belief distribution. Conveniently, point predictions rely on simple incentives, adapt flexibly to heterogeneous beliefs, and do not influence beliefs by providing anchors through pre-defined intervals.

We provide experimental evidence that pre-defined intervals indeed anchor (and potentially bias) elicited beliefs. Interestingly, we find that the bias acts in two different ways: Higher(lower) interval thresholds lead to higher(lower) mean values of the elicited distribution. Larger(smaller) intervals lead to larger(smaller) uncertainty expressed in the elicited distributions. MPP does not require any pre-defined thresholds and therefore can operate entirely without external anchors. It is possible that point predictions influence one another. To minimize such *internal* anchoring we elicit and score all point predictions simultaneously. It is, however, an open question if the experimental findings of our design are robust with respect to the number and extremity of quantile levels. If external anchors are wished for by the elicitor, a natural way to provide them would be by setting default responses for the MPP reports.

The bounded scores for eliciting the entire distribution or MPP can be of separate interest for expert forecasting. A profit maximizing expert with limited liability (Carroll, 2019; Osband, 1989), can be properly incentivized by bounded scores. In this context, additional complexity should be manageable, and eliciting the full density might be preferable.

The experimental evidence suggests that individuals are able to understand and complete the MPP procedure and to report informative estimates. Applying a wide range of criteria and application domains, we find that both methods have their merits and drawbacks so that an application

[20]Note that a similar extension is possible with the linear quantile scores to elicit all quantiles simultaneously in which case the infinite sum of scores converges to the Continuous Ranked Probability Score (CRPS) as shown in Laio and Tamea (2007). For incentive compatibility the continuous QSR requires bounded densities, the CRPS bounded support.

of one or the other should be evaluated depending on the domain and objective of the elicitation. Under homogeneous beliefs and if anchoring is desired or no concern, IP may be more suitable. The faster elicitation and the more positive self-reported emotional reactions suggest that MPP might be more suitable in long surveys and panels that otherwise could suffer from drop outs.

It is often argued that it is preferable to incentivize belief elicitation.[21] If incentives are infeasible or the incentivized elicitation is prone to adverse effects of stakes and hedging (Armantier and Treich, 2013), researchers commonly rely on unincentivized IP reports. We argue that in this case, MPP with hypothetical payoffs can be useful. In an additional experiment presented in Appendix C, we find no evidence that such hypothetical questions fundamentally change the conclusions of the main experiment, which suggests that MPP is applicable in surveys. In this additional experiment, we further show how extreme quantile reports induce similar responses to asking for the minimum or maximum.

This paper considers probabilistically sophisticated preferences on real-valued outcomes that allow beliefs to be represented by a single probability measure. For a related mechanism that considers ambiguity averse preferences for events see Schmidt (2019).

We abstracted from rounding in this study. However, rounding is a common feature observed for discrete probability questions in surveys and experiments (Bissonnette and de Bresser, 2018; Kleinjans and Soest, 2014; Manski and Molinari, 2010). While we would expect most of the findings on rounding to transfer to point predictions, including its useful aspects of signaling ambiguity or skill, the negative consequences (like systematic biases in the tails) are arguably less severe.

It is possible to ask for the maximum and minimum of the support before eliciting IP. Unfortunately, there exists no proper scoring rule for those properties, which disqualifies this procedure for many applications. Nevertheless, using unincentivized maximum and minimum statements or a single point estimate to construct intervals may improve the performance of belief elicitation using IP.

While theoretically any number of quantile levels could be elicited, eliciting three quantiles (e.g., at the levels of 0.25, 0.50 and 0.75) seems adequate for many applications. It directly identifies a measure of central tendency (the median) and of uncertainty (the interquantile range) without parametric assumptions. At the same time, it requires only a reasonable amount of time and cognitive capacity. If the elicitor is more interested in the tails of the distribution, other specifications of MPP might be more suitable.

[21] Schlag *et al.* (2015) provide a survey over several incentivized belief elicitation methods and their performances, Harrison (2014) compares non-incentivized to incentivized belief elicitation and finds a bias for non-incentivized methods across demographic boards, Blanco *et al.* (2010) consider how incentivized belief reports in the laboratory affect hedging against adverse payoff outcomes, and Gächter and Renner (2010) consider public good games and find that incentivized belief reports increase belief accuracy and influence decisions in these games.

Acknowledgments

We thank the participants of the Grüneburgseminar in Frankfurt, the ACDD in Strasbourg, the MM Research Colloquium 2017 in Hirschegg, the IPP Ideas Crunch in Mainz, and the European Meetings of the Econometric Society in Cologne for valuable comments and discussions. Furthermore, we thank Martin Dufwenberg, Tilmann Gneiting, Glenn Harrison, Simon Heß, Florian Hett, Tanjim Hossain, Bertrand Koebel, Michael Kosfeld, Matthias Schündeln, Ferdinand von Siemens, Johannes Wohlfahrt, Verena Wondratschek, and Basit Zafar for very helpful feedback. We thank the Forschungstopf of the Goethe University Frankfurt for funding the experiment. The work of Patrick Schmidt has been partially funded by the Klaus Tschira Foundation.

References

Alpert, M. and Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic and A. Tversky (eds.), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, pp. 294–305.

Altig, D., Barrero, J. M., Bloom, N., Davis, S. J., Meyer, B. and Parker, N. (2020). Surveying business uncertainty. *Journal of Econometrics*.

Armantier, O., Bruine de Bruin, W., Potter, S., Topa, G., Van Der Klaauw, W. and Zafar, B. (2013). Measuring inflation expectations. *Annual Review of Economics*, **5** (1), 273–301.

—, Topa, G., Van der Klaauw, W. and Zafar, B. (2017). An overview of the survey of consumer expectations. *Economic Policy Review*, **23** (2), 51–72.

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

Bellemare, C., Bissonnette, L. and Kröger, S. (2012). Flexible approximation of subjective expectations using probability questions. *Journal of Business & Economic Statistics*, **30** (1), 125–131.

Bellini, F. and Bignozzi, V. (2015). On elicitable risk measures. *Quantitative Finance*, **15** (5), 725–733.

Benjamin, D. J., Moore, D. A. and Rabin, M. (2017). *Biased Beliefs About Random Samples: Evidence from Two Integrated Experiments*. NBER working paper, National Bureau of Economic Research.

Bissonnette, L. and de Bresser, J. (2018). Eliciting subjective survival curves: Lessons from partial identification. *Journal of Business & Economic Statistics*, **36** (3), 505–515.

Blanco, M., Engelmann, D., Koch, A. K. and Normann, H.-T. (2010). Belief elicitation in experiments: Is there a hedging problem? *Experimental Economics*, **13** (4), 412–438.

Boero, G., Smith, J. and Wallis, K. F. (2015). The measurement and characteristics of professional forecasters' uncertainty. *Journal of Applied Econometrics*, **30** (7), 1029–1046.

Brenner, L. A., Koehler, D. J., Liberman, V. and Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, **65** (3), 212–219.

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

Budescu, D. V. and Du, N. (2007). Coherence and consistency of investors' probability judgments. *Management Science*, **53** (11), 1731–1744.

Carman, K. G. and Kooreman, P. (2014). Probability perceptions and preventive health care. *Journal of Risk and Uncertainty*, **49** (1), 43–71.

Carroll, G. (2019). Robust incentives for information acquisition. *Journal of Economic Theory*, **181**, 382–420.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, **74** (6), 1579–1601.

Chen, D. L., Schonger, M. and Wickens, C. (2016). oTree – an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, **9**, 88–97.

