
Gutenberg School of Management and Economics

& Research Unit “Interdisciplinary Public Policy”

Discussion Paper Series

# *Belief Elicitation with Multiple Point Predictions*

Markus Eyting, Patrick Schmidt

October 22, 2019

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*
Goethe University Frankfurt and Heidelberg Institute for Theoretical Studies

This version: October 22, 2019

**Abstract**

We consider beliefs about real-valued outcomes, and show how to elicit the entire subjective probability distribution with binarized scoring rules. Further, we propose a simple, incentive compatible elicitation mechanism - *multiple point predictions* - that partially identifies the subjective probability distribution. Simultaneously eliciting multiple point predictions with linear incentives reveals the subjective probability distribution without pre-defined anchors or probabilistic statements. In a laboratory experiment, we test belief elicitation with multiple point predictions and compare it to the standard approach of eliciting discrete probabilities on pre-defined intervals. We find that elicitation with point predictions is faster, more convenient and more predictive of subsequent behavior. In the absence of anchors, the elicited distributions show evidence of first order biases, but are less prone to overconfidence. Further, the distributions are less accurate for uninformed participants, and more accurate if participants have heterogeneous information.

***Keywords***— elicitation of subjective expectations; partial identification; quantiles; overconfidence; 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 modeling 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), which take the form of continuous probability distributions.[1] As the human mind is commonly not used to processing probability distributions, the elicitation can be challenging in practice. In this paper, we propose to elicit subjective probability distributions indirectly with linear incentive schemes and framed as point predictions.

In the first part, we provide a unified framework for the incentivized elicitation of subjective probability distributions with binarized scoring rules[2] (Harrison *et al.*, 2014; Hossain and Okui, 2013; Schlag and van der Weele, 2013; Smith, 1961). A common procedure is to divide the real line into intervals and elicit the discrete distribution on those intervals. This elicitation can be incentivized with the widely applied quadratic scoring rule (QSR) (Costa-Gomes *et al.*, 2014; Huck and Weizsäcker, 2002; McKelvey and Page, 1990; Nyarko and Schotter, 2002; Offerman *et al.*, 2009; Palfrey and Wang, 2009; Rutström and Wilcox, 2009). The obtained interval probabilities (IP) identify the cumulative distribution function (CDF) at the pre-defined interval thresholds. We generalize this procedure and show that any bounded density can be elicited without pre-defined intervals.

It is often argued that it is preferable to incentivize belief elicitation (e.g., Blanco *et al.*, 2010; Gächter and Renner, 2010; Harrison, 2014; Schlag *et al.*, 2015). The literature provides several elicitation mechanisms that reveal (aspects of) subjective probability distributions about real-valued variables (Demuynck, 2013; Harrison *et al.*, 2015; Hossain and Okui, 2013; Qu, 2012; Schlag and van der Weele, 2013). While elicitation mechanisms can be theoretically equivalent, they have different psychological implications and experimental evidence for the applicability of different methods is context dependent (Schlag *et al.*, 2015; Schotter and Trevino, 2014; Trautmann and van de Kuilen, 2015).

For events, no simple linear scoring rule truthfully elicits the event probability. For subjective probability distributions on the real line, however, linear incentives identify pre-defined CDF levels. We propose to elicit multiple point predictions (MPP) with asymmetric linear incentives to identify quantiles of the subjective probability distribution. We show that MPP can be used to elicit points of the subjective CDF under probabilistically sophisticated preferences (Machina and Schmeidler, 1992). While IP allow to choose at which outcome levels the CDF is revealed, MPP allow to choose

[1] Manski (2018) and Delavande *et al.* (2011) review examples in macro and development economics.
[2] Also referred to as binary lottery procedure.

at which probability levels the CDF is revealed. Without anchors or explicit probabilistic statements[3], MPP provide essentially the same amount of 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 relatively intuitive elicitation mechanism.

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

Moreover, IP can be influenced by bin effects (Benjamin *et al.*, 2017) and insufficient adjustments from anchors (Jacowitz and Kahneman, 1995; Tversky and Kahneman, 1975; Wright and Anderson, 1989). Especially for non-expert respondents that are less familiar with the subject these effects may lead to distorted reports. The choice of intervals may influence the reported subjective probability distribution. Therefore these intervals have to be chosen carefully. If beliefs vary strongly across participants, uniformly adequate intervals may be hard to find. Individual specific adaptation of intervals, on the other hand, can influence responses and complicate comparisons across individuals.

Finally, incentivized reports are often based on complex payoff functions (e.g., proper scoring rules) (Brier, 1950; Winkler, 1967). Some procedures show payoffs contingent on outcomes, which allows respondents to explore the incentive structure (Harrison *et al.*, 2014; Holt and Smith, 2016). Other procedures explicitly tell participants that it is optimal to report their "true beliefs". This recommendation of optimality is debatable, as it depends on decision theoretic assumptions

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

(compare Offerman *et al.*, 2009).

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 multiple quantiles without assuming bounded support or limiting tail behavior (see Theorem 1). Under simpler preferences, related scoring procedures have been considered in Jose and Winkler (2009) and Fissler and Ziegel (2016). Further, extreme asymmetric linear incentives can reveal the minimum and maximum of the distribution (see Proposition 3). To the best of our knowledge, that is the first mechanism that reveals a non-elicitable property (Bellini and Bignozzi, 2015) with binarized scoring rules. Finally, we discuss why MPP often provides sharper bounds on the CDF than IP (see Section 2.4).

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. 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. Despite the fact that the two mechanisms predict identical responses under probabilistically sophisticated preferences, we find strong evidence for differing response behaviors. Using various evaluation criteria, we conduct a thorough analysis of the advantages and disadvantages of both elicitation methods.

In forecasting, the elicited subjective distributions should be informative about the unknown outcome. We define calibration and accuracy in a joint probability space that contains elicited distributions and the uncertain outcome. Specifically, we test first and second order calibration using the Probability Integral Transform (PIT) (Dawid, 1984; Diebold *et al.*, 1998), and compare the information content and accuracy of the elicited distributions with the Continuous Ranked Probability Score (CRPS) (Gneiting and Raftery, 2007; Matheson and Winkler, 1976). Both measures are easily interpretable and consider the whole distribution.

The elicited distributions reveal systematic biases that depend on the application as well as the elicitation procedure. While the elicitation of IP leads to overconfident (too narrow) probability distributions, elicitation of MPP shows no evidence for such miscalibration. Distributions obtained from MPP are biased for some applications. This bias is less severe for IP. We further find that the distributions elicited by MPP are slightly more consistent with subsequently recorded behavior. Finally, we find that uninformed participants report more accurate beliefs with IP, and that informed participants with heterogeneous beliefs report more accurate beliefs with MPP. These results are consistent with the interval thresholds functioning as anchors for uninformed participants. 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 a concurrent paper, Palley and Bansal (2019) provide a related experimental comparison by explicitly eliciting quantiles. We rely on a more intuitive framing using MPP. Further, we analyse calibration and accuracy in a more general framework based on the realized value instead of empirical marginal distributions. We argue that, whenever available, objective Bayes probabilities should be used. However, except for experimentally constructed domains, there rarely exist valid Bayes probabilities. In this case, the comparison with the realized value is a valid benchmark at the cost of additional noise and technicality. Conveniently, the same analysis can be mirrored in economic applications when individual specific realizations are available.

In the following section, we provide theoretical background on property elicitation 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 defined as a CDF –or equivalently a 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) with the following procedure. 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. See Regularity Conditions 1 in the Appendix for details.[4] The remainder of this section tackles the question which properties of the distribution can be elicited by the mechanism 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 an agent with subjective probability $\mathbb{P}$ is $T(\mathbb{P})$. Further, a series of elicitation methods is called *essentially*

[4]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.

*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 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 mechanism 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. Thus, we provide the first case of revealing properties with binarized scores that are generally not elicitable (Bellini and Bignozzi, 2015).

## 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 $c_1, \dots, c_{n-1}$ that define the respective property $T_c(\mathbb{P}) = (\mathbb{P}(y \leq c_1), \mathbb{P}(c_1 < y \leq c_2), \dots, \mathbb{P}(y > c_{n-1}))$ and to apply the QSR for discrete probabilities. The eligible reports are probability vectors for $n$ possible outcome values, $\mathcal{X} := \{x \in [0,1]^n | \sum_i x_i = 1\}$. After issuing the report $x = (x_i, \dots, x_n)$ 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 $[0, 2]$. This elicitation mechanism is incentive compatible for the discrete probability distribution $T_c$ (Winkler, 1967). 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)**.** *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 individual underestimates or overestimates, and then deducted from an initial endowment $e$.

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

Under this scoring rule the expected score maximizing strategy is to report the quantile of $\mathbb{P}$ with level $\alpha = \frac{a}{a+b}$ (Schlaifer and Raiffa, 1961). If the payoff consists of lottery tickets, the incentive

structure is robust beyond risk neutrality (Hossain and Okui, 2013; Schlag and van der Weele, 2013).

Similar incentives (without binarized scores) framed as prediction tasks have been used in the experimental literature before. Dufwenberg and Gneezy (2000) and Kirchkamp and Reiß (2011) use the symmetric version with $a = b$ of the linear scoring rule in Equation (2) which is incentive compatible for the median.[5] Charness and Dufwenberg (2006) and Sapienza *et al.* (2013) use the score $s(x, y) = \mathbb{1}(|x-y| < d)$ which is incentive compatible for the midpoint of the modal interval of length $d$. This scoring rule obtains only two possible values and is therefore robust to risk attitude without binarizing. A single point forecast, however, does not allow to do meaningful inference on the subjective probability distribution.

