Eliciting Beliefs as Distributions in Online Surveys

Lucas Leemann[1], Lukas F. Stoetzer[2] and Richard Traunmüller[3]

[1] Department of Political Science, University of Zürich, Zürich, Switzerland. Email: leemann@ipz.uzh.ch
[2] Humboldt University of Berlin, Cluster of Excellence SCRIPTS, Berlin, Germany. Email: lukas.stoetzer@hu-berlin.de
[3] School of Social Sciences, University of Mannheim, Mannheim, Germany. Email: traunmueller@uni-mannheim.de

Abstract

Citizens' beliefs about uncertain events are fundamental variables in many areas of political science. While beliefs are often conceptualized in the form of distributions, obtaining reliable measures in terms of full probability densities is a difficult task. In this letter, we ask if there is an effective way of eliciting beliefs as distributions in the context of online surveys. Relying on experimental evidence, we evaluate the performance of five different elicitation methods designed to capture citizens' uncertain expectations. Our results suggest that an elicitation method originally proposed by Manski (2009) performs well. It measures average citizens' subjective belief distributions reliably and is easily implemented in the context of regular (online) surveys. We expect that a wider use of this method will lead to considerable improvements in the study of citizens' expectations and beliefs.

Keywords: prior elicitation, online survey research, citizen beliefs, measurement

1 Introduction

Citizens' beliefs about uncertain events are fundamental variables in many areas of political science, including work on attitudes (e.g., Zaller and Feldman 1992), cognitive biases (e.g., Gerber and Green 1999; Bartels 2002; Bullock 2009), ambivalence (e.g., Alvarez and Brehm 1997), misinformation (e.g., Berinsky 2017), or citizen forecasts (e.g., Murr 2011; Leiter *et al.* 2018), to name just a few. While beliefs are often theoretically conceptualized in the form of distributions, obtaining reliable measures of these beliefs in terms of full probability densities is a difficult task (Savage 1971; Garthwaite, Kadane, and O'Hagan 2005; Goldstein and Rothschild 2014). Most survey questions are focused on the first moment of an underlying distribution and thus miss important information about beliefs' variance or uncertainty.

The question we ask in this letter is whether there is an effective way to elicit average citizens' belief distributions in the context of online surveys? This paper discusses five different elicitation methods designed to capture citizens' uncertain expectations. We present experimental evidence and evaluate which question format is best suited to elicit continuous beliefs as distributions from regular (i.e., nonexpert) survey respondents. That is, we are interested in how well these methods capture subjective distributions when compared to a benchmark and which of these methods performs best.

Our results suggest that an elicitation method originally proposed by Manski (2009) performs well. It contains five sequential survey questions that reliably measure average citizens' subjective belief distributions and that are easily implemented in the context of regular online surveys. They are also easy and quick to answer and, hence, not too cost-intensive in online surveys. We expect that a wider use of this method will lead to considerable improvements in the study of citizens' expectations and beliefs and, therefore, to important political science theories. In addition, it should also prove a useful tool to Bayesians who wish to elicit subjective prior distributions from nonexperts (Gill and Walker 2005).

***Political Analysis* (2021)**
vol. 29: 541–553
DOI: 10.1017/pan.2020.42

Published
4 February 2021

Corresponding author
Lucas Leemann

Edited by
Daniel Hopkins

To illustrate the use of the method in an applied example, we elicit people's expectations about the 2020 U.S. presidential election. Eliciting citizens' beliefs is a common element in citizen forecasts (Murr 2011), for which it would be valuable to distinguish between citizens who are more certain (i.e., who have narrow belief distributions) from those who are less certain about the election outcome (i.e., who have wide belief distributions) and weight them accordingly. Hence, we ask respondents to provide their full belief distribution concerning Donald Trump's likely vote share in the November 2020 election. In Section 5, we describe how the elicitation methods discussed in this letter can be applied to this practical task. Based on the Manski question format, we find that respondents expect a popular vote share of 48% for Donald Trump with a standard deviation of 5.5%. We further find considerable differences in both expectations and uncertainties between Democrats (44%, sd 4.6%) and Republicans (52%, sd 8.1%).

The remainder of this letter proceeds as follows. The next section discusses the elicitation process. Section 3 then presents the experimental setup and the five elicitation approaches we evaluate. Section 4 presents the results. Section 5 provides a brief illustration using Trump's vote share in the November 2020 election as an example. Section 6 concludes.

2 Eliciting Beliefs as Distributions

The elicitation of beliefs as distributions has a long tradition in statistics, psychology, and economics. In political science, Bayesians seek to elicit prior distributions from *experts* to inform their statistical models (Gill and Walker 2005; Gill and Freeman 2013). However, the process of eliciting probability distributions described in this literature usually is a time-consuming enterprise that requires careful effort even when it is used to learn about the beliefs of experts who may already be familiar with probabilities.

What makes the elicitation of beliefs so difficult is that average people are not used to expressing themselves in easily quantifiable ways. Many citizens are unlikely to be familiar with the concept of probability and not used to expressing their expectations in terms of distributions. Lengthy elicitation protocols also do not scale well to the number of respondents required for testing political science theories about citizens' expectations and are unlikely to be part of nationally representative surveys. Thus, the central challenge is how to best translate what people think into probability distributions within the confines of standard survey methodology.

Formally, an elicitation process can involve up to four steps (Garthwaite *et al.* 2005). In the *setup* step, the problem is defined and respondents are recruited and trained in the key concepts and procedures. *Elicitation* is the key step where the respondent is asked to provide information about his or her subjective belief. In the *fitting* step, this elicited information is converted into a probability distribution. The final step assesses the *adequacy* of the elicited distribution and provides an opportunity for correction. The challenge we address in this letter is how to implement these steps in the context of regular online surveys, where time and scale concerns as well as limited researcher–respondent interaction render the use of full elicitation protocols impractical.