Clements, M. P. (2014). Forecast uncertainty - ex ante and ex post: US inflation and output growth. *Journal of Business & Economic Statistics*, **32** (2), 206–216.

Costa-Gomes, M. A., Huck, S. and Weizsäcker, G. (2014). Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect. *Games and Economic Behavior*, **88**, 298–309.

— and Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies*, **75** (3), 729–762.

Cox, J. C. and Oaxaca, R. L. (1995). Inducing risk-neutral preferences: Further analysis of the data. *Journal of Risk and Uncertainty*, **11** (1), 65–79.

Croson, R. (1999). The disjunction effect and reason-based choice in games. *Organizational Behavior and Human Decision Processes*, **80** (2), 118–133.

— (2000). Thinking like a game theorist: Factors affecting the frequency of equilibrium play. *Journal of Economic Behavior & Organization*, **41** (3), 299–314.

Croushore, D. D. (1993). Introducing: The survey of professional forecasters. *Business Review-Federal Reserve Bank of Philadelphia*, pp. 3–15.

De Mel, S., McKenzie, D. and Woodruff, C. (2008). Returns to capital in microenterprises: Evidence from a field experiment. *The Quarterly Journal of Economics*, **123** (4), 1329–1372.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **32** (1-2), 12–22.

Delavande, A. (2008). Measuring revisions to subjective expectations. *Journal of Risk and Uncertainty*, **36** (1), 43–82.

—, Giné, X. and McKenzie, D. (2011). Measuring subjective expectations in developing countries: A critical review and new evidence. *Journal of Development Economics*, **94** (2), 151–163.

— and Kohler, H.-P. (2009). Subjective expectations in the context of HIV/AIDS in Malawi. *Demographic research*, **20**, 817.

Demuynck, T. (2013). A mechanism for eliciting the mean and quantiles of a random variable. *Economics Letters*, **121** (1), 121–123.

Diebold, F. X., Tay, A. S. and Wallis, K. F. (1999). Evaluating density forecasts of inflation: The Survey of Professional Forecasters. In R. F. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, New York: Oxford University Press, pp. 76–90.

Dillon, B. (2016). Measuring subjective probability distributions, working paper.

Dominitz, J. (2001). Estimation of income expectations models using expectations and realization data. *Journal of Econometrics*, **102** (2), 165–195.

— and Manski, C. F. (1997). Using expectations data to study subjective income expectations. *Journal of the American Statistical Association*, **92** (439), 855–867.

— and — (2011). Measuring and interpreting expectations of equity returns. *Journal of Applied Econometrics*, **26** (3), 352–370.

Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, **30** (2), 163–182.

Engelberg, J., Manski, C. F. and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, **27** (1), 30–41.

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

Fissler, T. and Ziegel, J. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, **44** (4), 1680–1707.

Fox, C. R. and Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, **51** (9), 1417–1432.

Gächter, S. and Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, **13** (3), 364–377.

Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, **100** (470), 680–701.

Giordani, P. and Söderlind, P. (2003). Inflation forecast uncertainty. *European Economic Review*, **47** (6), 1037–1059.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106** (494), 746–762.

— and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, **1**, 125–151.

— and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102** (477), 359–378.

Gouret, F. and Hollard, G. (2011). When Kahneman meets Manski: Using dual systems of reasoning to interpret subjective expectations of equity returns. *Journal of Applied Econometrics*, **26** (3), 371–392.

Greiner, B. *et al.* (2004). The online recruitment system orsee 2.0-a guide for the organization of experiments in economics. *University of Cologne, Working paper series in economics*, **10** (23), 63–104.

Guiso, L., Jappelli, T. and Pistaferri, L. (2002). An empirical analysis of earnings and employment risk. *Journal of Business & Economic Statistics*, **20** (2), 241–253.

Haran, U., Moore, D. A. and Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making*, **5** (7), 467–476.

Harrison, G. W. (2014). Hypothetical surveys or incentivized scoring rules for eliciting subjective belief distributions?, working Paper 2014-05, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.

—, Martínez-Correa, J. and Swarthout, J. T. (2013). Inducing risk neutral preferences with binary lotteries: A reconsideration. *Journal of Economic Behavior & Organization*, **94**, 145–159.

—, — and Swarthout, J. T. (2014). Eliciting subjective probabilities with binary lotteries. *Journal of Economic Behavior & Organization*, **101**, 128–140.

—, —, Swarthout, J. T. and Ulm, E. R. (2015). Eliciting subjective probability distributions with binary lotteries. *Economics Letters*, **127**, 68–71.

—, —, — and Ulm, E. R. (2017). Scoring rules for subjective probability distributions. *Journal of Economic Behavior & Organization*, **134**, 430–448.

Hill, R. V. (2010). Liberalisation and producer price risk: Examining subjective expectations in the ugandan coffee market. *Journal of African Economies*, **19** (4), 433–458.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association*, **70** (350), 271–289.

Holt, C. A. and Smith, A. M. (2016). Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. *American Economic Journal: Microeconomics*, **8** (1), 110–139.

Hossain, T. and Okui, R. (2013). The binarized scoring rule. *The Review of Economic Studies*, **80** (3), 984–1001.

Huck, S. and Weizsäcker, G. (2002). Do players correctly estimate what others do? Evidence of conservatism in beliefs. *Journal of Economic Behavior & Organization*, **47** (1), 71–85.

Hurd, M., Van Rooij, M. and Winter, J. (2011). Stock market expectations of dutch households. *Journal of Applied Econometrics*, **26** (3), 416–436.

Jacowitz, K. E. and Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, **21** (11), 1161–1166.

Jose, V. R. R. and Winkler, R. L. (2009). Evaluating quantile assessments. *Operations research*, **57** (5), 1287–1297.

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

Kaufmann, K. and Pistaferri, L. (2009). Disentangling insurance and information in intertemporal consumption choices. *American Economic Review*, **99** (2), 387–92.

Kirchkamp, O. and Reiss, J. P. (2011). Out-of-equilibrium bids in first-price auctions: Wrong expectations or wrong bids. *The Economic Journal*, **121** (557), 1361–1397.

Kleinjans, K. J. and Soest, A. V. (2014). Rounding, focal point answers and nonresponse to subjective probability questions. *Journal of Applied Econometrics*, **29** (4), 567–585.

Lahiri, K. and Teigland, C. (1987). On the normality of probability distributions of inflation and GNP forecasts. *International Journal of Forecasting*, **3** (2), 269–279.

—, — and Zaporowski, M. (1988). Interest rates and the subjective probability distribution of inflation forecasts. *Journal of Money, Credit and Banking*, **20** (2), 233–248.

Laio, F. and Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, **11** (4), 1267–1277.

Lichtenstein, S. and Fischhoff, B. (1977). Do those who know more also know more about how much they know. *Organizational Behavior and Human Performance*, **20** (2), 159–183.

Machina, M. J. and Schmeidler, D. (1992). A more robust definition of subjective probability. *Econometrica*, **60** (4), 745–780.

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

— (2018). Survey measurement of probabilistic macroeconomic expectations: Progress and promise. *NBER Macroeconomics Annual*, **32** (1), 411–471.

— and Molinari, F. (2010). Rounding probabilistic expectations in surveys. *Journal of Business & Economic Statistics*, **28** (2), 219–231.

— and Neri, C. (2013). First-and second-order subjective expectations in strategic decision-making: Experimental evidence. *Games and Economic Behavior*, **81**, 232–254.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, **22** (10), 1087–1096.

McKelvey, R. D. and Page, T. (1990). Public and private information: An experimental study of information pooling. *Econometrica*, **58** (6), 1321–1339.