The following theorem shows how to elicit multiple quantiles simultaneously. 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 participant 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 that the agent has a subjective probability with strictly positive density.*

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

*(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 $\alpha_i^*$ such that*

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

See Appendix A for details. To avoid eliciting distorted quantiles, the initial endowment $e$ has to be chosen large enough to result in positive scores for every value of $y$ with positive subjective probability. In applications, a smaller value of $e$ might be desirable to increase the incentives to exert effort. Point *(ii)* of Theorem 1 allows to bound the error in terms of the probability level.

[5] Haruvy *et al.* (2007) apply a piecewise constant version of the linear symmetric incentives, which incentivizes a report that deviates slightly from the median.

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 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 $\alpha_i$ 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).

In Figure 1, we see an example of CDF points that can be identified by IP with interval thresholds $c = (-12, -6, 0, 6, 12)$. The example 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.[6] The grey area bounds the set of consistent CDFs.[7]

[6]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).

[7]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 $C_i$.

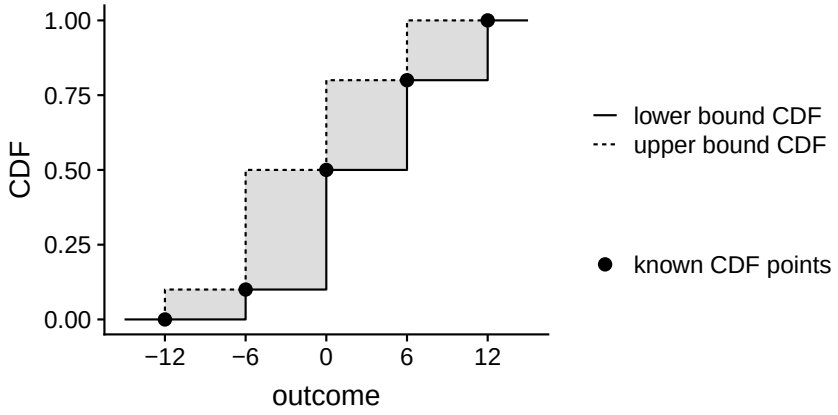



Figure 1: **Elicited CDF points and the space of consistent CDFs.** In this example five CDF points were elicited with IP on the interval thresholds $c = (-12, -6, 0, 6, 12)$ , and the elicited probabilities are $(0, 0.1, 0.4, 0.3, 0.2, 0)$. The space of consistent CDFs is denoted by any weakly increasing function that crosses the five given black dots. The possible CDF values are depicted as gray rectangulars.

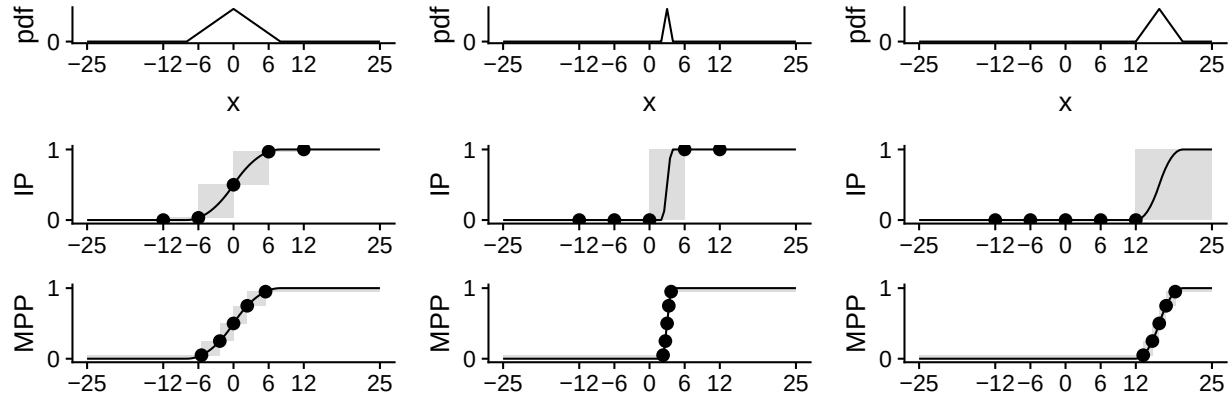



Figure 2: **Elicited CDF points and the space of consistent CDFs for three examples.** Each column contains the true belief as pdf in the first row, and the elicited CDF points and the space of consistent CDFs in the second and third row for IP with $c = (-12, -6, 0, 6, 12)$ and MPP with $\alpha = (0.05,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

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. As an example, we consider three individuals with different beliefs in Figure 2. 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 no 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 no information about the expectations beyond 12%. MPP do not allow to bound the CDF from below. 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 (illustrated in Figure 1). In the following we analyze which method provides sharper bounds. We assume that the support of the distribution is bounded.[8] For the mean property, it follows 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.[9]

Generally, it is harder to find valid bounds on measures of dispersion.[10] Conveniently, elicitation of MPP identifies the interquantile range between the elicited quantiles. In the elicitation of IP, quantiles are only pinned down in their respective intervals, thus the size of the bounds on any interquantile range is either the length of the elicited intervals (if both quantiles are located in the same interval) or twice that amount (if the quantiles are located in different intervals).

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

[9]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.

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

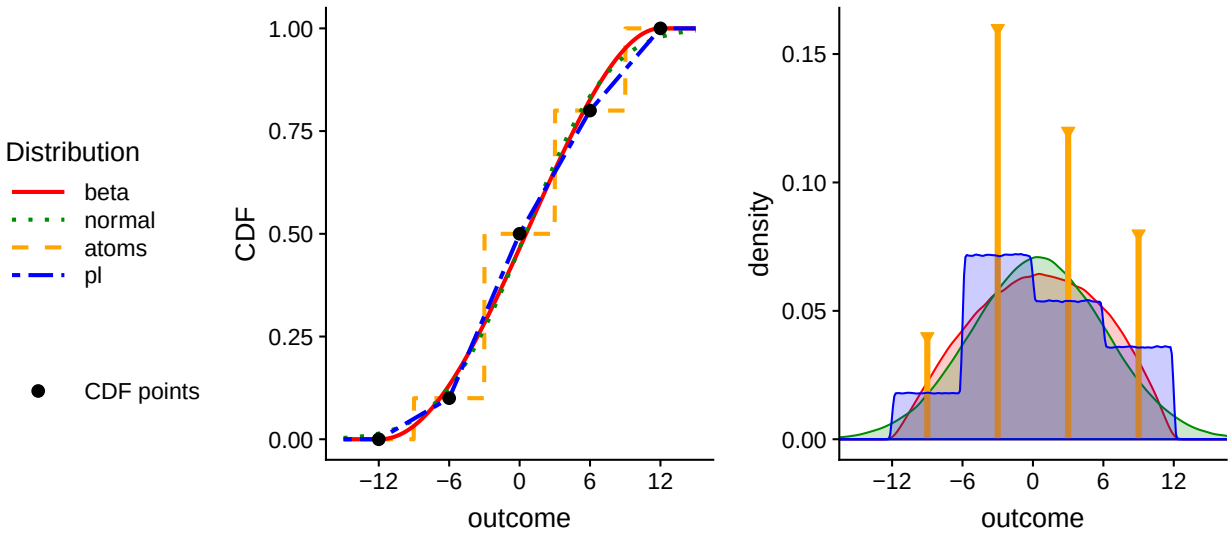



Figure 3: **Elicited CDF points and parametric distributions.** The fitted distributions are illustrated as CDFs (left plot) and densities (right plot).

## 2.5 Parametric Assumptions

When the bounds on the subjective probability distribution are too broad, one can rely on additional assumptions to fit unique parametric distributions to the elicited CDF points. We call the resulting distribution a *predictive distribution*. We consider four commonly applied procedures that are illustrated in Figure 3. The *atoms* distribution assumes a discrete distribution, where the mass is located at the midpoint of each elicited interval[11] (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.

Further, we fit predictive distributions by minimizing

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

where $x_i$ and $\alpha_i$ are obtained by the reports and $F(\cdot, \theta)$ is the CDF of the distribution for the parameter $\theta$. 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

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

Neri, 2013; Neri, 2015).[12]

Figure 3 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 underestimates the uncertainty.

# 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 August 2018. In total we recruited 282 subjects through the online system ORSEE (Greiner, 2004) divided into 14 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 Experimental Treatments

The experiment comprised four $(2 \times 2)$ different treatments as illustrated in Figure 4a. 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. Furthermore, the treatments differed with respect to the information updates that were presented to the subjects during the experiment, which allows to analyze the performance under different information environments.

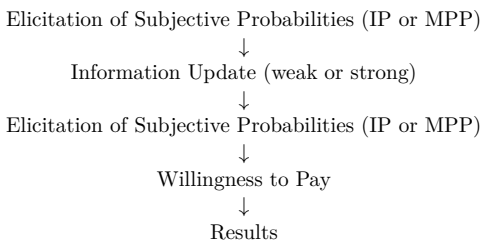
## 3.2 Domains

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

[12]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).

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

(a) Experimental treatments




(b) Elicitation procedure within each domain

Figure 4: Summary of experimental treatments and elicitation procedure within each of the five domains

4b 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 subject's belief 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. We use this willingness to pay as consistency check for the elicited beliefs.

For a summary of the domains and the used information updates see Table 1. 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.

[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 1: Information Updates

### 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. For details on the experiment and instructions, see the online supplement.