Traditional elicitation methods come in three basic forms (Spetzler and Stael von Holstein 1975). In each of these three forms, subjects are asked questions and the answers represent points on a cumulative distribution function. In so-called P-methods, subjects are provided with fixed values referring to the quantity of interest and asked to assign *probabilities* attached to these values (e.g., what is the probability that the value is below *x*?). In V-methods, subjects are instead provided with predefined probabilities and asked to assign *values* to them (e.g., at what value are half of the observations below or above that value?). PV-methods are more difficult and simultaneously integrate both approaches. For instance, respondents may be asked to draw a graph of a probability distribution. In this letter, we evaluate several ways to implement these methods with online survey questions.

https://doi.org/10.1017/pan.2020.42 Published online by Cambridge University Press

Given humans' difficulties with probabilities, eliciting beliefs as distributions is as much a psychological problem as it is a statistical one. Many cognitive human biases are well known: representativeness, availability, anchoring biases, the law of small numbers as well as hindsight biases (Tversky and Kahneman 1971, 1973, 1974; Kynn 2008). But it is important to distinguish those biases in beliefs from biases introduced by elicitation methods. Psychological research suggests that while people are generally capable of estimating proportions, modes, and medians, they are less proficient at assessing the means of highly skewed distributions (Peterson and Miller 1964) and often have serious misconceptions about variances (Garthwaite *et al*. 2005). People are reasonably good at quantifying their opinions as credible intervals but have the tendency to imply a greater degree of confidence than is justifiable (Wallsten and Budescu 1983; Cosmides and Tooby 1996).

3 Experimental Set-Up

In the following, we evaluate a set of elicitation question formats. For a proper evaluation of elicitation methods, we need an objective benchmark against which to judge the derived beliefs. To this end, we run a number of experiments where we instill objective distributions and assess which format yields beliefs that are most consistent with these objective benchmark distributions.

Instead of working with arbitrary numbers, we rely on an example of citizens' beliefs about hypothetical election results. Note that this experimental evaluation is different from an actual elicitation process where we would not instill a prior but rather try to elicit a pre-existing belief. To illustrate the actual usage of the method to a political science audience, we provide an example of an actual elicitation process further below. In the following presentation of our experiments, we proceed along the four steps of the elicitation process described in the previous section: setup, elicitation, fitting, and adequacy check.[1]

3.1 The Setup Step

We ran experiments with a total of about 3,600 participants. We relied on Amazon Mechanical Turk (MTurk), which is widely used for scientific purposes (Berinsky, Huber, and Lenz 2012; Mason and Suri 2012; Thomas and Clifford 2017). We recruited workers advertising a study on *surveys*, *opinion polls*, and *charts*. MTurk allowed us to carry out the experiments in a short time period and at a low cost. While MTurk samples may be special, they are comparable to other online samples. Mullinix *et al.* (2015) analyze treatment effects obtained from 20 experiments implemented on a population-based sample and MTurk. The results reveal considerable similarity between effects obtained from convenience and nationally representative population-based samples. Coppock (2018) replicates 15 survey experiments and compares the estimates based on random samples to estimates based on an MTurk sample. In general, the two sets of estimates overlap. These findings may not be surprising because just like MTurk, many online survey panels actually consist of semi-professional survey takers who are experienced in completing online tasks and are perhaps younger and more educated (see, e.g., Berinsky *et al.* 2012, p. 358), in addition to being paid for their participation. Since we are specifically interested in eliciting beliefs in the context of online surveys, these particular respondent characteristics do not concern us much.

We presented respondents with 100 results from hypothetical local elections that we randomly drew from a prespecified distribution. By exposing respondents to these draws, we manipulated the objective belief distribution[2] along two factors: symmetric *versus* asymmetric and small *versus* large variance. We rely on a beta distribution in all four conditions but vary the shape parameters

[1] An "adequacy check" gives respondents the chance to review and check their elicited belief distribution and correct themselves in case this elicited distribution does not adequately represent their belief.

[2] This works under the assumption that in our hypothetical example, respondents hold no priors over the result.

https://doi.org/10.1017/pan.2020.42 Published online by Cambridge University Press

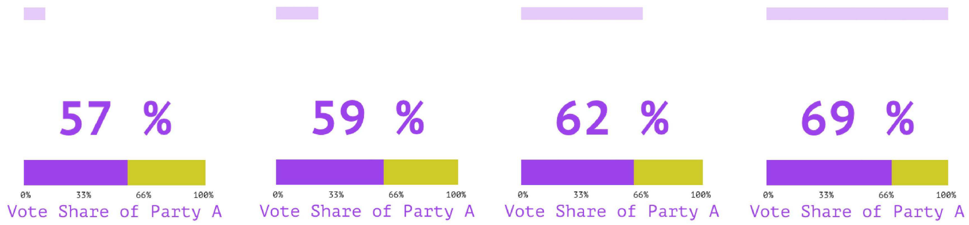

Figure 1. Four still frames from the GIF. Each still frame is shown for about half a second. The black bar at the top is a progress bar.

of the distribution. The symmetric small-variance distribution is $\mathcal{B}(60,60)$ and the respective asymmetric distribution is $\mathcal{B}(60,30)$. For the large-variance condition, we rely on $\mathcal{B}(30,30)$ for the symmetric and on $\mathcal{B}(30,15)$ for the asymmetric distribution. The hypothetical results of 100 election simulations are presented as a short GIF where each frame shows one election outcome and is displayed for about half a second. Four random draws are illustrated in Figure 1. This approach follows Goldstein and Rothschild (2014), who also rely on this form of visualization to present the distribution. The goal is to treat these distributions as the objective truth and to identify which question format elicits beliefs that are closest to the true distribution.