McKenzie, D., Gibson, J. and Stillman, S. (2013). A land of milk and honey with streets paved with gold: Do emigrants have over-optimistic expectations about incomes abroad? *Journal of Development Economics*, **102**, 116–127.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, **115** (7), 1330–1338.

Neri, C. (2015). Eliciting beliefs in continuous-choice games: A double auction experiment. *Experimental Economics*, **18** (4), 569–608.

Nyarko, Y. and Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica*, **70** (3), 971–1005.

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

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

Osband, K. (1989). Optimal forecasting incentives. *Journal of Political Economy*, **97** (5), 1091–1112.

Palley, A. and Bansal, S. (2019). Is it better to elicit quantile or probability judgments to estimate a continuous distribution? *Kelley School of Business Research Paper*, (17-44).

Qu, X. (2012). A mechanism for eliciting a probability distribution. *Economics Letters*, **115** (3), 399–400.

Rutström, E. E. and Wilcox, N. T. (2009). Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior*, **67** (2), 616–632.

Sapienza, P., Toldra-Simats, A. and Zingales, L. (2013). Understanding trust. *The Economic Journal*, **123** (573), 1313–1332.

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

Schlag, K. H., Tremewan, J. and Van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, **18** (3), 457–490.

— and van der Weele, J. J. (2013). Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality. *Theoretical Economics Letters*, **3** (01), 38.

Schlaifer, R. and Raiffa, H. (1961). *Applied statistical decision theory*. Cambridge, Mass.: Harvard Business School.

Schmidt, P. (2019). Elicitation of ambiguous beliefs with mixing bets. *arXiv e-prints*, arXiv:1902.07447.

Schotter, A. and Trevino, I. (2014). Belief elicitation in the laboratory. *Annual Review of Economics*, **6** (1), 103–128.

Selten, R., Sadrieh, A. and Abbink, K. (1999). Money does not induce risk neutral behavior, but binary lotteries do even worse. *Theory and Decision*, **46** (3), 213–252.

Smith, C. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society. Series B (Methodological)*, **23** (1), 1–37.

Trautmann, S. T. and van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, **125** (589), 2116–2135.

Tversky, A. and Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*, Springer, pp. 141–162.

Vargas Hill, R. (2009). Using stated preferences and beliefs to identify the impact of risk on poor households. *The Journal of Development Studies*, **45** (2), 151–171.

Wang, Y. (2014). Dynamic implications of subjective expectations: Evidence from adult smokers. *American Economic Journal: Applied Economics*, **6** (1), 1–37.

Wilcox, N. T. (2006). Theories of learning in games and heterogeneity bias. *Econometrica*, **74** (5), 1271–1292.

Winkler, R. L. (1967). The assessment of prior distributions in bayesian analysis. *Journal of the American Statistical Association*, **62** (319), 776–800.

Wright, W. F. and Anderson, U. (1989). Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. *Organizational Behavior and Human Decision Processes*, **44** (1), 68–82.

Zarnowitz, V. and Lambros, L. A. (1987). Consensus and uncertainty in economic prediction. *Journal of Political Economy*, **95** (3), 591–621.

Appendix

A Proofs

We provide formal statements of the main results. Let the state space be denoted by $\Omega = \mathbb{R} \times [0,1]$, where each state $\omega = (y, r)$ consists of the real valued y and some randomization device outcome r. An event E is a subset of Ω. Let $\mathcal{P}$ bet the set of eligible probability measures for y on $\mathbb{R}$. Let $\mathcal{X}$ be the action space of the agent, and $x \in \mathcal{X}$ the report issued by the agent. A function $s : \mathcal{X} \times \mathbb{R} \mapsto \mathbb{R}$ is called a scoring rule. For any scoring rule within the binary lottery procedure in Section 2, every report $x \in \mathcal{X}$ can be associated with a binary act $M_{E(x)}m$ that pays prize M if the event $E(x) = \{(y, r) \in \Omega \mid s(x, y) > r\}$ realizes and m otherwise.

We assume probabilistically sophisticated preferences as introduced in Machina and Schmeidler (1992) with some measure $\mathbb{P}_0$ for the real-valued outcome y and where the randomisation device r is perceived as independent.

Regularity Conditions 1 (Probabilistic Sophistication (Machina and Schmeidler, 1992))**.** *There exists a probability measure $\mathbb{P}_\Omega$ on Ω such that for all events E and E' and all payoffs $M \succ m$*

$$M_E m \succeq M_{E'} m \iff \mathbb{P}_\Omega(E) \geq \mathbb{P}_\Omega(E'),$$

and $\mathbb{P}_\Omega = \mathbb{P}_0 \times U[0,1]$ for some $\mathbb{P}_0 \in \mathcal{P}$.

Regularity Conditions 1 hold for a subjective expectation maximizing agent, if the unknown utility function depends only on the obtained prize and is otherwise independent of the uncertain outcome y.

Proposition 1 (density)**.** *For absolutely continuous measures $\mathcal{P}$ with densities that are $\mathbb{P}_0$-almost surely bounded by $B \in \mathbb{R}$ and for the scoring rule*

$$s(p, y) = 2p(y) - \int_\Omega p(w)^2 dw + B,$$

any element of $\arg\max_{p \in \mathcal{P}} M_{E(p)}m$ with $E(p) = \{(y, r) \in \Omega \mid s(p, y) > 3Br\}$ is a density of $\mathbb{P}_0$.

Note that the result naturally extends to discrete measures. In fact, IP in Section 2.1 is a special case of Proposition 1, as for discrete probability distributions $p(y) = p_k$ and thus $s(p, y) = 2p_k - \sum p_k^2 + 1$. For discrete distributions a random draw on $[0, 2]$ instead of $[0, 3]$ suffices as $p_k \geq p_k^2$ for $p_k \in [0, 1]$.

Proof. As $\int_\Omega p(w)^2 dw \leq \int_\Omega Bp(w)dw = B$ and $p(y) \geq 0$, it holds that $s(p,y) \in [0, 3B]$. Thus, the agent maximizes

$$\mathbb{P}_\Omega(s(p,y) > 3Br) = \mathbb{E}_{y\sim\mathbb{P}_0}[\mathbb{E}_{r\sim U[0,1]}[\mathbb{1}(s(p,y) > 3Br)]] = \mathbb{E}_{y\sim\mathbb{P}_0}[\tfrac{1}{3B}s(p,y)].$$

Consider for any density p the term

$$\Delta p = \mathbb{E}[s(p_0,y)] - \mathbb{E}[s(p,y)],$$

where p_0 is a density of $\mathbb{P}_0$. It holds that

$$\begin{aligned}\Delta p &= \int 2p_0(y)^2 dy - \int p_0(w)^2 dw - (\int 2p(y)p_0(y)dy - \int p(w)^2 dw) \\ &= \int p_0(y)^2 - 2p(y)p_0(y) + p(y)^2 dy \\ &= \int (p_0(y) - p(y))^2 dy \geq 0\end{aligned}$$

If $p \in \arg\max_{p\in\mathcal{P}} \mathbb{E}_{y\sim\mathbb{P}_0}[s(p,y)]$, then $\Delta p = 0$ and thus $p = p_0$ Lebesgue-almost surely. Consequently, p is also a density of $\mathbb{P}_0$. □

Theorem 1 (multiple quantiles)**.** *Consider $a_i, b_i > 0$ for $i = 1, \dots, n$. Let $\mathcal{P}$ be a class of absolutely continuous probability distributions with finite moment and strictly positive density on their support. Consider the scoring rule*

$$s_e(x,y) = \frac{1}{ne}\sum_{i=1}^{n} \max(s_{a_i,b_i}(x_i,y), 0)$$

with

$$s_{a_i,b_i}(x_i,y) = \begin{cases} e - a_i \cdot |x_i - y| & \text{if } x \leq y \text{ (underestimation)}, \\ e - b_i \cdot |x_i - y| & \text{if } x > y \text{ (overestimation)}. \end{cases} \tag{4}$$

Let $x^*(e) = (x_1^*(e), \dots, x_n^*(e)) = \arg\max_{x\in\mathbb{R}^n} M_{E(x)}m$ with $E(x) = \{(y,r) \in \Omega \mid s_e(x,y) > r\}$ and $\alpha_i = a_i/(a_i + b_i)$ for $i = 1, \dots, n$.