## 4 Results

We analyze several outcome measures $Z$. All results are based on simple averages or differences in means within each domain (and information update) separately. To provide some robustness with

respect to the chosen fits, we consider all fitting methods described in Section 2.5. Here, we only show the best performing fit for each treatment. See Appendix B for the full set of 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 modeling*. Forecasting strives to accurately describe the unknown outcome, whereas in economic modeling the subjective probabilities should accurately represent the belief of an agent.

A common approach for the evaluation of elicitation mechanisms is the comparison of the resulting subjective probabilities with objective probabilities (Schlag *et al.*, 2015, Table 1). However, many applications of interest do not allow for valid objective probabilities, because the conditional distribution of the uncertain outcome - given the information set of the participant - cannot be stated. Consequently, the predictive distributions cannot be compared to valid objective probability distributions.

Figure 5 depicts a comparison between the reported subjective CDF values and the objective Bayes CDF values. This kind of analysis is not possible for the temperature or dots domain, where no natural Bayes probabilities are available. Arguably, most uncertain entities in economics have such an ambiguous character. The linear fits provide first suggestive evidence for overconfidence of the IP reports in the ball and the number domain and for underconfident MPP reports in the dice domain. In the following we provide a more general analysis of first and second order biases based on the PIT that does not require the existence of valid Bayes probabilities.

Instead, we understand the outcome $y$ and the predictive distribution $\mathbb{P}$ as realizations in a joint probability space[14]. This allows us to analyze calibration and accuracy without unduly strong assumptions on the data generating process.

## 4.1 Calibration

We begin with an analysis of the *calibration* of the obtained predictive distributions. Calibration refers to the statistical consistency between the predictive distribution and the observation (Gneiting *et al.*, 2007). A sample of predictive distributions $\mathbb{P}$ is called *calibrated for* $y$, if each $\mathbb{P}$ constitutes the conditional distribution of $y$ given some (unknown) information set.[15]

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

[15]No assumptions are made on the variables that construct the information set. Lichtendahl *et al.* (2013) and Hora (2004) use closely related definitions of calibration that do not consider information sets. A similar analysis with additional structure on the information environment can be found in Grushka-Cockayne *et al.* (2016).

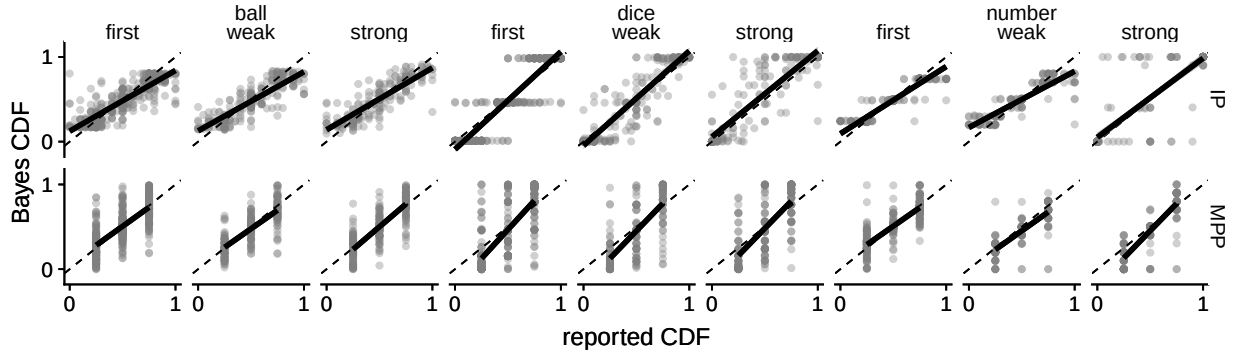



Figure 5: **Comparison with Bayesian distributions.** For the IP treatment, the reported probabilities are plotted against the Bayes probabilities at the elicited threshold levels. For the MPP treatment, the elicited quantile levels are plotted against the reported quantile levels of the Bayes distribution. Throughout, the Bayes distribution is computed based on the individual and time specific information. Linear fits with slopes smaller (larger) than the slope of the dashed line suggest overconfidence (underconfidence).

While the calibration of point forecasts could be assessed with a simple analysis of the forecast error, predictive distributions require a more sophisticated approach. We propose to analyze calibration with the first and second moment of the Probability Integral Transform (PIT), where the first moment reflects bias in position of the distribution (optimism and pessimism) and the second moment in spread of the distribution (overconfidence and underconfidence). The PIT is defined as the probability that the predictive distribution $\mathbb{P}$ assigns to the interval below the realized outcome $y$:

$$PIT(\mathbb{P}, y) = \mathbb{P}((-\infty, y]).$$

If the predictive distributions are calibrated, the PIT is uniformly distributed between zero and one (Diebold *et al.*, 1998; Gneiting and Ranjan, 2013). If the outcome is consistently larger than postulated by the predictive distribution, the average PIT is close to one. If the outcome is consistently smaller, the average PIT is closer to zero. We call a sample of predictive distributions *first order calibrated* if the average PIT is equal to $\frac{1}{2}$. First order calibration is a necessary condition for calibration.

Figure 6 depicts the average PIT. All results are transformed such that a coefficient of zero indicates first order calibration. A value of 0.5 indicates that the whole probability mass is underestimating and a value of $-0.5$ implies that the whole probability mass is overestimating the realized value. We can reject first order calibration in the dots and temperature domain for the MPP treatment. The strong information update erases the bias in the temperature domain. The

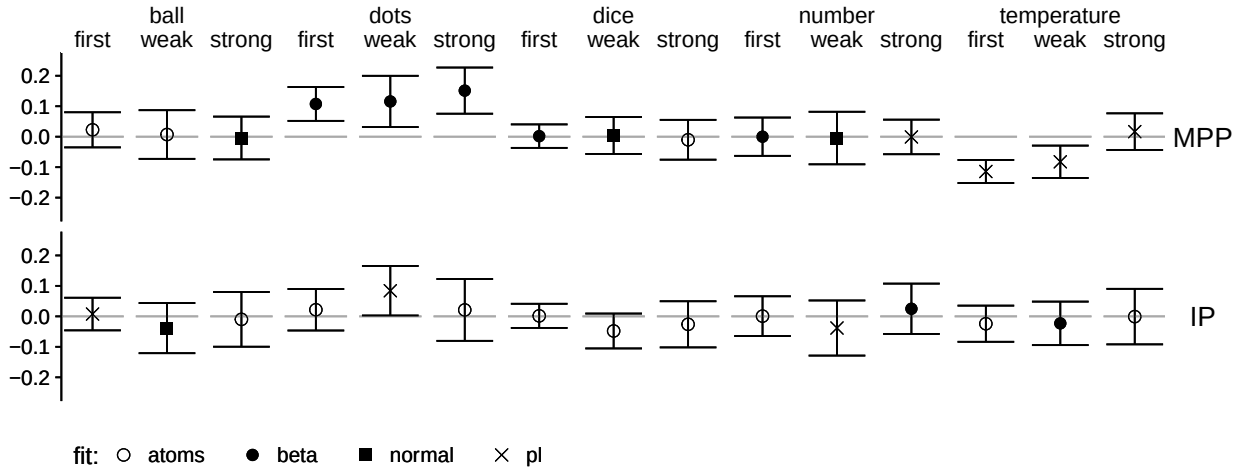



Figure 6: **First-order calibration.** The target variable is $Z := PIT(\mathbb{P}, y) - 0.5$. Throughout, the error bars show 95% confidence intervals. Participants are pooled in the first round of elicitation (*first*), and distinguished after receiving the information update (*weak* and *strong*).

IP treatment shows no evidence against first order calibration. The observed pattern is consistent with the intervals providing helpful anchors in the first round that alleviate biases in perception. The results after the information update suggest that the first order miscalibration in the MPP treatment can be reduced by additional information.

A similar analysis of the variance of the PIT can be used to analyze if predictive distributions over- or underpredict uncertainty. If the outcome is consistently more extreme (smaller or larger) than postulated by the predictive distribution, the PIT is more likely to be close to either zero or one. If the outcome is consistently less extreme, the PIT is on average closer to 0.5. A uniformly distributed PIT has variance $\frac{1}{12}$. We call a sample of predictive distributions *second order calibrated* if the variance of the PIT is equal to $\frac{1}{12}$. Predictive distributions that are overconfident (too narrow) have a larger PIT variance, predictive distributions with an overly wide spread have smaller PIT variance.

Figure 7 shows the transformed variance of the PIT. A value of 0 denotes a well-calibrated second moment of the PIT. A value of $-1$ is obtained for the minimal PIT variance of 0 and a value of 1 would imply that the variance of the PIT is twice as large as under second order calibration. In general, second order calibration is less robust with respect to the chosen fit than first order calibration. (See Figure 11 and Table 4 in Appendix B for results depicting all fits simultaneously.)