Each respondent is then also randomly assigned to an elicitation question format in a simple *between-subjects design* (one question format per respondent). There is balance across question types with respect to a number of socio-economic variables (see Section A3 in the online appendix). We also employ two questions that serve as attention checks, and each question is correctly answered by about 75% of the respondents. Here, we show results for all respondents that answered both questions correctly, which is about 60% of the original sample. The same tables based on all respondents are shown in the online appendix (see Section A5 in the online appendix). There is no substantive difference between the two.

3.2 The Elicitation Step: Comparing Five Question Formats

The literature proposes different question formats to elicit univariate distributions (e.g., O'Hagan *et al.* 2006, Chapter 5.2). Here, we compare five common question formats that elicit different elements of a distribution and pose varying levels of cognitive demand. Two main selection criteria guided our choice of formats: (a) general question type and (b) ease of implementation in the context of online surveys. Based on a review of the relevant literature, we found that different question formats refer to different aspects of the belief distribution. Some present fixed intervals and ask for probabilities, others directly elicit quantile values or rely on a mix of both (see the distinction of P-methods and V-methods mentioned above). While most elicitation methods are purely verbal, others make use of visualization. Thus, our goal was to include one method of each general type.

Equally important is the second goal: to evaluate only such methods that are easily implemented in online surveys because they follow a simple question format. In addition, we also take advantage of the fact that online surveys provide us with the ability to use simple visual tools. But we will not consider elicitation protocols that demand close researcher–respondent interaction (e.g., Morris, Oakley, and Crowe 2014) or rely on incentivized elicitation methods that are often used in economic laboratory experiments (for an overview, see, e.g., Schlag, Tremewan, and Van der Weele 2015).

Here, we only briefly discuss each format. We present precise question wording in the online appendix (Section A1).

Interval Question (Wide and Narrow). These questions ask about the *probabilities* of fixed intervals (with the two versions varying the width of the interval values). More specifically, respondents are first asked to indicate the most likely value and then to provide us with the probability

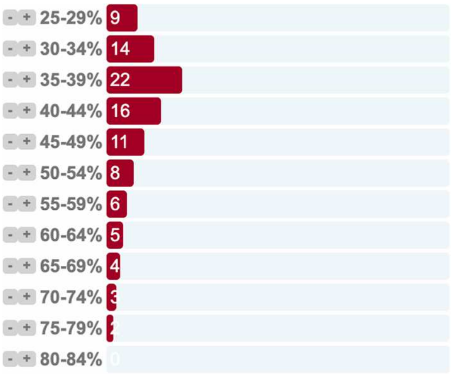

Figure 2. Screenshot of a Balls and Bins question. Illustration after hypothetical respondent has allocated all 100 balls.

Table 1. Question formats.

Question format	Most difficult concept R's need to know?	How many questions?	Adequacy check?
Interval (Wide and Narrow)	Quantiles	3	×
Quantile	Median	4	✓
Manski	Percentages	5	×
Bins & Balls	Percentages	1/100	✓

that a vote outcome will be lower than 40% (45% in the narrow format) and the probability that it will be higher than 60% (55% in the narrow format).

Quantile Question. The second question format asks respondents to provide three quantile *values*: the median, the first quartile, and the third quartile. This question format also provides an adequacy check. It ends by showing people their three responses (P_{25}, P_{50}, P_{75}) and asking them whether they think that a random draw is equally likely to fall into any of these intervals: $0 - P_{25}$, $P_{25} - P_{50}$, $P_{50} - P_{75}$, $P_{75} - 1$. Respondents can then correct themselves if they wish to do so. Thus, the fourth elicitation step is possible and respondents can assess the adequacy of their responses.

Manski Question. The third hybrid question format relies on work by Manski (2009) and asks for both, values *and* probabilities. Specifically, it first asks for three values along the distribution (the most likely value as well as the expected lower and upper bounds) and then asks respondents to provide probabilities for their elicited lower and upper bounds.

Bins and Balls. The last question is the latest addition to elicitation methods and takes advantage of the fact that a large number of surveys are being carried out online and, hence, allow for completely new question formats. Bins and Balls follows a proposal by Goldstein and Rothschild (2014) and is a visual tool for specifying a distribution where respondents have to place 100 balls into bins of a specific range (see Figure 2; see also Delavande and Rohwedder 2008). Balls are placed in a bin by the respondent's clicking on the + and – symbols. Since respondents are able to directly see the implied distribution, this is akin to an implicit adequacy check.

All five question formats differ in their complexity for respondents but also in how easily they can be implemented. Some of these questions lend themselves to adding an adequacy check at the end, others do not. Table 1 allows us to compare the different formats. The number of questions is ill-defined for the Bins and Balls format as it is one question but requires respondents to provide 100 inputs to distribute the virtual balls.

The Interval question and the Quantile question are particularly demanding as they require an understanding of quantiles. The Manski method is similar but can be expected to be less demanding since it translates the task into easier terms (means, maximums, and minimums) well. Finally, the Bins and Balls Question requires the least of respondents but its implementation is the most demanding for researchers. The question formats further differ on whether they allow for an adequacy check. In the Quantile question, for example, respondents can incorrectly place the upper quartile below the median. This can signal a wrong understanding of the question. In the next section, we investigate the accuracy of the elicited beliefs and discuss the experimental results.

3.3 The Fitting Step

To estimate a respondent's belief, we assume a flexible parametric distribution for his or her beliefs and estimate the parameters of the distribution such that it closely mimics the observed indicators for the different question formats. Because the sampling space of our experiment is bound between 0 and 1, we employ a beta distribution as our parametric assumption. The beta distribution has two shape parameters: α and β. We provide the derivation of the Interval question format as an example here. We present the derived likelihood functions for the other formats in the online appendix (see Section A2).