(i) For $i = 1, \dots, n$ it holds that

$$x_i^*(e) \to q_{\alpha_i}(\mathbb{P}_0) \text{ for } e \to \infty,$$

where $q_{\alpha_i}(\mathbb{P}_0)$ denotes the quantile at level α_i of the distribution $\mathbb{P}_0$.

(ii) Consider a fixed endowment $e \in \mathbb{R}$ and the according optimal point prediction x_i^* for $i = 1, \dots, n$. If the tails of $\mathbb{P}_0$ are bounded with $\mathbb{P}_0(y \leq x_i^* - e/b_i) < c_1$ and $\mathbb{P}_0(y > x_i^* + e/a_i) < c_2$,

then

$$x_i^* = q_{\alpha_i^*}(\mathbb{P}_0)$$

for some level α_i^* *such that*

$$-\alpha_i c_2 < \alpha_i^* - \alpha_i < (1-\alpha_i)c_1.$$

Note that the result can be extended to densities that are not strictly positive and to discrete distributions under more complicated notation. In this case, some quantiles are set-valued and the best response converges to an element of the set.

Proof. For notational convenience we define the score associated with the ith point prediction as $s_i^+ = \max(s_{a_i,b_i}(x_i, y), 0)$ and thus $s_e(x, y) = \frac{1}{ne}\sum s_i^+$. As $0 \leq s_i^+ \leq e$ it follows that $0 \leq s_e(x, y) \leq 1$, and the agent maximizes

$$\mathbb{P}_\Omega(s_e(x, y) > r) = \mathbb{E}_{y\sim\mathbb{P}_0}[\mathbb{E}_{r\sim U[0,1]}[\mathbb{1}(s_e(x, y) > r)]] = \mathbb{E}_{y\sim\mathbb{P}_0}[s_e(x, y)].$$

We can investigate each point forecast x_i separately as

$$\begin{aligned}
\arg\max_{x\in\mathbb{R}^n} \mathbb{E}[s_e(x, y)] =& \arg\max_{x\in\mathbb{R}^n} \mathbb{E}[\frac{1}{ne}\sum s_i^+] \\
=& \arg\max_{x\in\mathbb{R}^n} \sum \mathbb{E}[s_i^+] \\
=& (\arg\max_{x_1\in\mathbb{R}} \mathbb{E}[s_1^+], \ldots, \arg\max_{x_n\in\mathbb{R}} \mathbb{E}[s_n^+])
\end{aligned}$$

where the last equation holds true as the expected score $\mathbb{E}[s_i^+]$ does not depend on x_j with $j \neq i$.

The score s_i^+ is non-zero for $x_i - e/b_i < y < x_i + e/a_i$. Thus,

$$\mathbb{E}[s_i^+] = \int_{x_i - e/b_i}^{x_i} e + b_i(y - x_i)f(y)dy + \int_{x_i}^{x_i + e/a_i} e + a_i(x_i - y)f(y)dy,$$

where f denotes a density function of $\mathbb{P}_0$. We denote the cdf of $\mathbb{P}_0$ with F and compute the derivative

$$\begin{aligned}
\frac{\partial}{\partial x_i}\mathbb{E}[s_i^+] &= \int_{x_i}^{x_i + e/a_i} a_i f(y)dy - \int_{x_i - e/b_i}^{x_i} b_i f(y)dy \\
&= a_i[F(x_i + e/a_i) - F(x_i)] - b_i[F(x_i) - F(x_i - e/b_i)] \\
&= b_i F(x_i - e/b_i) + a_i F(x_i + e/a_i) - (a_i + b_i)F(x_i)
\end{aligned}$$

by Leibniz rule. The first order condition is

$$F(x_i) = \frac{b_i F(x_i - e/b_i) + a_i F(x_i + e/a_i)}{a_i + b_i}.$$

The second derivative is

$$\frac{\partial^2}{(\partial x_i)^2}\mathbb{E}[s_i^+] = b_i f(x_i - e/b_i) + a_i f(x_i + e/a_i) - (a_i + b_i) f(x_i).$$

Case (i):

As F is a cdf, $F(x_i - e/b_i) \to 0$ and $F(x_i + e/a_i) \to 1$ for $e \to \infty$. Thus, the first order condition implies $F(x_i) \to \frac{a_i}{a_i + b_i} = \alpha_i$. Given our assumptions, F is strictly monotone and continuous on the support and we can conclude that $x_i \to q_{\alpha_i}(\mathbb{P})$.

Consider the second order condition to show that the first order condition is sufficient. It holds that $\lim_{e\to\infty} f(x_i - e/b_i) = \lim_{e\to\infty} f(x_i + e/a_i) = 0$ and

$$\frac{\partial^2}{(\partial x_i)^2}\mathbb{E}[s_i^+] = -(a_i + b_i) f(x_i) < 0$$

as f is strictly positive on the support and an off-support x_i cannot be optimal.

Case (ii):

Define the constants c_1 and c_2 such that $F(x_i^* - e/b_i) < c_1$ and $1 - F(x_i^* + e/a_i) < c_2$. It follows that

$$\begin{aligned} F(x_i^*) &< \frac{b_i c_1 + a_i}{a_i + b_i} = \alpha_i + c_1 \frac{b_i}{a_i + b_i} = \alpha_i + c_1(1 - \alpha_i), \\ F(x_i^*) &> \frac{a_i(1 - c_2)}{a_i + b_i} = \alpha_i(1 - c_2). \end{aligned}$$

Thus, the error (in terms of the quantile level) can be bounded with

$$-\alpha_i c_2 < F(x_i^*) - \alpha_i < c_1(1 - \alpha_i).$$

□

Proposition 2. *If $\mathcal{P}$ contains only distributions with bounded support of length B and $e >$*

$B \max(a_1, b_1, \ldots, a_n, b_n)$, *it follows that*

$$x^* = (q_{\alpha_1}(\mathbb{P}_0), \ldots, q_{\alpha_n}(\mathbb{P}_0)).$$

Proof. If the support of f is bounded with length B then, $e > Bb_i$ guarantees that $F(x_i - e/b_i) = 0$ and $e > Ba_i$ that $F(x_i + e/a_i) = 1$. In this case, the first order condition reduces to $F(x_i) = \alpha_i$. The second order condition holds as $f(x_i - e/b_i) = 0$ and $f(x_i + e/a_i) = 0$ and $f(x_i) > 0$. □

We define the minimum property

$$\min : \mathcal{P} \mapsto \Omega : \mathbb{P} \mapsto \inf\{x \in \Omega \mid \mathbb{P}(x) > 0\},$$

and analogously the maximum property

$$\max : \mathcal{P} \mapsto \Omega : \mathbb{P} \mapsto \sup\{x \in \Omega \mid \mathbb{P}(x) > 0\}.$$

Proposition 3. *If $b_i \to \infty$ and $\frac{e}{b_i} \to \infty$, the best response converges to the minimum of the support, i.e.,*

$$x_i^* \to \min(\mathbb{P}_0).$$

If $a_i \to \infty$ and $\frac{e}{a_i} \to \infty$, the best response converges to the maximum of the support.