In line with other studies, elicitation of IP lead to overly confident predictive distributions[16],

[16]A fact referred to as overconfidence, or more specifically as overprecision (compare Alpert and Raiffa,

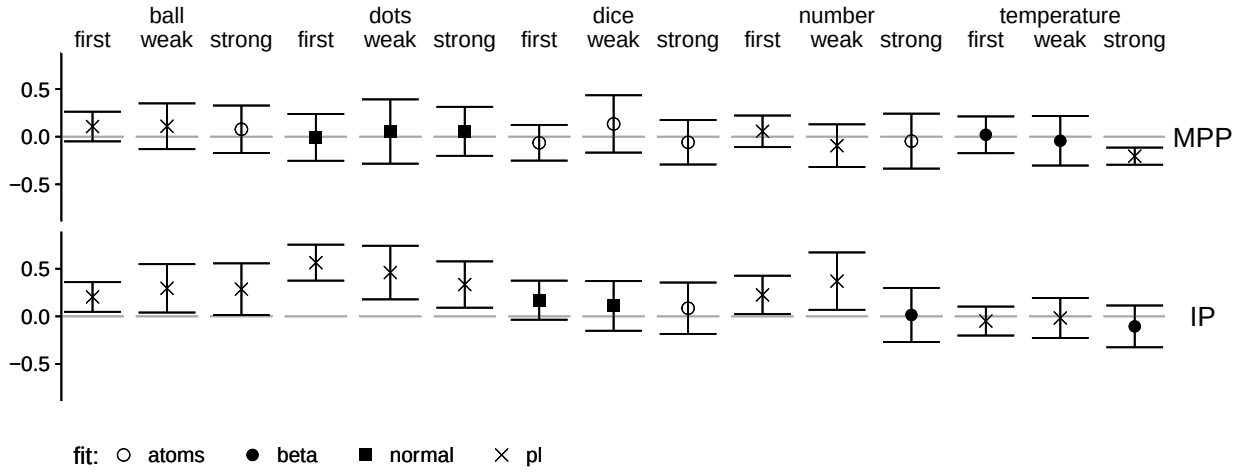



Figure 7: **Second order calibration.** The target variable is $Z := 12(PIT(\mathbb{P}, y) - \hat{m}(d))^2 - 1$, where $\hat{m}(d)$ constitutes the estimated mean for each domain and information update. A positive value indicates overconfidence.

with many domains indicating variances that are at least 25% larger than under calibration even after choosing the most suitable fit. This overconfidence seems to be absent for the elicitation of MPP. The information updates do not seem to affect this result. Figure 7 shows consistent evidence for overconfidence in eight out of fifteen domains for the IP treatment and in no domain for the MPP treatment. Several mechanisms could explain this difference in overconfidence. The four applied fits might not be sufficiently flexible. Further, probability assessments of outer intervals are prone to rounding towards zero and thereby underrepresenting uncertainty. Finally, participants may simply report more focused probabilistic statements (for IP) than indirectly inferred from their behavior (for MPP).

To sum up, MPP are more prone to first order miscalibration for uninformed participants. Under additional information, the first order calibration improves. Considering second order calibration, IP reports are prone to overconfidence. This does not apply to MPP reports.

At this point it is worth noting that the calibration analysis above considers different elicitation and fit methods for a random participant and a specific domain. As each participant encountered each 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 miscalibration is, for example, perfectly consistent with participants following some rule of thumb that is calibrated across the relevant decisions in daily life, while being uncalibrated at the specific tasks encountered in our experiment.

1982; Brenner *et al.*, 1996; Fox and Clemen, 2005; Haran *et al.*, 2010; Lichtenstein and Fischhoff, 1977).

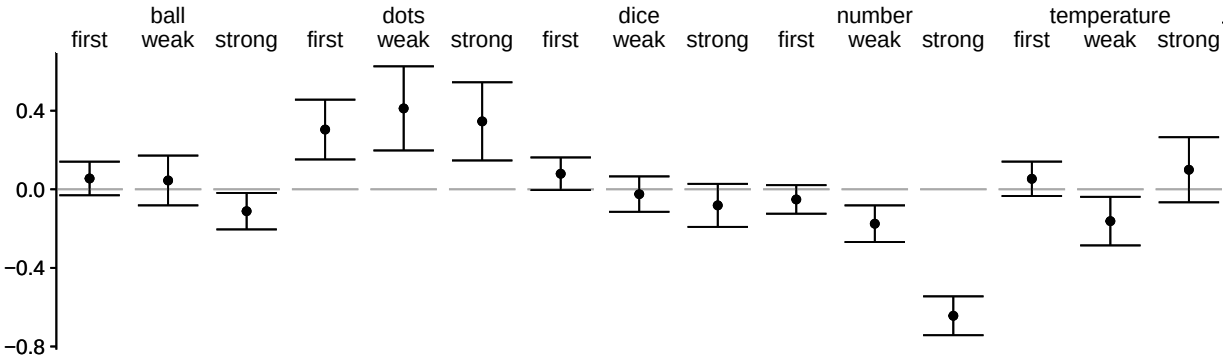



Figure 8: **Difference in accuracy.** The target variable is denoted as $Z := CRPS(\mathbb{P}, y)$. The plot depicts differences in average scores for the best performing fit divided by the average score in the IP treatment, so that negative values indicate superior accuracy of MPP.

## 4.2 Accuracy

Accuracy and calibration are different concepts, which do not necessarily agree as to which elicitation method performs best. An elicitation method can provide perfectly calibrated distributions that provide little useful information because they lack sharpness. We evaluate accuracy with *proper scoring rules* that can assess sharpness and calibration simultaneously (Gneiting and Raftery, 2007). Among calibrated distributions, larger information sets imply lower[17] expected scores (Holzmann and Eulert, 2014). Thus, average scores can rank elicitation methods in terms of their information content.

We argue against using local proper scoring rules like the logarithmic score that are unstable with respect to the distributional assumption as they depend only on the density at the realized value. Instead, we propose to use the Continuous Ranked Probability Score (CRPS) which can be stated with $\alpha$-quantiles $q_\alpha(\mathbb{P})$ of the predictive distribution (Laio and Tamea, 2007) as

$$CRPS(\mathbb{P}, y) = \int_0^1 (q_\alpha(\mathbb{P}) - y)(\mathbb{1}(y \leq q_\alpha(\mathbb{P})) - \alpha) d\alpha.$$

This lends the CRPS a convenient interpretation: The CRPS depicts the average proper linear score across the quantiles of the predictive distribution. As such, lower values indicate supreme accuracy and for predictive distributions that put all mass on one point, the CRPS reduces to the absolute error.

The average CRPS is depicted in Figure 8. Before the information update both methods seem

[17]Note that we use positively oriented scores for elicitation in Section 2 and will use negatively oriented scores for comparing predictive performance in the following.

to provide equally accurate predictive distributions, except for the dots domain, where the first moment miscalibration analyzed in Section 4.1 translates into worse scores for MPP. After the information update, the evidence suggests that MPP provide at least equally accurate predictive distributions for all but the dots domain. In most domains, the difference in accuracy is more favorable for MPP under more heterogeneous information (after the strong information update). Arguably, IP could potentially benefit here if thresholds adapted to the current information.

To summarize, the first order miscalibration analyzed in Section 4.1 led to less accurate beliefs in the dots domain. In all other domains, MPP provided at least as accurate beliefs. Throughout, strong information was positively correlated with the relative performance between the MPP and IP reports.

Thus, in terms of both, calibration and accuracy, MPP performed well under heterogeneous information sets. 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 Willingness to Pay

For economic modelling an elicitation mechanism should provide accurate evidence on the belief of an agent. The elicited probabilities cannot be compared to the unobservable beliefs. However, 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 economic action*. In our experiment we elicited if the participant is willing to pay a certain amount of credits for the unknown outcome. As described in Section 3.2, we offered the participants an amount of credits in exchange for receiving the uncertain outcome $y$ in credits after the second round of belief elicitation.[18]

Figure 9 depicts the difference in average consistency. On average both elicitation methods provide a high consistency rate of about 50% to 85% across both elicitation methods (compare Figure 13 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 likely to be consistent with subsequent behavior.

[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 willingness to pay between the two different elicitation mechanisms.

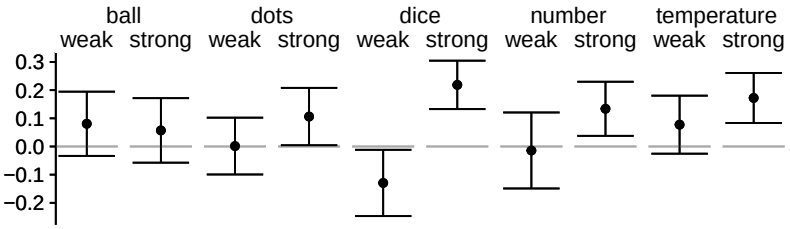



Figure 9: **Difference in consistency with willingness to pay.** The plot depicts the average difference between the best performing fits in the ratio of consistent behavior normalized by the consistency rate in the IP treatment, where positive values indicate superior consistency of MPP. The subjective distribution and the willingness to pay decision are consistent, 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 responses are inconsistent.

Table 2: **Perception of experiment**

| | 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 |

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 ($n = 282$). The $p$-values are derived from an unpaired two-sample $t$-test with two-sided alternative.

## 4.4 Applicability

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.

Table 2 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 convenient with MPP elicitation, an important property for preventing drop-outs and ensuring high take-up rates in panel data sets.

# 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 modeling 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. 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.

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, forthcoming; 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 eliciting point predictions is convenient and reliable. Applying a wide range of criteria, we find that both methods have their merits and drawbacks so that an application of one or the other should be evaluated depending on the domain and objective of the elicitation. We find that MPP is more adequate under heterogeneous beliefs that complicate the construction of uniformly optimal pre-defined intervals. Under homogeneous beliefs and if anchoring is desired or no concern, IP may be more suitable. In contrast to elicitation using IP, we do not observe overconfidence (too narrow distributions) when asking for MPP, suggesting that overconfidence may at least partly be a relic of the elicitation design, instead of an intrinsic property of preferences. 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.