We observe three values for the Interval question. Respondents report the mean value of their beliefs and the probabilities of observing a value below and above a certain threshold. We denote the mean with y_i and the two ($k \in (1,2)$) probabilities with p_{i1} and p_{i2}. The interval values depend on the question format and are denoted with $c = [c_1, c_2]$, where in the wide version $c = [40\%, 60\%]$ and in the narrow version $c = [45\%, 55\%]$. We assume that the values are measured with normal measurement error:[3]

$$y_i \sim \mathcal{N}(\mu_y, \sigma_y^2) \tag{1}$$

$$p_{i1} \sim \mathcal{N}(\mu_{p_1}, \sigma_p^2) \tag{2}$$

$$p_{i2} \sim \mathcal{N}(\mu_{p_2}, \sigma_p^2). \tag{3}$$

The expectations μ_y are calculated from the assumed parametric belief distribution. Here, we use the same distribution as in the data-generating process—a beta distribution. The beta distribution is relatively flexible and well-suited for our example with vote shares constrained on the unit interval. The expectation for the mean from the beta is given by the two shape parameters α and β:

$$\mu_y = \frac{\alpha}{\alpha + \beta}. \tag{4}$$

The expected probabilities are given by the cumulative density function of the beta distribution, which we denote with $Q(\cdot, \alpha, \beta)$.

$$\mu_{p_1} = Q(c_1, \alpha, \beta) \tag{5}$$

$$\mu_{p_2} = 1 - Q(c_2, \alpha, \beta). \tag{6}$$

With this model, we can define the likelihood for the observed data. As we assume that all responses are identically and independently normal distributed, the likelihood is the product of

[3] Including a measurement model extends existing approaches that only minimize the squared error between observed and theoretical expected value (e.g., Morris *et al.* 2014). This can open the modelling framework up for a set of extensions, for example, correlated and heteroscedastic errors, hierarchical structures, and Bayesian estimation.

https://doi.org/10.1017/pan.2020.42 Published online by Cambridge University Press

three normal distributed measurements y_i, p_{i1}, and p_{i2} for each of the respondents.[4] To obtain maximum likelihood estimates of the parameters $\alpha, \beta, \sigma_p, \sigma_y$, the log-likelihood function is maximized using R's `optim` function. The estimates yield an estimate of the average beliefs for a specific condition. In the experiments, we can then identify the question format that will yield average belief estimates closest to the true values.

3.4 The Adequacy Check Step

Assessing the adequacy of the elicited distribution by giving respondents the chance to review and correct their belief distributions is difficult, because the fitting is done "outside" of the survey software and only after the answers have been collected. But for some formats, it is still possible to provide the opportunity for correction using question filters based either on respondents' answers or on visual question formats. The Quantile question, for instance, presents respondents with the quartiles they provided and asks if election results are equally likely to fall within each of them.[5] The Bins and Balls format asks respondents to "draw" their distribution and thus provides immediate feedback.

4 Results

To evaluate the different elicitation methods, we now compare the elicited beliefs to the benchmark of true objective distributions. Each column in Figure 3 stands for a combination of conditions (small/large variance and symmetric/asymmetric distribution). While we look at both symmetric and asymmetric true distributions, the asymmetric scenarios are likely to be more relevant in practice. This is because the only symmetric beta distributions are those distributions where the two shape parameters are exactly equal to each other. The five rows contain the different elicitation methods. We focus on the *average* elicited belief across all respondents and present the same figure with each *individual* belief distribution in the online appendix (see Figure A3).[6]

We find that most question formats are unbiased when the true distribution is symmetric, that is, they are able to provide the correct first moment. With asymmetric distributions, there is some bias towards .5 but its extent varies across question formats. It is especially evident for the Bins and Balls format. Looking at the second moment, we find that the two Interval questions tend to provide beliefs that are too wide in both, the symmetric and asymmetric scenarios. Thus, after simply eyeballing the plots, it seems that overall the Manski question and the Quantile question come closest to the true distributions.

To evaluate the question formats more formally, we turn to the results in Table 2 where we illustrate for each combination of experimental factors: the implied parameters of the elicited priors, the Kullback–Leibler divergence, the number of observations, and the *p*-value of a likelihood-ratio test on whether the estimated parameters differ from the true values of the parameter. The smaller the KL divergence, the closer the elicited prior is to the true distribution. We also present a figure with the sum of the KL divergence over all four experimental conditions (see Figure 4). Here, we again only show the results when averaging across all respondents. The online appendix contains the results for individual respondents (see Figure 3).

If we take the sum of all four experimental settings, the Manski question scores the smallest value for the Kullback–Leibler divergence, $KL = 0.28$. It is followed by the Quantile question ($KL = 0.65$) and Bins and Balls ($KL = 0.91$). The two Interval questions perform worst ($KL = 1.31$ for the wide and $KL = 1.48$ for the narrow interval). As mentioned above, in practice asymmetric

[4] We provide a more detailed description of the likelihood function in Section A2 in the online appendix.

[5] In online appendix (Section A4), we compare the results of the Quantile question with and without the adequacy check and find that the adequacy check can considerably improve the result.