Proof. If $b_i \to \infty$ and $\frac{e}{b_i} \to \infty$, we observe that

$$F(x_i) = \frac{F(x_i - e/b_i) + \frac{a_i}{b_i} F(x_i + e/a_i)}{\frac{a_i}{b_i} + 1} \to 0.$$

As f is strictly positive the quantile for every level is unique and for every $c \in \{t \in \mathbb{R} | p_0(t) > 0\}$ there exists a level $\alpha_i^* \in (0, 1)$ such that $q_{\alpha_i^*}(\mathbb{P}_0) = c$.

First, consider the case $\min(\mathbb{P}_0) = -\infty$. Take some $c \in \mathbb{R}$. As $F(y < c) > 0$, there exists b_i, e such that the first order condition implies $F(x_i) < F(c)$ and consequently $x_i < c$. Thus, $F(x_i) \to 0$ implies $x_i \to -\infty$.

Now consider the case of a finite $\min(\mathbb{P}_0)$. For any $c > \min(\mathbb{P}_0)$, there exists b_i, e such that $\alpha_i < F(c)$ and consequently $x_i < c$. As $x_i \geq \min(\mathbb{P}_0)$ for all b_i, e, it follows that $F(x_i) \to 0$ implies $x_i \to \min(\mathbb{P}_0)$.

Again, we check that the second order condition is fulfilled

$$\frac{\partial^2}{(\partial x_i)^2} \mathbb{E}[s_i^+] = b_i(f(x_i - e/b_i) - f(x_i)) - a_i f(x_i) < 0.$$

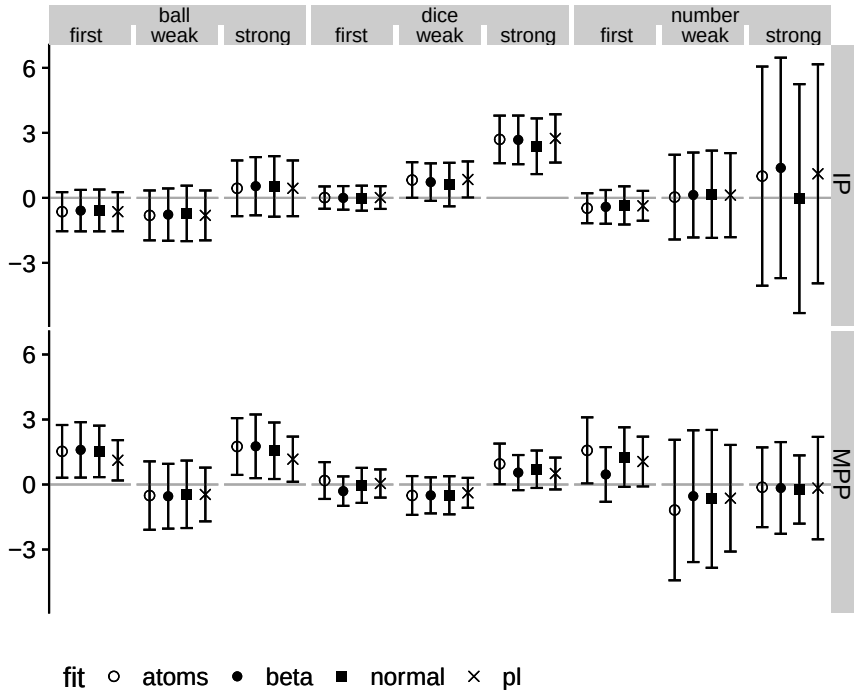
It follows that

$$x_i^* \to \min(F) \text{ for } b_i \to \infty \text{ and } \frac{b_i}{e} \to \infty.$$

A similar arguments gives $x_i^* \to \max(F)$ for $a_i \to \infty$ and $\frac{a_i}{e} \to \infty$.

□

Figure 11: **Mean: Comparison between fitted distributions and Bayesian distributions.**

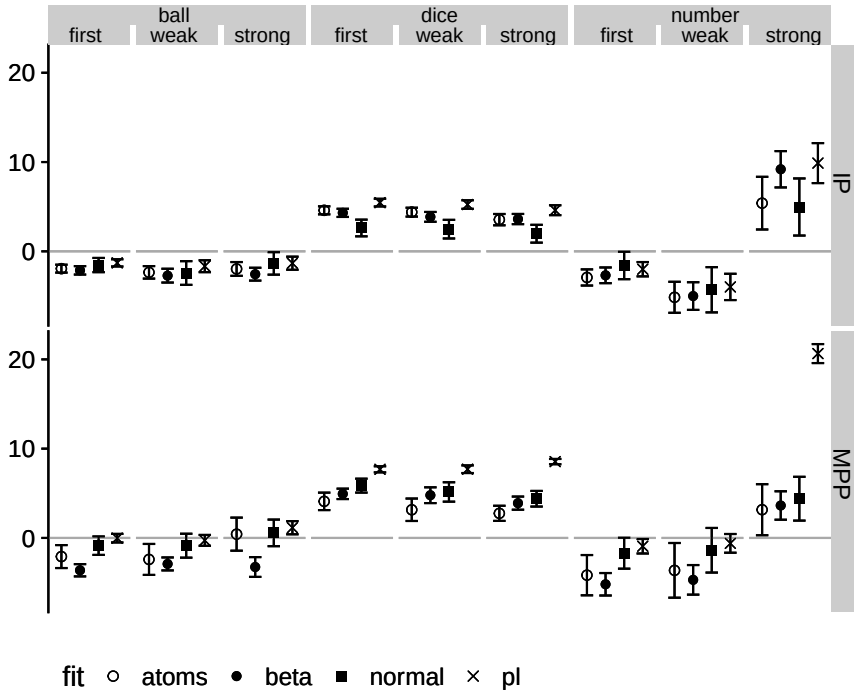


The target variable is $Z := mean(\mathbb{P}) - mean(\mathbb{P}_{bayes})$. The MPP and IP value indicate the bias of the extracted mean forecast.