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 MPP can still be useful with hypothetical payoffs. In an additional experiment presented in Appendix C, we find that such hypothetical questions provide informative reports, 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 that allow beliefs to be represented by a single probability measure. For a related mechanism that considers ambiguity averse preferences 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 25, 50 and 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. 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. Another interesting question is, how belief elicitation via MPP performs with different subject pools.

## 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.

Armantier, O., 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.

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.

Carroll, G. D. (forthcoming). Robust incentives for information acquisition. *Journal of Economic Theory*.

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.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, **147** (2), 278–292.

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.

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

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, **39** (4), 863–883.

—, 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., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69** (2), 243–268.

— 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.

— and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, **7**, 1747–1782.

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. (2004). The online recruitment system ORSEE 2.0 - a guide for the organization of experiments in economics, working Paper Series in Economics, University of Cologne.

Grushka-Cockayne, Y., Jose, V. R. R. and Lichtendahl Jr, K. C. (2016). Ensembles of overfit and overconfident forecasts. *Management Science*, **63** (4), 1110–1130.

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.

Haruvy, E., Lahav, Y. and Noussair, C. (2007). Traders' expectations in asset markets: Experimental evidence. *American Economic Review*, **97** (5), 1901–1920.

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.

Holzmann, H. and Eulert, M. (2014). The role of the information set for forecasting with applications to risk management. *The Annals of Applied Statistics*, **8** (1), 595–621.

Hora, S. C. (2004). Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science*, **50** (5), 597–604.

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.

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.

Lichtendahl, K. C., Grushka-Cockayne, Y. and Winkler, R. L. (2013). Is it better to average probabilities or quantiles? *Management Science*, **59** (7), 1594–1611.

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.

Palfrey, T. R. and Wang, S. W. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behavior & Organization*, **71** (2), 98–109.

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.

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 statement 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 $\Omega$. 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 $M_E m$ that pays prize $M$ if the event $E = \{(y, r) \in \Omega \mid s(x, y) > r\}$ realizes and $m$ otherwise.

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

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

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

Regularity Conditions 1 hold for an expected 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 $\arg\max_{p \in \mathcal{P}} M_{s(p,y)>3Br} m$ is a density of $\mathbb{P}_0$.*

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

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

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*

$$s(x,y) = \frac{1}{ne} \sum_{i=1}^{n} \max(s_i(x_i, y), 0)$$

*with*

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

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

*(i) It holds that*

$$x^* \to (q_{\alpha_1}(\mathbb{P}_0), \dots, q_{\alpha_n}(\mathbb{P}_0)) \text{ for } e \to \infty.$$

*(ii) Consider an optimal response $x_i^*$. If $F(x_i^* - e/b_i) < c_1$ and $1 - F(x_i^* + e/a_i) < c_2$, then*

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

*for some level $\alpha^*$ such that*

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

*Proof.* As $0 \leq s(x,y) \leq 1$, the agent maximizes

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

Trivially, $\arg\max_{x\in\mathcal{X}} \mathbb{E}[\sum s(x,y)] = \sum \arg\max_{x_i\in\mathcal{X}_i} \mathbb{E}[s_i(x_i,y)]$ as the score $s_i$ does not depend on $x_j$ with $j \neq i$. We can investigate $\arg\max_{x_i\in\mathcal{X}_i} \mathbb{E}[s_i(x_i,y)]$ for fixed $i = 1, \dots, n$.

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

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

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

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

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

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

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

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

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

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

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 $e > B \max(a_1, b_1, \ldots, a_k, b_k)$, 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 $e > Ba$ 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 $f(x_i - e/b) = 0$ and $f(x_i + e/a) = 0$ 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 $\alpha^* \in (0,1)$ such that $q^*_\alpha(\mathbb{P}_0) = c$.

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

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

It follows that

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

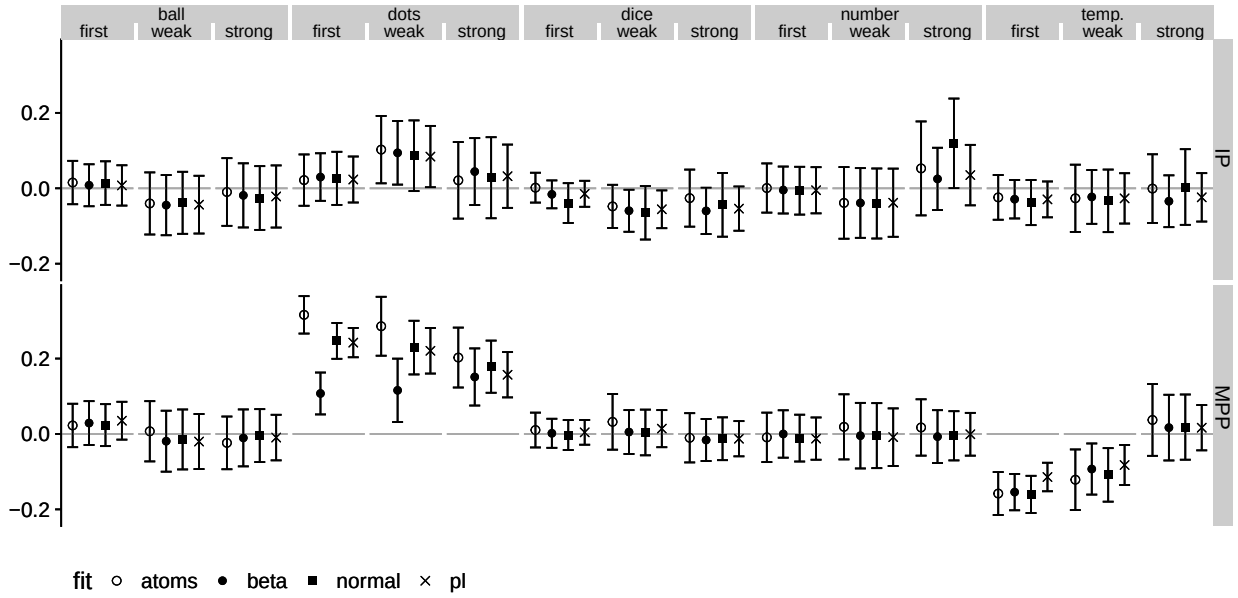
A similar arguments gives $x_i^* \to \max(F)$ for $a_i \to \infty$ and $\frac{a_i}{e} \to \infty$.

□

# 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 10: **First-order calibration - PIT average**



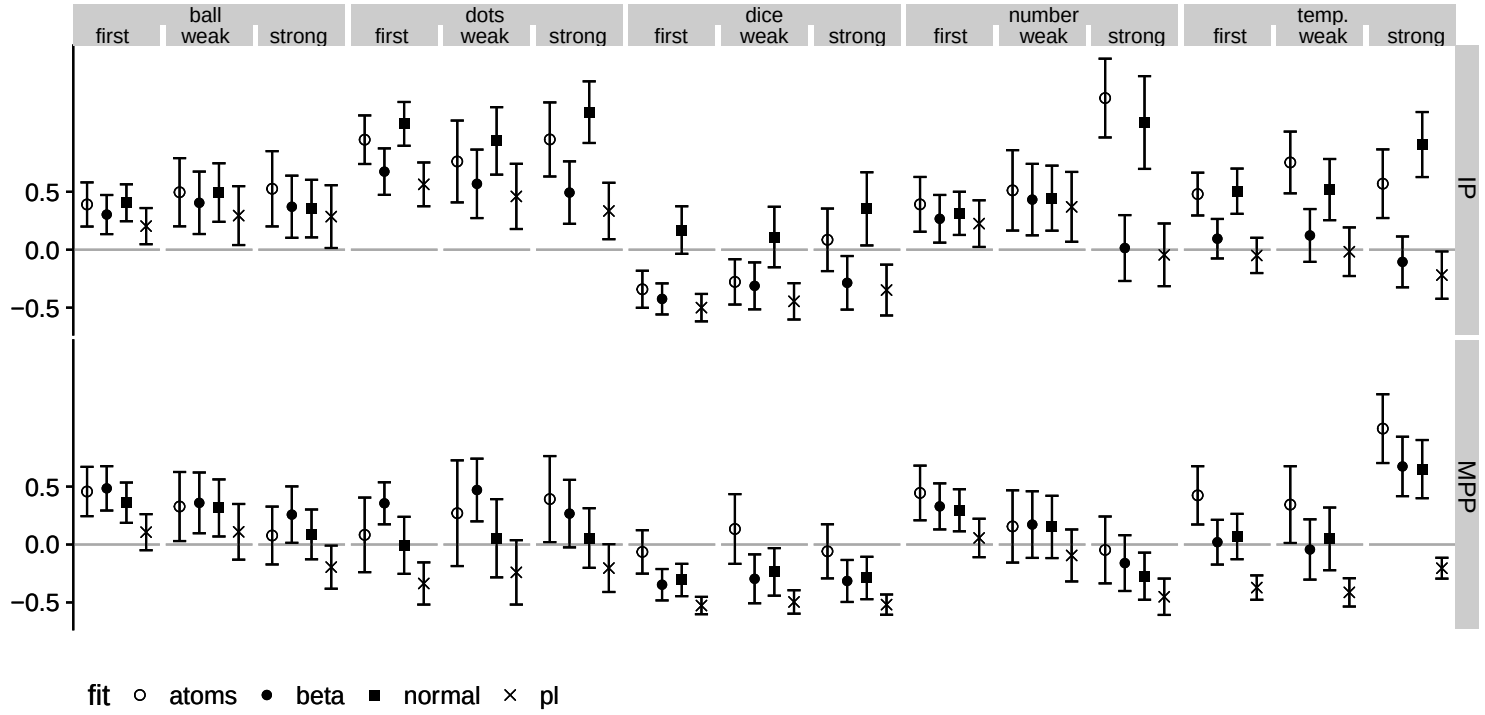

The target variable is $Z := PIT(\mathbb{P}, y) - 0.5$.