[6] A full replication package is available, see Leemann, Traunmueller, and Stoetzer (2020).

https://doi.org/10.1017/pan.2020.42 Published online by Cambridge University Press

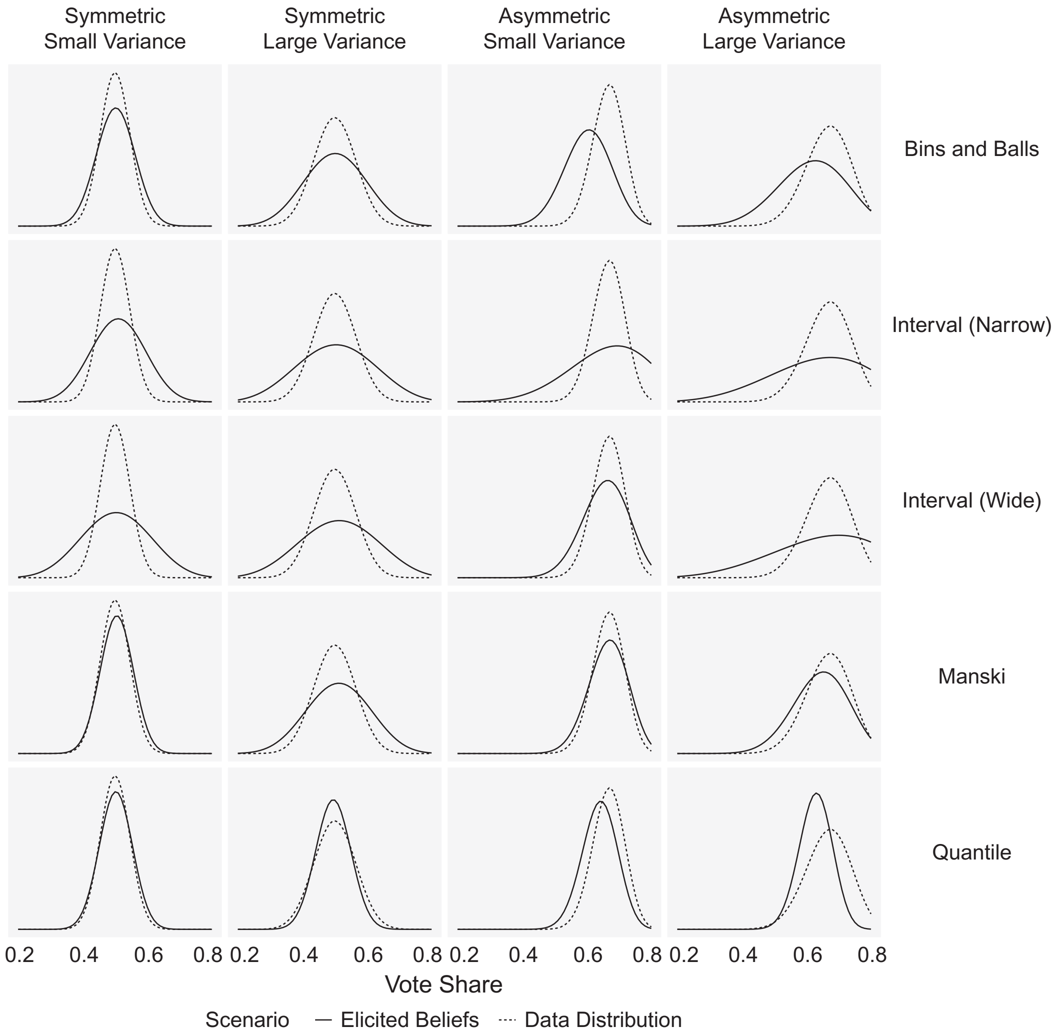

Figure 3. Comparison of question formats. The dotted line indicates the true distribution and the black solid line shows the average of the elicited distributions.

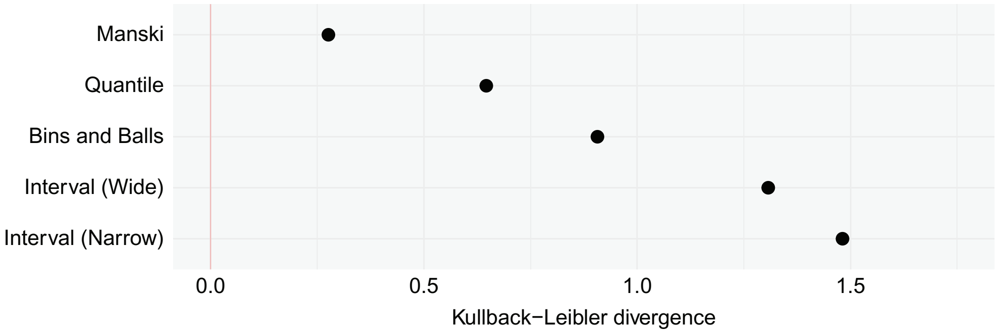

Figure 4. Summed Kullback–Leibler divergence. The Manski question performs best across the five experimental scenarios.

scenarios are more frequent and the Manski question format also beats all other alternatives for this case.

Based on these experiments we conclude that the Manski question format outperforms the other elicitation methods. In principle, it seems intuitive that eliciting more points along a

https://doi.org/10.1017/pan.2020.42 Published online by Cambridge University Press

Table 2. Experimental comparison of five elicitation methods across four scenarios. Implied parameters of elicited priors, Kullback–Leibler divergence, *p*-value of a likelihood-ratio test and number of observations shown.

method	alpha	beta	KL	lr	N
Quantile	42.35	43.10	0.04	0.44	65
Bins and Balls	13.59	13.51	0.13	0.00	60
Manski	13.01	12.41	0.15	0.00	62
Interval (Narrow)	8.54	8.41	0.28	0.00	61
Interval (Wide)	8.68	8.27	0.29	0.00	69

(a) Symmetric, large variance

method	alpha	beta	KL	lr	N
Quantile	48.41	47.90	0.01	0.69	119
Manski	48.75	47.72	0.02	0.39	112
Bins and Balls	35.67	35.38	0.06	0.00	107
Interval (Narrow)	18.02	17.35	0.27	0.00	126
Interval (Wide)	11.03	10.88	0.45	0.00	115

(c) Symmetric, small variance

method	alpha	beta	KL	lr	N
Manski	19.97	11.07	0.06	0.00	65
Bins and Balls	12.75	7.96	0.23	0.00	48
Interval (Narrow)	5.95	3.39	0.41	0.00	68
Interval (Wide)	5.39	2.88	0.46	0.00	65
Quantile	55.23	32.69	0.47	0.00	65