B Results

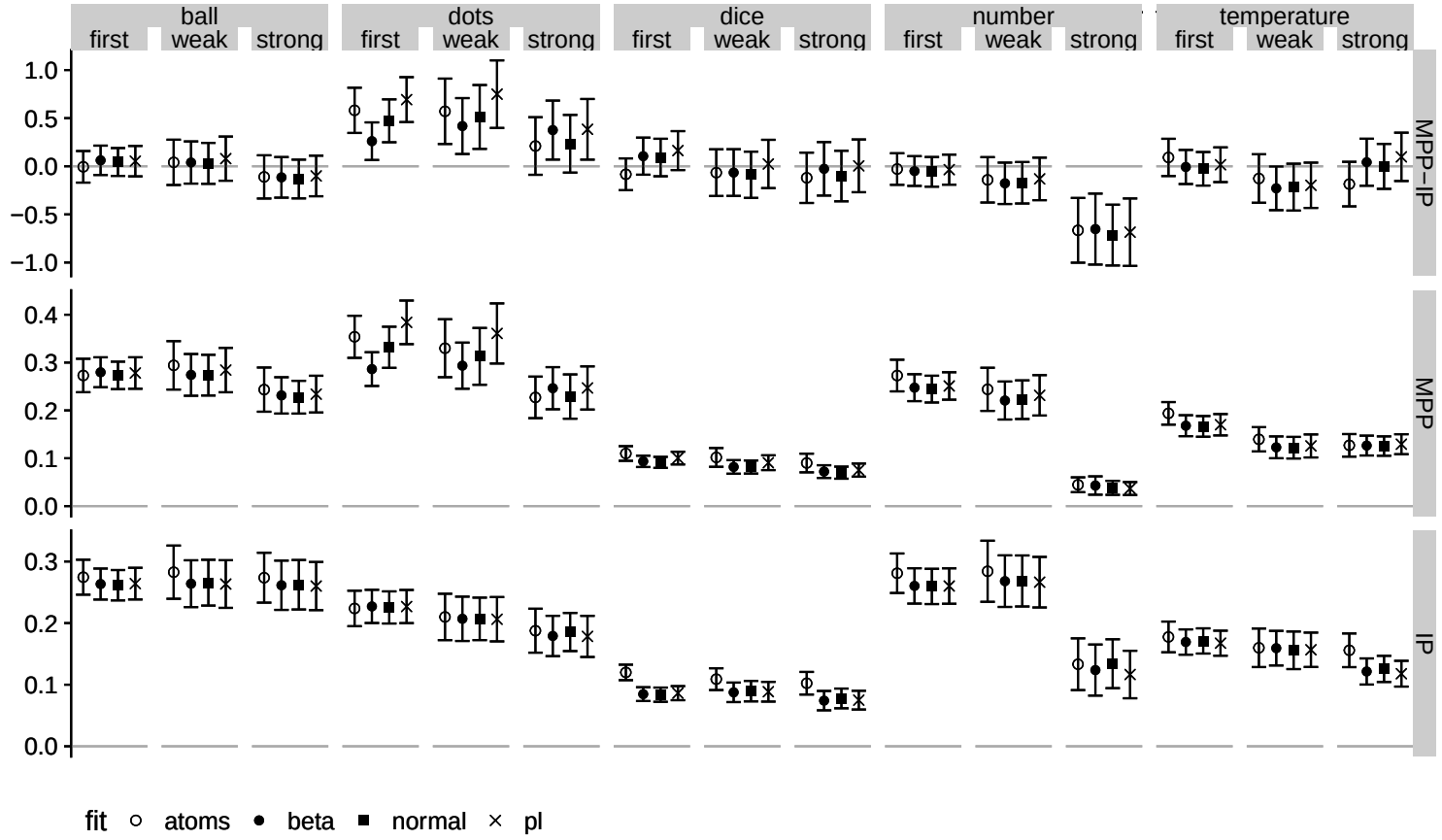
This section provides the full set of results, complementing the plots in Section 4 that show the best performing fit only. The interpretation of the results remains unaffected. Throughout, the error bars show 95% confidence interval in figures and parentheses show the respective p-values in tables.

Figure 12: **Standard deviation: Comparison between fitted distributions and Bayesian distributions.**



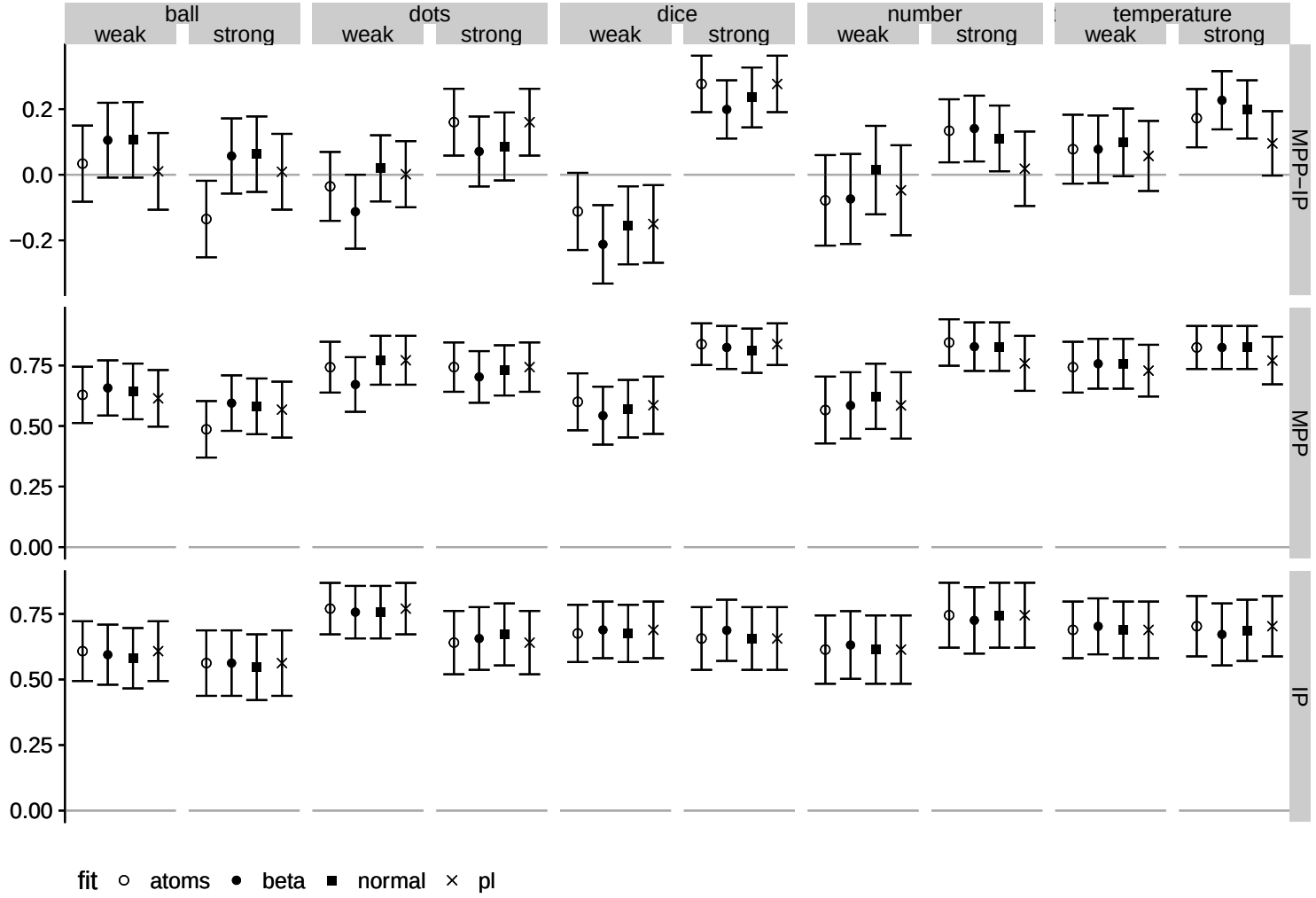
The target variable is $Z := sd(\mathbb{P}) - sd(\mathbb{P}_{bayes})$. The MPP and IP value indicate the bias of the extracted standard deviation forecast.

Figure 13: **Difference in accuracy - absolute error**



The target variable is denoted as $Z := |median(\mathbb{P}) - y|$. The MPP and IP value indicate average absolute error normalized by the highest possible absolute error within each domain. The $MPP - IP$ value indicates the average difference normalized by IP, where negative values indicate superior accuracy of MPP.