Table 3: **First order calibration - PIT average**

| | IP | | | | MPP | | | |
|---|---|---|---|---|---|---|---|---|
| | atoms | beta | norm | pl | atoms | beta | norm | pl |
| $info = strong$ | | | | | | | | |
| ball | -0.01(0.83) | -0.02(0.66) | -0.03(0.54) | -0.02(0.6) | -0.02(0.5) | -0.01(0.78) | 0(0.9) | -0.01(0.75) |
| dots | 0.02(0.68) | 0.04(0.32) | 0.03(0.6) | 0.03(0.45) | 0.2(0) | 0.15(0) | 0.18(0) | 0.16(0) |
| dice | -0.03(0.49) | -0.06(0.06) | -0.04(0.3) | -0.05(0.07) | -0.01(0.76) | -0.02(0.57) | -0.01(0.66) | -0.01(0.59) |
| number | 0.05(0.4) | 0.02(0.55) | 0.12(0.05) | 0.03(0.38) | 0.02(0.65) | -0.01(0.84) | 0(0.88) | 0(0.98) |
| temperature | 0(0.99) | -0.03(0.32) | 0(0.95) | -0.02(0.46) | 0.04(0.44) | 0.02(0.7) | 0.02(0.68) | 0.02(0.58) |
| $info = weak$ | | | | | | | | |
| ball | -0.04(0.33) | -0.04(0.27) | -0.04(0.35) | -0.04(0.26) | 0.01(0.86) | -0.02(0.64) | -0.01(0.72) | -0.02(0.59) |
| dots | 0.1(0.03) | 0.09(0.03) | 0.09(0.07) | 0.08(0.04) | 0.29(0) | 0.12(0.01) | 0.23(0) | 0.22(0) |
| dice | -0.05(0.1) | -0.06(0.04) | -0.06(0.07) | -0.06(0.03) | 0.03(0.39) | 0.01(0.86) | 0(0.9) | 0.01(0.57) |
| number | -0.04(0.42) | -0.04(0.4) | -0.04(0.39) | -0.04(0.4) | 0.02(0.66) | 0(0.92) | 0(0.92) | -0.01(0.82) |
| temperature | -0.03(0.55) | -0.02(0.52) | -0.03(0.43) | -0.03(0.43) | -0.12(0) | -0.09(0.01) | -0.11(0) | -0.08(0) |
| $first\ round$ | | | | | | | | |
| ball | 0.02(0.6) | 0.01(0.77) | 0.01(0.64) | 0.01(0.78) | 0.02(0.44) | 0.03(0.33) | 0.02(0.41) | 0.04(0.17) |
| dots | 0.02(0.53) | 0.03(0.35) | 0.03(0.46) | 0.02(0.45) | 0.32(0) | 0.11(0) | 0.25(0) | 0.24(0) |
| dice | 0(0.93) | -0.02(0.39) | -0.04(0.14) | -0.01(0.4) | 0.01(0.66) | 0(0.93) | 0(0.89) | 0(0.8) |
| number | 0(0.98) | 0(0.88) | -0.01(0.84) | -0.01(0.87) | -0.01(0.79) | 0(1) | -0.01(0.73) | -0.01(0.66) |
| temperature | -0.02(0.42) | -0.03(0.26) | -0.04(0.21) | -0.03(0.22) | -0.16(0) | -0.15(0) | -0.16(0) | -0.11(0) |

The dependent variable is $Z := PIT(\mathbb{P}, y) - 0.5$. Tests are two-sided one sample $t$-tests.

Figure 11: **Second order calibration - PIT variance**



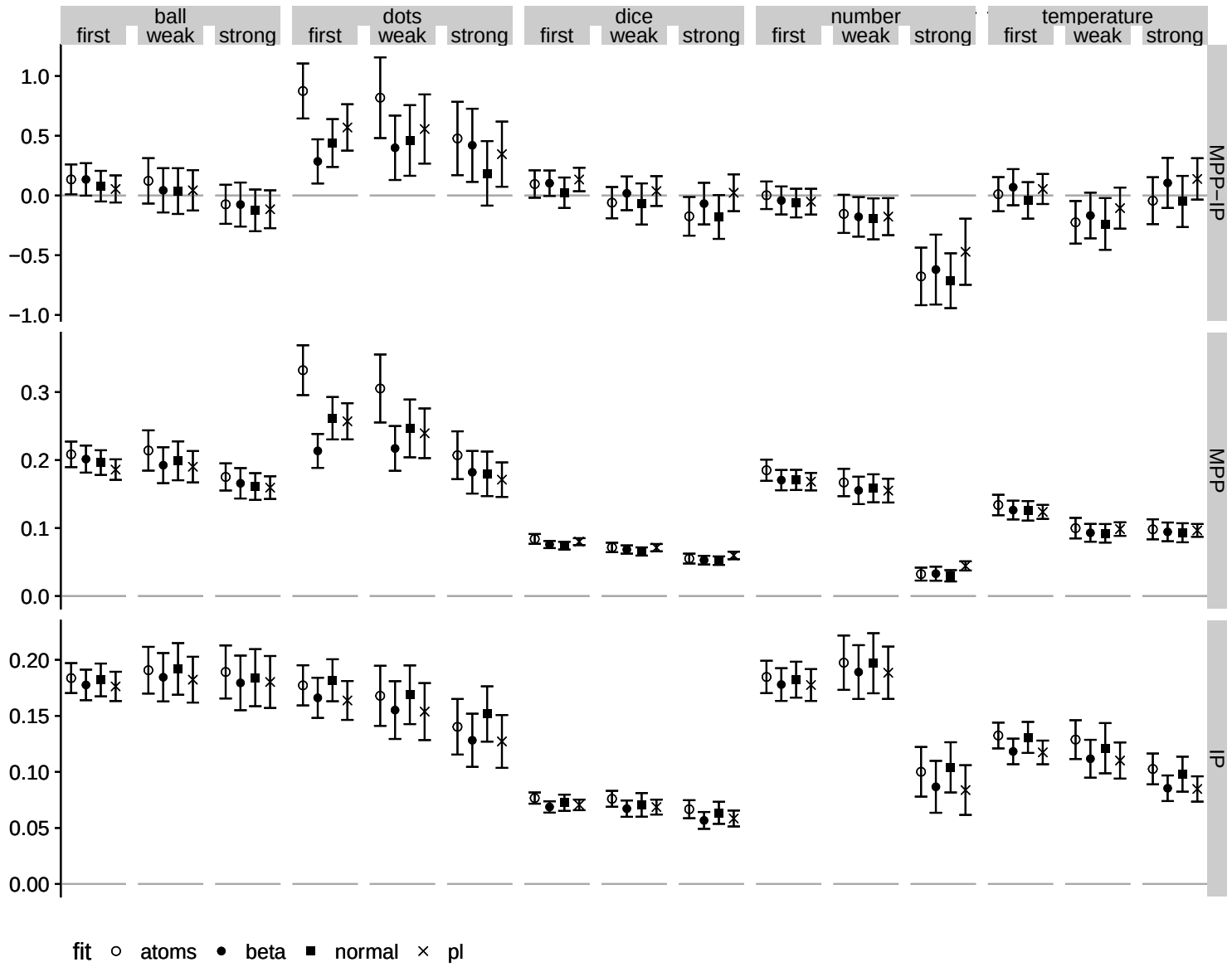

The target variable is $Z := 12(PIT(\mathbb{P}, y) - \hat{m}(d))^2 - 1$, where $\hat{m}(d)$ constitutes the estimated mean for each domain and information update.

Table 4: **Second order calibration - PIT variance**

| | IP | | | | MPP | | | |
|---|---|---|---|---|---|---|---|---|
| | atoms | beta | norm | pl | atoms | beta | norm | pl |
| $info = strong$ | | | | | | | | |
| ball | 0.5(0) | 0.4(0.01) | 0.4(0.01) | 0.3(0.04) | 0.1(0.54) | 0.3(0.04) | 0.1(0.42) | -0.2(0.04) |
| dots | 1(0) | 0.5(0) | 1.2(0) | 0.3(0.01) | 0.4(0.04) | 0.3(0.07) | 0.1(0.67) | -0.2(0.05) |
| dice | 0.1(0.53) | -0.3(0.02) | 0.4(0.03) | -0.3(0) | -0.1(0.62) | -0.3(0) | -0.3(0) | -0.5(0) |
| number | 1.3(0) | 0(0.92) | 1.1(0) | 0(0.75) | 0(0.75) | -0.2(0.19) | -0.3(0.01) | -0.5(0) |
| temperature | 0.6(0) | -0.1(0.34) | 0.9(0) | -0.2(0.04) | 1(0) | 0.7(0) | 0.7(0) | -0.2(0) |
| $info = weak$ | | | | | | | | |
| ball | 0.5(0) | 0.4(0) | 0.5(0) | 0.3(0.02) | 0.3(0.03) | 0.4(0.01) | 0.3(0.01) | 0.1(0.37) |
| dots | 0.8(0) | 0.6(0) | 0.9(0) | 0.5(0) | 0.3(0.24) | 0.5(0) | 0.1(0.75) | -0.2(0.09) |
| dice | -0.3(0.01) | -0.3(0) | 0.1(0.41) | -0.4(0) | 0.1(0.38) | -0.3(0.01) | -0.2(0.02) | -0.5(0) |
| number | 0.5(0) | 0.4(0.01) | 0.4(0) | 0.4(0.02) | 0.2(0.32) | 0.2(0.24) | 0.2(0.26) | -0.1(0.4) |
| temperature | 0.8(0) | 0.1(0.29) | 0.5(0) | 0(0.87) | 0.3(0.04) | 0(0.74) | 0(0.72) | -0.4(0) |
| $first\ round$ | | | | | | | | |
| ball | 0.4(0) | 0.3(0) | 0.4(0) | 0.2(0.01) | 0.5(0) | 0.5(0) | 0.4(0) | 0.1(0.18) |
| dots | 0.9(0) | 0.7(0) | 1.1(0) | 0.6(0) | 0.1(0.61) | 0.4(0) | 0(0.95) | -0.3(0) |
| dice | -0.3(0) | -0.4(0) | 0.2(0.1) | -0.5(0) | -0.1(0.5) | -0.3(0) | -0.3(0) | -0.5(0) |
| number | 0.4(0) | 0.3(0.01) | 0.3(0) | 0.2(0.03) | 0.4(0) | 0.3(0) | 0.3(0) | 0.1(0.5) |
| temperature | 0.5(0) | 0.1(0.27) | 0.5(0) | 0(0.52) | 0.4(0) | 0(0.84) | 0.1(0.49) | -0.4(0) |