(b) Asymmetric, large variance

method	alpha	beta	KL	lr	N
Manski	38.63	19.36	0.04	0.18	112
Interval (Wide)	28.39	14.79	0.11	0.00	120
Quantile	49.08	27.64	0.13	0.00	118
Bins and Balls	27.21	17.98	0.50	0.00	121
Interval (Narrow)	9.38	4.69	0.52	0.00	133

(d) Asymmetric, small variance

distribution would also result in a better measurement of respondents' beliefs.[7] However, we would be mistaken to equate the number of questions with an elicitation's methods performance. The Quantile question, for instance, asks for three quantities and adds an adequacy check. Without this adequacy check (an option that we test, see Table A5 in the online appendix), it would ask just as many questions as the two Interval questions—yet the performance across these formats clearly differs. Without the adequacy check, the Quantile question outperforms the two Interval questions in the symmetric scenario with large variance (KL-distance of .07 *vs.* .28 and .29) but has more problems than the two Interval questions in the asymmetric scenario with large variance (KL-distance of .98 *vs.* .41 and .46). One possibility why the Quantile question fares better in the symmetric case than in the asymmetric case is that in the symmetric case respondents can rely on equal distance of the first and third quantile to the median as an informal adequacy check. This possibility does not exist when the belief is asymmetric.

In sum, the Manski format provides a fairly effective approach to prior elicitation that is straightforward to implement because it only requires five questions to measure respondents' beliefs. In addition, the Manski question takes marginally less time (median completion time was 170 s) than the Quantile question (200 s) and Bins and Balls (199 s), but more time than the two Interval questions (149 and 157 s, respectively). These differences are not statistically significant (more detail is provided in Section A8 in the online appendix).

[7] Note that this is different from the classical notion of reducing measurement error by have multiple measures of the *same* underlying quantity.

https://doi.org/10.1017/pan.2020.42 Published online by Cambridge University Press

A valid question is whether our results are sensitive to the composition of the sample used. For example, it is unclear whether respondents on MTurk pay more or less attention than "normal" respondents. While some argue that MTurkers are less attentive and try to complete tasks as quickly as possible, others argue that workers are more attentive because they are paid for these tasks. Several studies have looked into the properties of MTurk samples and found them to perform equally well to other online samples (Mullinix *et al.* 2015; Coppock 2018). We provide analyses that probe into the effects of respondents' attention in the online appendix (Section A5). Comparing attentive respondents (i.e., those that passed the attention checks[8]) to all respondents, we find that attentive respondents are slightly closer to the objective distributions than the complete sample. More importantly however, the relative performance of the five elicitation methods is not affected by respondents' level of attentiveness. We find similar results for the distinction between sophisticated and unsophisticated respondents (as proxied by political interest, see Section A6 in the online appendix).

On a final note, we only find limited evidence for any systematic biases in respondents' beliefs. In particular, we only observe over-confidence (i.e., respondents' tendency to be more certain than the objective data would warrant and, therefore, assign variances that are too narrow) in the case of the Quantile question. In the symmetric scenario with large variance, for instance, the true standard deviation is $\sigma = .064$, but the average elicited distribution yields a standard deviation of only $\sigma = .051$.[9] For all other formats, respondents actually express beliefs that are *less* certain than the objective benchmark would demand (i.e., the elicited distributions are too wide).

Before concluding this letter, we provide an illustrating application where we use these different techniques to elicit people's subjective beliefs about a future outcome.

5 Application: What Vote Share Will Trump Receive in November 2020?

In the applied setting of an actual elicitation process, researchers would of course not instill an objective distribution. Instead they would seek to elicit the beliefs that respondents already hold about a subject matter. To illustrate the relative performance of the five elicitation methods in a more realistic setting, this section provides an example where we ask respondents to indicate their beliefs about the upcoming presidential election. This closely mimics an actual elicitation exercise.

We report the results from an online survey carried out on MTurk with 500 participants. Each respondent was asked what their belief was of the popular vote share that Donald Trump will garner in November 2020. As with the main experiments presented above, we again only offer respondents one randomly assigned question format in a simple between-subjects design. For each question format, we estimate the underlying belief distribution of the full sample and then, because we expect clear partisan differences, separately for Democrats and Republicans.

There are two main results that can be gleaned from Figure 5 (we provide more detail on the estimated beliefs in Table A10 in the online appendix). First, which question format is used for eliciting respondents' beliefs clearly matters. The average expected 2020 popular vote share for Donald Trump differs from one format to the other. In addition, the elicitation methods differ vastly in the belief variances they produce (see Section A7 in the online appendix for more details). In the experiments, the Manski question format performed best in retrieving objective belief distributions. When eliciting pre-existing subjective beliefs we do not know the true distribution and hence cannot assess each method's precision. What we can assess, is how clearly the signal

[8] We use two screening questions to detect inattentive respondents (Berinsky, Margolis, and Sances 2014). One asks respondents to recall the founding organization mentioned in the introductory text and the second question asks them to recall the colors of the plots in the GIF (see Figure 1). The latter is especially relevant as it is directly tied to the communication of the true distribution.