Figure 14: **Consistency of willingness to pay and subjective probabilities**



The target variable is 1 if the mean of the predictive distribution is above the offer and the offer was accepted or if the mean is below and the offer was rejected. Otherwise, the dependent variable is 0 indicating inconsistent behavior. The MPP and IP value indicate the average ratio of consistent behavior within each treatment and domain. The $MPP - IP$ coefficient indicates the average difference normalized by the IP value, where positive values indicate superior consistency of MPP.

		MPP			IP		
topic	info	0.25	0.5	0.75	1	2	3
ball	first	3	1.2	-2.2	-0.2	-2.1	-2.8
		(1.45)**	(1.38)	(1.36)	(0.91)	(1.3)	(1.21)**
	weak	0.4	-1.6	-5.9	1.6	-2.7	-5.8
		(1.56)	(1.76)	(2.08)**	(1.34)	(1.78)	(1.61)***
	strong	-1.6	0.0	1.9	1.6	1.2	-0.2
		(1.85)	(1.83)	(1.92)	(1.6)	(1.95)	(1.59)
dice	first	-10.6	-6.8	7.6	-12.8	-4.2	12.6
		(1.86)***	(2.14)***	(1.84)***	(0.92)***	(1.24)***	(0.95)***
	weak	-11.5	-6.5	3.5	-8.4	-1.4	12.3
		(2.31)***	(3.1)**	(2.99)	(1.08)***	(2.34)	(1.32)***
	strong	-8.5	-2.4	5.6	-3.2	10.4	11.8
		(2.65)***	(3.37)	(2.66)**	(0.86)***	(3.35)***	(2.02)***
number	first	4.1	0.5	-2.1	2.0	-1.0	-3.9
		(1.38)***	(0.94)	(1.15)*	(0.84)**	(0.78)	(0.73)***
	weak	-2.3	-3.7	-8	6.0	-0.6	-4.4
		(1.88)	(1.73)**	(2.34)***	(2.06)**	(1.6)	(1.48)***
	strong	-11.4	-5.7	2.9	3.6	1.5	0.8
		(1.71)***	(2.06)**	(2.64)	(4.2)	(3.62)	(2.52)

Table 6: **Comparison between reports and Bayesian values per task.** Columns *0.25*, *0.5* and *0.75* show the differences between issued point forecast and the respective true Bayesian quantile in percentage points. Columns *1*, *2* and *3* show the differences of the issued interval probability and respective true Bayesian probabilities in percent. Brackets show standard deviations of the differences. *p<0.1; **p<0.05; ***p<0.01

Additionally, we provide quantitative evidence for the deviations from the Bayesian values in Figure 4, per task (Table 6) and aggregated over all tasks (Table 7).

	MPP			IP		
info	0.25	0.5	0.75	1	2	3
first	-1.6	-1.9	1.4	-4.1	-2.5	2.4
	(0.99)	(0.97)*	(0.92)	(0.62)***	(0.68)***	(0.71)***
weak	-4.7	-4.0	-3.1	-0.8	-1.6	1.1
	(1.19)***	(1.38)***	(1.51)**	(0.94)	(1.15)	(1.04)
strong	-6.8	-2.5	3.5	0.5	4.6	4.4
	(1.28)***	(1.5)	(1.39)**	(1.37)	(1.75)**	(1.23)***

Table 7: **Comparison between reports and Bayesian values on aggregate.** Columns *0.25*, *0.5* and *0.75* show the differences between issued point forecast and the respective true Bayesian quantile in percentage points. Columns *1*, *2* and *3* show the differences of the issued interval probability and respective true Bayesian probabilities in percent. Brackets show standard deviations of the differences. *p<0.1; **p<0.05; ***p<0.01

C Additional experiment: Unincentivized reports

In this section we provide the results of an additional experiment which was conducted with a sample of 89 participants after the main experiment.

In this additional experiment, we applied the same two treatments as before (MPP and IP, weak and strong information updates). The participants were asked for the number of heads out of 100 coin flips. The weak and strong information update was the number of heads in the first 50 and 90 flips, respectively.

Since participants of that additional experiment had completed the main experiment before, they were already aware of the general procedure when entering the additional experiment. After having completed the main experiment we therefore informed them about the differences that applied to the remainder of the respective session. The crucial difference to the main experiment is that the participants were explicitly told that these reports would not influence their payments. (Compare the explanation in Section S2.5 of the supplementary document.)

As such we can investigate if hypothetical incentives can be used to elicit beliefs. While this is common practice in applied studies with probability questions, little is known about the reliability of unincentivized reports that use the incentive structure merely as an explanation and communication device.

Further, we elicited more CDF points in this experiment. For MPP we elicited seven quantiles for the levels

$$\alpha = (\tfrac{1}{101}, \tfrac{1}{11}, \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{3}{4}, \tfrac{10}{11}, \tfrac{100}{101}).$$

For IP, we first asked for the minimum and maximum (*What is the minimum/maximum number of times that the coin shows "heads" out of this* 100 *coin flips?*), divided this interval into seven

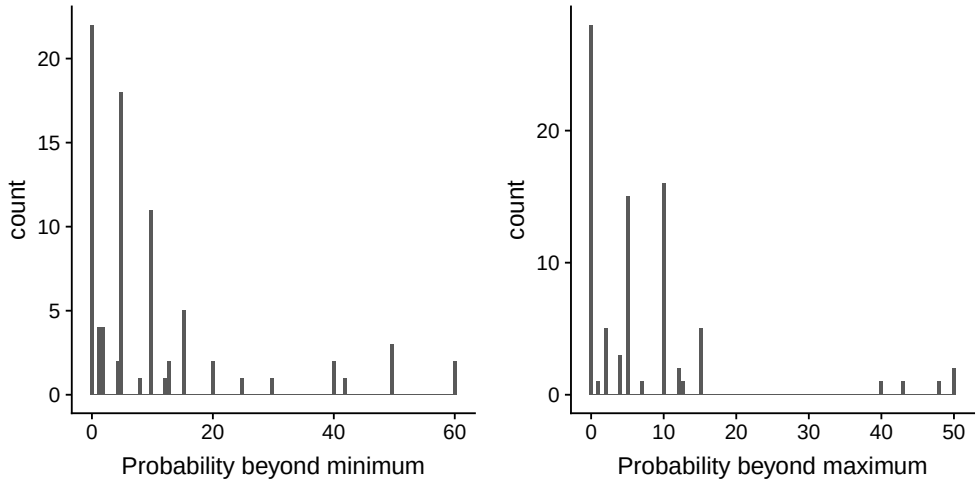


Figure 15: **Reported probabilities beyond minimum and maximum.** Histogram of reported probabilities (in %) for the interval beyond the minimum (left plot) and the maximum (right plot) in the IP treatment.

equidistant subintervals, and finally elicited the nine probabilities on the generated intervals.

Note, that while these changes render a direct comparison between behaviour in the main experiment and behaviour in this additional experiment impossible, this additional experiment further contributes to the comparison of MPP versus IP. We aim to provide first evidence that belief elicitation using MPP can be beneficial in settings beyond experimental work (e.g., unincentivized surveys).

C.1 Reports for minimum and maximum

We find that about 80% of participants report positive probabilities for the intervals beyond the minimum and maximum report. Similar patterns were observed in Delavande *et al.* (2011) and Dominitz and Manski (1997). Thus, the reported probabilities and the minimum and maximum estimates are not consistent.

In Figure 15, we see that most participants report positive probabilities that are relatively small. In over 80% of the cases the reported mass does not exceed 0.1. One interpretation of the results would be that participants neglect the tails and issue extreme quantiles instead of the actual minimum/maximum.