The target variable is $Z := 12(PIT(\mathbb{P}, y) - \hat{m}(d))^2 - 1$, where $\hat{m}(d)$ constitutes the estimated mean for each domain and information update. Tests are two-sided one sample $t$-tests.

Figure 12: **Difference in accuracy - CRPS**



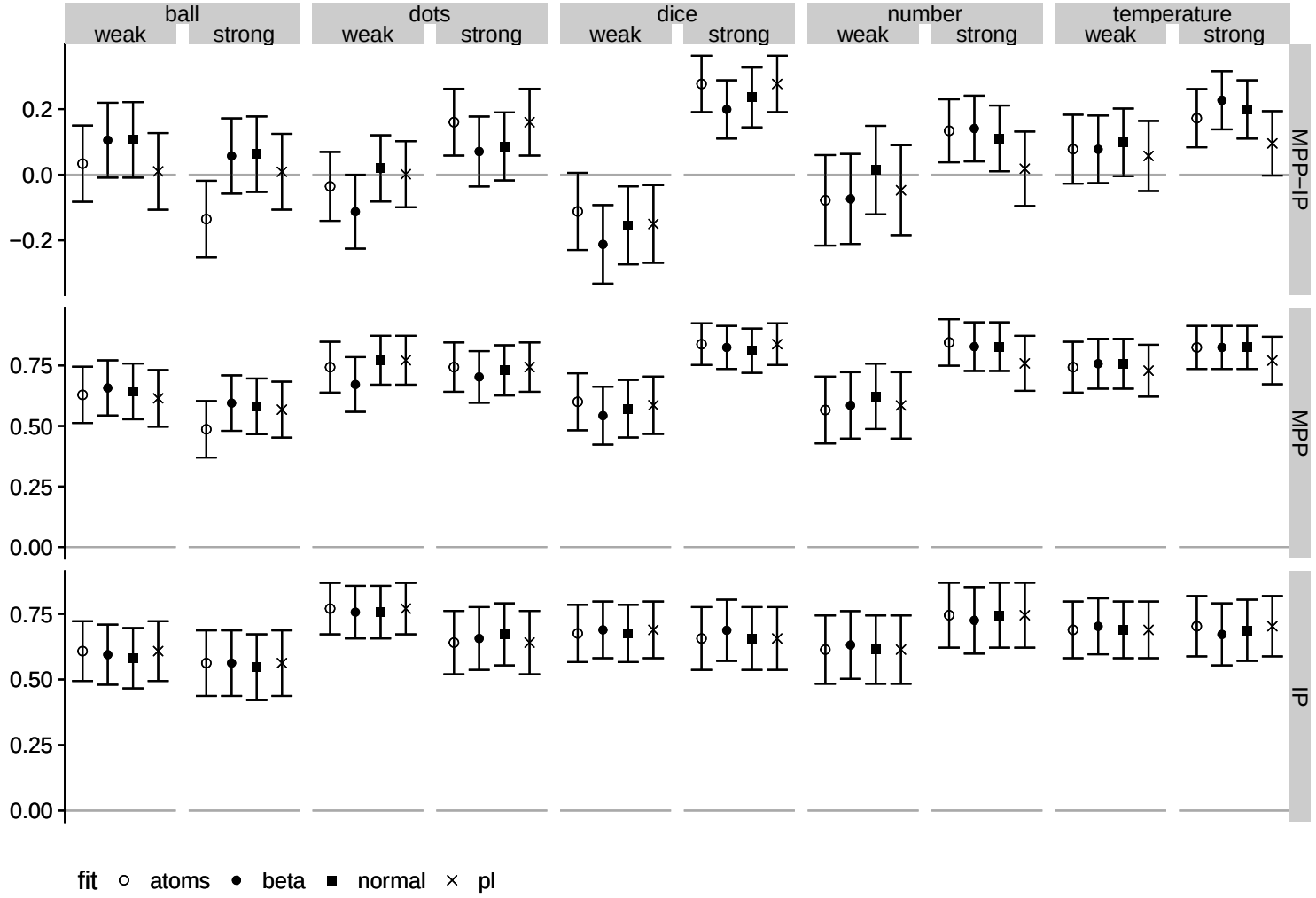

The target variable is denoted as $Z := CRPS(\mathbb{P}, y)$. The $MPP$ and $IP$ value indicate average CRPS normalized by the highest possible CRPS within each domain. The $MPP - IP$ value indicates the average score difference normalized by the $IP$ score, where negative values indicate superior accuracy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

Table 5: **Difference in accuracy - CRPS**

| | IP | | | | MPP - IP | | | |
|---|---|---|---|---|---|---|---|---|
| | atoms | beta | norm | pl | atoms | beta | norm | pl |
| $info = strong$ | | | | | | | | |
| ball | 11.3 | 10.8 | 11 | 10.8 | -0.8(0.46) | -0.8(0.5) | -1.4(0.24) | -1.3(0.22) |
| dots | 14 | 12.8 | 15.2 | 12.7 | 6.7(0.01) | 5.4(0.02) | 2.8(0.25) | 4.4(0.04) |
| dice | 4 | 3.4 | 3.8 | 3.5 | -0.7(0.08) | -0.2(0.52) | -0.7(0.1) | 0.1(0.81) |
| number | 10 | 8.7 | 10.4 | 8.4 | -6.8(0) | -5.4(0) | -7.4(0) | -3.9(0.01) |
| temperature | 3.1 | 2.6 | 2.9 | 2.5 | -0.1(0.72) | 0.3(0.4) | -0.1(0.7) | 0.4(0.18) |
| $info = weak$ | | | | | | | | |
| ball | 11.4 | 11.1 | 11.5 | 10.9 | 1.4(0.28) | 0.5(0.7) | 0.4(0.75) | 0.5(0.65) |
| dots | 16.8 | 15.5 | 16.9 | 15.4 | 13.7(0) | 6.2(0.01) | 7.8(0.01) | 8.6(0) |
| dice | 4.6 | 4 | 4.2 | 4.1 | -0.3(0.45) | 0.1(0.83) | -0.3(0.5) | 0.2(0.61) |
| number | 19.7 | 18.9 | 19.7 | 18.8 | -3(0.11) | -3.4(0.08) | -3.8(0.06) | -3.3(0.06) |
| temperature | 3.9 | 3.4 | 3.6 | 3.3 | -0.9(0.04) | -0.6(0.15) | -0.9(0.07) | -0.4(0.31) |
| $first\ round$ | | | | | | | | |
| ball | 11 | 10.7 | 10.9 | 10.6 | 1.5(0.08) | 1.4(0.1) | 0.9(0.31) | 0.6(0.42) |
| dots | 17.7 | 16.6 | 18.2 | 16.4 | 15.5(0) | 4.7(0.01) | 8(0) | 9.3(0) |
| dice | 4.6 | 4.1 | 4.3 | 4.2 | 0.4(0.17) | 0.4(0.11) | 0.1(0.76) | 0.6(0.03) |
| number | 18.5 | 17.8 | 18.2 | 17.8 | 0(0.98) | -0.7(0.56) | -1.2(0.38) | -0.9(0.43) |
| temperature | 4 | 3.5 | 3.9 | 3.5 | 0(0.9) | 0.2(0.45) | -0.2(0.66) | 0.2(0.48) |

The target variable is denoted as $Z := CRPS(\mathbb{P}, y)$. The $IP$ value indicate average CRPS within each domain for the $IP$ treatment. The $MPP - IP$ value indicates the average score difference normalized by the $IP$ score, where negative values indicate superior accuracy of MPP. Tests are two-sided two sample $t$-tests.

Figure 13: **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.