[9] The variance of the beta distribution is $\sigma^2 = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$. The standard deviation σ is more intuitive because it is on the original scale of the variable, in our case vote shares.

https://doi.org/10.1017/pan.2020.42 Published online by Cambridge University Press

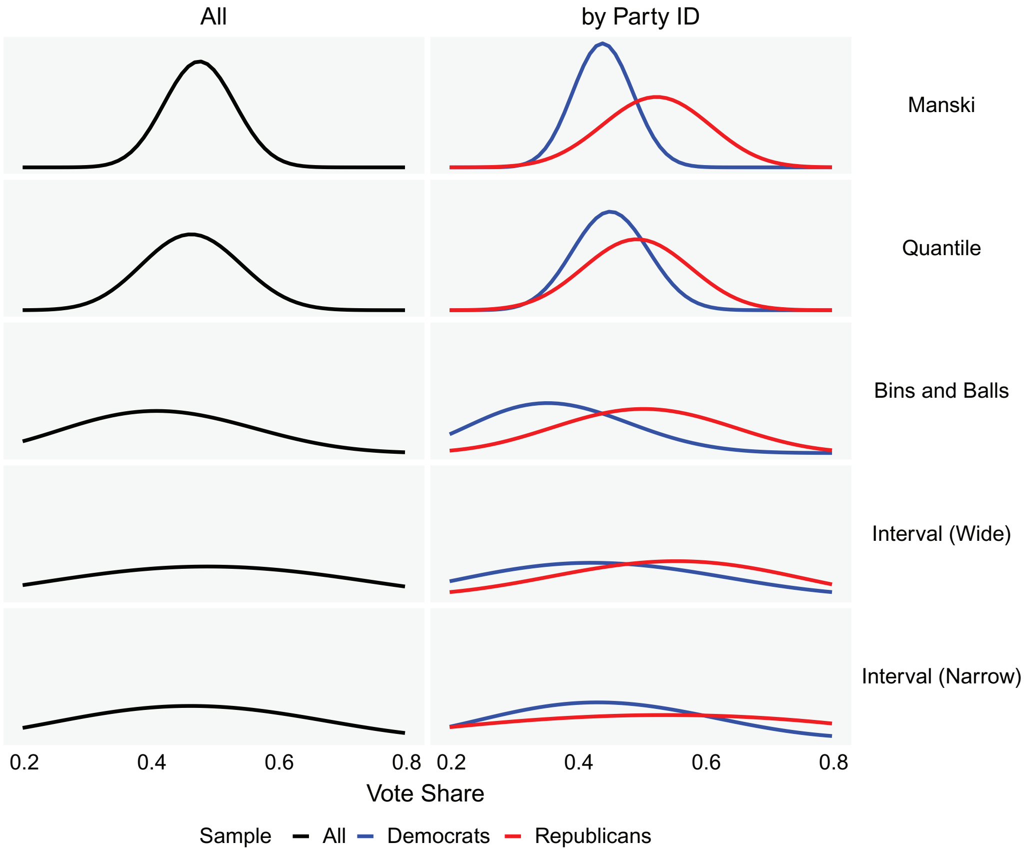

Figure 5. Elicited beliefs over Trump's popular vote share in November.

is measured, that is, which formats provide plausible results with low variance. In the 2020 Trump election example, we find that the ordering in performance is similar to the ordering found in the experiments. Given what we know about vote shares in U.S. presidential elections, the variances provided by the two Interval questions and Bins and Balls are much too wide to be of any substantive use. The average beliefs elicited by the Interval question have a standard deviation of $\sigma = 17.9\%$ (wide) and $\sigma = 16.3\%$ (narrow) while the beliefs produced by Bins and Balls have standard deviation of $\sigma = 13.1\%$. In line with the experimental results, both the Manski and Quantile question formats provide reasonable results $\sigma = 5.4$ and $\sigma = 7.5\%$, respectively), but the Manski question format again performs best.

The second result is that there are clear partisan differences in the beliefs and expectations about the upcoming 2020 presidential election. We think this is in line with prior literature (Lebo and Cassino 2007; Kuru, Pasek, and Traugott 2017; Madson and Hillygus 2019). Relying on the preferred Manski question format, we find that Republicans have a more optimistic belief about Donald Trump's expected popular vote share (52.3%) than Democrats (44.1%). At the same time, Republicans are less certain about the election outcome than Democrats. The standard deviation of Republicans' belief is $\sigma = 8.1\%$ compared to only $\sigma = 4.7\%$ for Democrats.

6 Conclusion

This research note has empirically evaluated five different question formats for prior elicitation in the context of online surveys. For each format, we derived the estimators to recover the shape parameters describing respondents beliefs and ran experiments to compare the relative performance of these elicitation methods. We find that a set of questions originally proposed by Manski (2009) performs very well.

https://doi.org/10.1017/pan.2020.42 Published online by Cambridge University Press

This is good news for applied researchers who seek to study citizens beliefs as distributions. While all five types of elicitation methods are fairly easy to implement, the Manski question is especially straightforward as it only consist of asking people for five numbers (most likely value, lower and upper bound and the probabilities associated with the two bounds). Since it is purely verbal, there is no need for programming—unlike other elicitation methods, such as the Bins and Balls method recently proposed by (Goldstein and Rothschild 2014) which requires programming in Java script. In addition, the Manski format seems to perform in a similar fashion across different subgroups defined by political sophistication, which can be a relevant consideration. Finally, there is one caveat that needs mention: we assumed throughout that citizens' beliefs follow a unimodal distribution. While this is a reasonable assumption in many circumstances, it is possible that one would want to elicit multimodal beliefs. In such situation, the Bins and Balls format would allow researchers to do so, but the estimation methods must be adapted accordingly.

Data Availability

Supplementary materials for this article are available on the Cambridge Core website. For Dataverse replication materials, see Leemann *et al*. (2020).

Supplementary Material

For supplementary material accompanying this paper, please visit
https://dx.doi.org/10.1017/pan.2020.42.

Acknowledgments

Many thanks to Daniel Bischof, Tim Hicks, Patrick Kraft, Andreas Murr, Simon Munzert, Ana Petrova, and David Rothschild for their valuable comments and help. We would also like to thank the editor for his guidance and help and the two anonymous referees for their comments. We thank Lucien Baumgartner for his impeccable research assistance. This research was partly funded by the Department of Political Science at the University of Zürich.

References

Alvarez, R. M., and J. Brehm. 1997. "Are Americans Ambivalent Towards Racial Policies?" *American Journal of Political Science* 41:345–374.