In the MPP treatment, we can use the extreme quantiles to approximate minimum/maximum reports (compare Proposition 3). Indeed, we see in Figure 16 that the lowest elicited quantile and the minimum report behave in a similar manner. Before the information update only a small

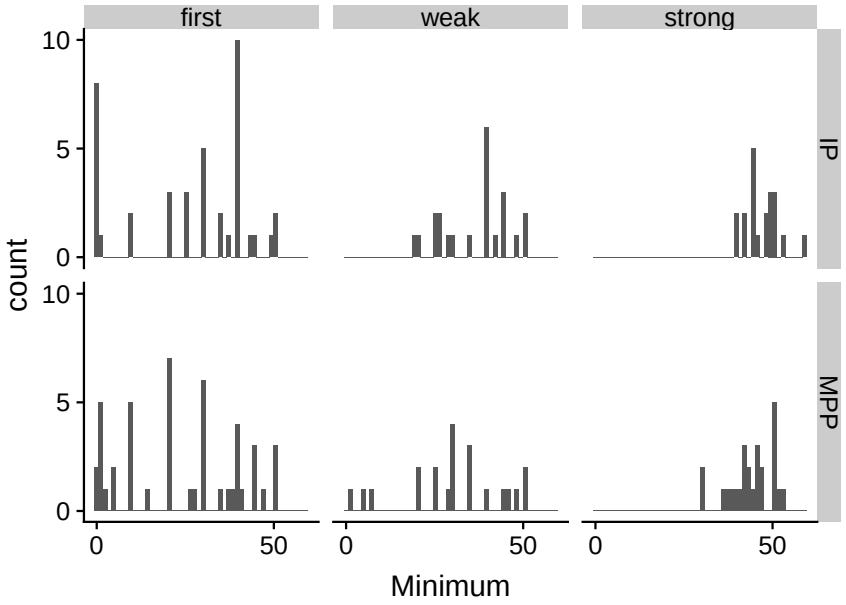


Figure 16: **Minimum reports.** Histogram of minimum reports for IP and extreme quantile reports for MPP in the first round, and after the weak and strong information update.

ratio of participants reports the true minimum at 0. Clearly, the information updates are largely incorporated logically, as both updates correctly increase the reported minimums and the strong information does so more heavily. We obtain similar results for the maximum, which are not shown here.

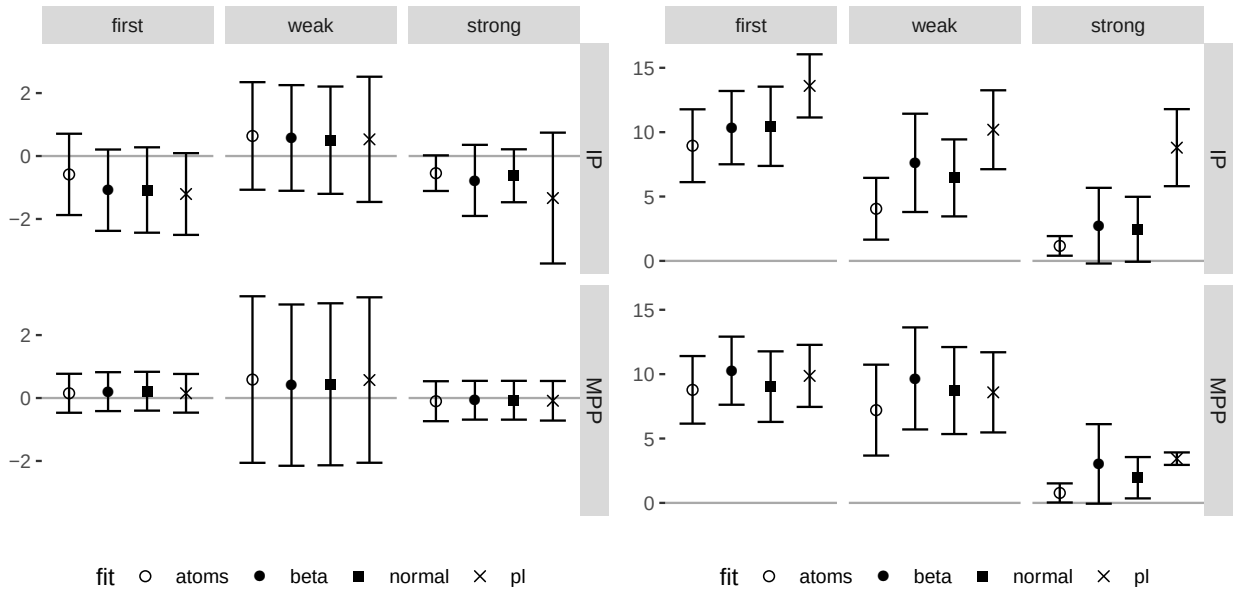
C.2 Biases and accuracy

We apply the same analysis as in the main part of the paper. Two observations for each elicitation method were deleted as the mean of their predictive distribution was below 30 or above 70 indicating that the participants either misunderstood or put no thought in the task. Figure 17a shows that there is no evidence for a biased mean of the predictive distribution.

Figure 17b suggests that participants overestimate uncertainty for the 100 coin flips, but are mostly able to judge the distribution for the 10 remaining coin flips after the strong information update. A possible explanation is that participants overestimate the uncertainty of a binomial distribution with many observations (Benjamin *et al.*, 2017).

The average absolute error of the median in Figure 18 shows no evidence against equal accuracy of both elicitation procedures.

In summary, the experiment provides no evidence that the absence of incentives, the elicitation



(a) **Mean.** (b) **Standard deviation.**

Figure 17: **Comparison between fitted distributions and Bayesian distributions.** For the mean plot, the target variable is $Z := mean(\mathbb{P}) - mean(\mathbb{P}_{bayes})$. For the standard deviation plot the target variable is $Z := sd(\mathbb{P}) - sd(\mathbb{P}_{bayes})$. Throughout, the error bars show 95% confidence intervals. Participants are pooled in the first round of elicitation (*first*), and distinguished after receiving the information update (*weak* and *strong*).

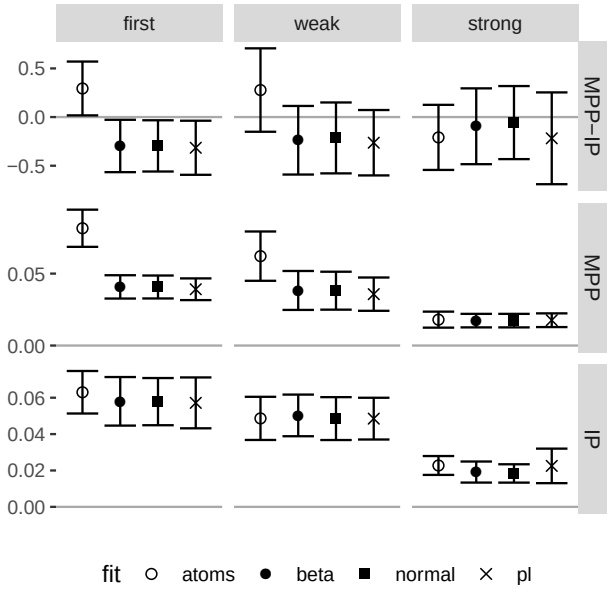


Figure 18: **Difference in accuracy.** The target variable is denoted as $Z := |median(\mathbb{P}) - y|$. The MPP and IP value indicate average absolute error normalized by the highest possible absolute error within each domain. The $MPP - IP$ value indicates the average difference normalized by IP, where negative values indicate superior accuracy of MPP.

of additional CDF points, and flexible support for the IP treatment fundamentally change the conclusions of the main experiment.