Table 6: **Consistency of willingness to pay and subjective probabilities**

| | IP | | | | MPP - IP | | | |
|---|---|---|---|---|---|---|---|---|
| | atoms | beta | norm | pl | atoms | beta | norm | pl |
| $info = strong$ | | | | | | | | |
| ball | 0.56 | 0.56 | 0.55 | 0.56 | -0.08(0.38) | 0.03(0.71) | 0.03(0.69) | 0.01(0.95) |
| dots | 0.64 | 0.66 | 0.67 | 0.64 | 0.1(0.2) | 0.05(0.56) | 0.06(0.46) | 0.1(0.2) |
| dice | 0.66 | 0.69 | 0.66 | 0.66 | 0.18(0.02) | 0.14(0.06) | 0.15(0.04) | 0.18(0.02) |
| number | 0.75 | 0.73 | 0.75 | 0.75 | 0.1(0.2) | 0.1(0.21) | 0.08(0.3) | 0.01(0.87) |
| temperature | 0.7 | 0.67 | 0.69 | 0.7 | 0.12(0.1) | 0.15(0.04) | 0.14(0.06) | 0.07(0.38) |
| $info = weak$ | | | | | | | | |
| ball | 0.61 | 0.59 | 0.58 | 0.61 | 0.02(0.8) | 0.06(0.44) | 0.06(0.45) | 0.01(0.94) |
| dots | 0.77 | 0.76 | 0.76 | 0.77 | -0.03(0.7) | -0.09(0.26) | 0.01(0.84) | 0(0.99) |
| dice | 0.68 | 0.69 | 0.68 | 0.69 | -0.08(0.35) | -0.15(0.07) | -0.1(0.2) | -0.1(0.2) |
| number | 0.61 | 0.63 | 0.61 | 0.61 | -0.05(0.61) | -0.05(0.62) | 0.01(0.93) | -0.03(0.76) |
| temperature | 0.69 | 0.7 | 0.69 | 0.69 | 0.05(0.48) | 0.05(0.47) | 0.07(0.37) | 0.04(0.61) |

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 IP coefficient depicts baseline averages for the IP treatment and the MPP - IP coefficient shows average difference in the ratio of consistent behavior, where positive values indicate superior consistency of MPP. Tests are two-sided two sample $t$-tests.

# 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.

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 S1.5 of the supplementary document.) As such we can investigate if hypothetical incentives can be used to elicit beliefs if incentives are infeasible. 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, subsequently divided this interval into seven equidistant subintervals, and finally elicited the nine probabilities on the generated intervals.

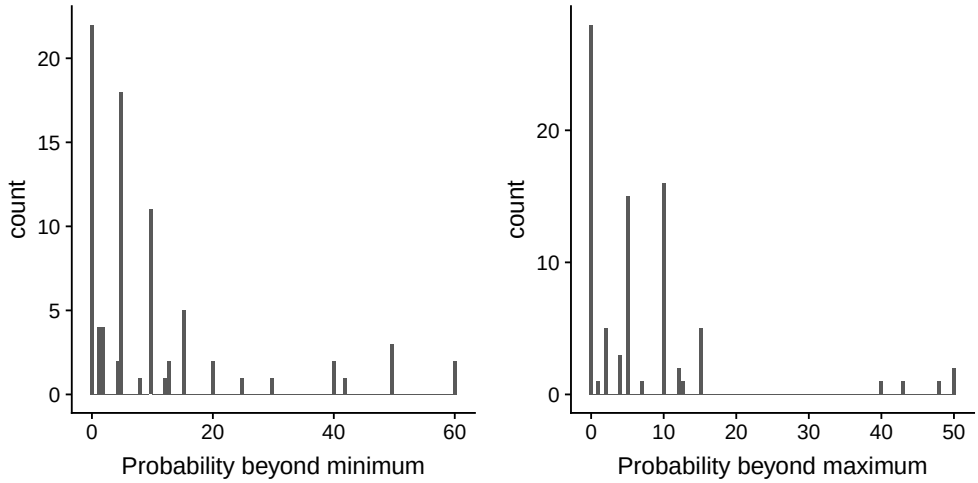
## C.1 Reports for minimum and maximum

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 minimum and maximum are not consistent.

In Figure 14, we see that while most participants report positive probabilities, those 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 **??**). And indeed, we see in Figure 15 that the lowest elicited quantile and the minimum report behave in a similar manner. Before the information update only a small 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

Figure 14: **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.

strong information does so more heavily.

## C.2 Calibration and accuracy

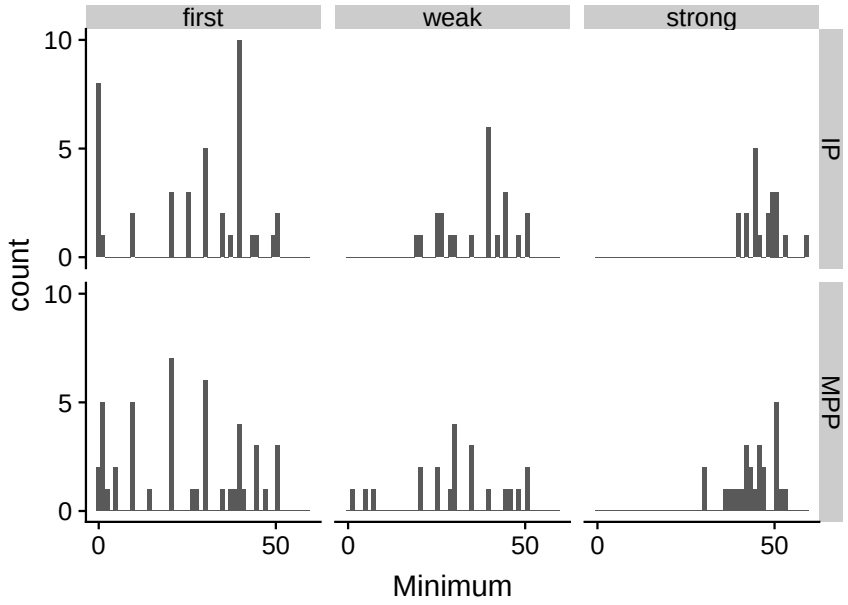
We apply the same analysis as in the main part of the paper. Two observations for each elicitation method were deleted as the mean of their predictive distribution was below 30 or above 70 indicating that the participants either misunderstood or put no thought in the task. Figure 16 shows that most domains exhibit no evidence against first order calibration with the exception of the first round of MPP elicitation, where there is some weak evidence against it.

Figure 17 suggests that participants overestimate uncertainty for the 100 coin flips, but are mostly able to judge the distribution for the 10 remaining coin flips. This underconfidence is stronger for MPP. A possible explanation is that participants overestimate the uncertainty of a binomial distribution with many observations (Benjamin *et al.*, 2017). This effect may be partially cancelled out by the tendency to report overconfidently with IP.

The average CRPS scores in Figure 18 show no evidence against equal accuracy, however MPP performed up to 30% better before the information update and after the strong information update.

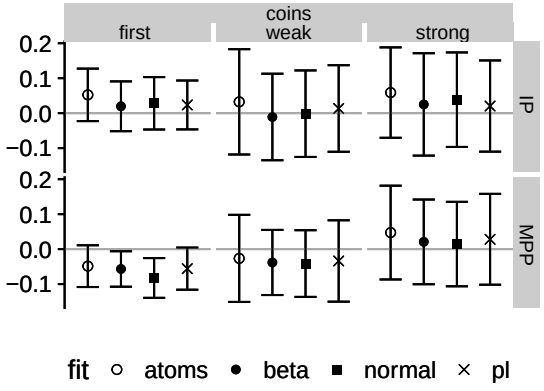
In summary, the experiment provides no evidence that the absence of incentives, the elicitation of additional CDF points, and flexible support for the IP treatment fundamentally change the conclusions of the main experiment.

Figure 15: **Minimum reports**



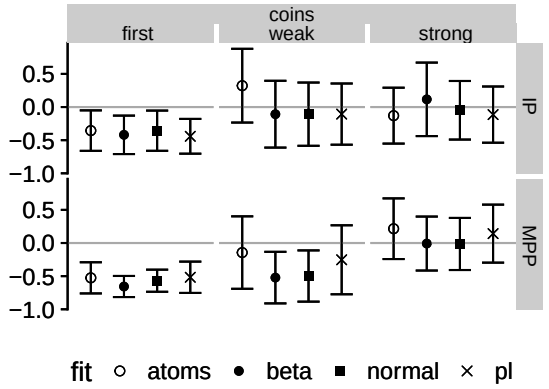

Histogram of minimum reports for IP and extreme quantile reports for MPP in the first round, and after the weak and strong information update.

Figure 16: **First-order calibration - PIT average**



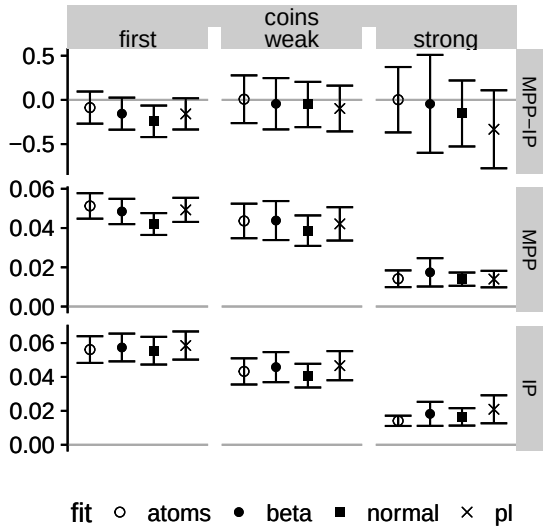

The target variable is $Z := PIT(\mathbb{P}, y) - 0.5$.

Figure 17: **Second order calibration - PIT variance**




The target variable is $Z := 12(PIT(\mathbb{P}, y) - \hat{m}(d))^2 - 1$, where $\hat{m}(d)$ constitutes the estimated mean for each domain and information update.

Figure 18: **Difference in accuracy - CRPS**




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