Bartels, L. M. 2002. "Beyond the Running Tally: Partisan Bias in Political Perceptions." *Political Behavior* 24(2):117–150.

Berinsky, A. J. 2017. "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science* 47(2):241–262.

Berinsky, A. J, G. A. Huber, and G. S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon. com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Berinsky, A. J, M. F. Margolis, and M. W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-administered Surveys." *American Journal of Political Science* 58(3):739–753.

Bullock, J. G. 2009. "Partisan Bias and the Bayesian Ideal in the Study of Public Opinion." *The Journal of Politics* 71(3):1109–1124.

Coppock, A. 2018. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7:1–16.

Cosmides, L., and J. Tooby. 1996. "Are Humans Good Intuitive Statisticians after all? Rethinking Some Conclusions from the Literature on Judgment under Uncertainty." *Cognition* 58(1):1–73.

Delavande, A., and S. Rohwedder. 2008. "Eliciting Subjective Probabilities in Internet Surveys." *Public Opinions Quarterly* 72(5):866–891.

Garthwaite, P. H., J. B. Kadane, and A. O'Hagan. 2005. "Statistical Methods for Eliciting Probability Distributions." *Journal of the American Statistical Association* 100(470):680–701.

Gerber, A., and D. Green. 1999. "Misperceptions About Perceptual Bias." *Annual Review of Political Science* 2(1):189–210.

Gill, J., and J. R. Freeman. 2013. "Dynamic Elicited Priors for Updating Covert Networks." *Network Science* 1(1):68–94.

Gill, J., and L. D. Walker. 2005. "Elicited Priors for Bayesian Model Specifications in Political Science Research." *The Journal of Politics* 67(3):841–872.

Goldstein, D. G., and D. Rothschild. 2014. “Lay Understanding of Probability Distributions.” *Judgment & Decision Making* 9(1):1–14.

Kuru, O., J. Pasek, and M. W. Traugott. 2017. “Motivated Reasoning in the Perceived Credibility of Public Opinion Polls.” *Public Opinion Quarterly* 81(2):422–446.

Kynn, M. 2008. “The ‘Heuristics and Biases’ Bias in Expert Elicitation.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(1):239–264.

Lebo, M. J., and D. Cassino. 2007. “The Aggregated Consequences of Motivated Reasoning and the Dynamics of Partisan Presidential Approval.” *Political Psychology* 28(6):719–746.

Leemann, L., R. Traunmueller, and L. F. Stoetzer. 2020. “Replication Data for: Eliciting Beliefs as Distributions in Online Surveys.” https://doi.org/10.7910/DVN/GEC2LS, Harvard Dataverse, V1, UNF:6:AJOSTeuI2rt9XmpP7/2rlg== [fileUNF].

Leiter, D., A. Murr, E. R. Ramirez, and M. Stegmaier. 2018. “Social Networks and Citizen Election Forecasting: The More Friends the Better.” *International Journal of Forecasting* 34(2):235–248.

Madson, G. J., and D. S. Hillygus. 2019. “All the Best Polls Agree with Me: Bias in Evaluations of Political Polling.” *Political Behavior* 5:1–18.

Manski, C. F. 2009. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

Mason, W, and S. Suri. 2012. “Conducting Behavioral Research on Amazon’s Mechanical Turk.” *Behavior Research Methods* 44(1):1–23.

Morris, D. E., J. E. Oakley, and J. A. Crowe. 2014. “A Web-Based Tool for Eliciting Probability Distributions from Experts.” *Environmental Modelling & Software* 52:1–4.

Mullinix, K. J., T. J. Leeper, J. N. Druckman, and J. Freese. 2015. “The Generalizability of Survey Experiments.” *Journal of Experimental Political Science* 2(2):109–138.

Murr, A. 2011. “‘Wisdom of Crowds’? A Decentralised Election Forecasting Model that Uses Citizens’ Local Expectations.” *Electoral Studies* 30(4):771–783.

O’Hagan, A. *et al.* 2006. *Uncertain Judgements: Eliciting Experts’ Probabilities*. Boca Raton, FL: John Wiley & Sons.

Peterson, C., and A. Miller. 1964. “Mode, Median, and Mean as Optimal Strategies.” *Journal of Experimental Psychology* 68(4):363.

Savage, L. J. 1971. “Elicitation of Personal Probabilities and Expectations.” *Journal of the American Statistical Association* 66(336):783–801.

Schlag, K. H., J. Tremewan, and J. J. Van der Weele. 2015. “A Penny for Your Thoughts: A Survey of Methods for Eliciting Beliefs.” *Experimental Economics* 18(3):457–490.

Spetzler, C. S., and C.-A. S. Stael von Holstein. 1975. “Exceptional Paper—Probability Encoding in Decision Analysis.” *Management Science* 22(3):340–358.

Thomas, K. A., and S. Clifford. 2017. “Validity and Mechanical Turk: An Assessment of Exclusion Methods and Interactive Experiments.” *Computers in Human Behavior* 77:184–197.

Tversky, A., and D. Kahneman. 1971. “Belief in the Law of Small Numbers.” *Psychological Bulletin* 76(2):105.

Tversky, A., and D. Kahneman. 1973. “Availability: A Heuristic for Judging Frequency and Probability.” *Cognitive Psychology* 5(2):207–232.

Tversky, A., and D. l. Kahneman. 1974. “Judgment under Uncertainty: Heuristics and Biases.” *Science* 185(4157):1124–1131.

Wallsten, T. S., and D. V. Budescu. 1983. “State of the Art—Encoding Subjective Probabilities: A Psychological and Psychometric Review.” *Management Science* 29(2):151–173.

Zaller, J., and S. Feldman. 1992. “A Simple Theory of the Survey Response: Answering Questions and Revealing Preferences.” *American Journal of Political Science* 36:579–616.

https://doi.org/10.1017/pan.2020.42 Published online by Cambridge University Press