CESifo WORKING PAPERS

8823
2021

Original Version: January 2021
This Version: January 2022

Online Belief Elicitation Methods

Valeria Burdea, Jonathan Woon

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 8823*

Online Belief Elicitation Methods

Abstract

How well do incentivized belief elicitation procedures work in online settings? We evaluate the quality of beliefs elicited from online respondents, comparing several characteristics of two widely used complex elicitation mechanisms (the Binarized Scoring Rule - BSR - and a stochastic variation of the Becker-deGroot-Marschak mechanism - BDM) against a at fee baseline for a variety of beliefs (induced probabilities, first-order factual knowledge, second-order knowledge of others). We find that the at-fee method requires the least amount of time, the BDM is the most difficult to understand, and that there are no differences in the average accuracy of induced beliefs across conditions. However, the methods are significantly different in terms of the frequency of first-order and second-order beliefs reported at exactly 50%: the at-fee method leads to the most mass on this belief, followed by BDM and BSR. Regarding induced beliefs, we also find that less-educated participants' accuracy is higher in the complex incentives treatments, and that attention, numeracy, and education are positively associated with the quality of these beliefs across methods. Our results suggest that the quality of beliefs elicited in online environments may depend less on the formal incentive compatibility properties of the elicitation procedure (whether the procedure prevents "dishonest" reporting) than on the difficulty of comprehending the task and how well incentives induce cognitive effort (thereby inducing subjects to quantify or construct their beliefs).

JEL-Codes: C810, C890, D830, D910.

Keywords: belief elicitation, incentives, online experiment.

Valeria Burdea
Department of Economics
University of Munich (LMU) / Germany
valeria.burdea@econ.lmu.de

Jonathan Woon
Department of Political Science
University of Pittsburgh / PA / USA
woon@pitt.edu

January 22, 2022
Financial support from the University of Pittsburgh is gratefully acknowledged. The research has been approved by the University of Pittsburgh IRB (STUDY19100046). Data and analysis materials (R code) will be made available online upon publication.

1 Introduction

Uncertainty pervades important economic, management, political, and life decisions such as whether to invest in untested new technologies, which political candidate to vote for, or whether to vaccinate against infectious diseases. Measuring probabilistic expectations can help us better understand and manage these important social interactions and decision processes. For example, consumer expectations about various economic outcomes can be used by central banks in their forecasting models (Armantier et al., 2017), or beliefs about the future value of education can reveal important pathways to reducing enrollment gaps in further education (Belfield et al., 2020). Similarly, measuring beliefs about performance differences across genders can help us better address gender discrimination in the workplace (Coffman, Exley and Niederle, 2021), while those about others' cooperativeness can shed light on the relations between economic heterogeneity, beliefs and behavior (Martinangeli, 2021). Equally important, the elicitation of subjective beliefs can help researchers test and distinguish between different models of judgment and behavior such for how people form biased beliefs (Coutts, 2019), for why they cooperate in prisoner's dilemma games (Peeters and Vorsatz, 2021) or for people's attitudes towards ambiguity (Baillon and Bleichrodt, 2015).

But eliciting subjective beliefs is far from trivial. Indeed, considerable effort has been devoted to developing incentive-compatible mechanisms for measuring these, both theoretically (e.g., Allen, 1987; Brier et al., 1950; DuCharme and Donnell, 1973; Grether, 1981; Holt, 2007; Hossain and Okui, 2013; Karni, 2009; Savage, 1971) as well as empirically (e.g., Burfurd and Wilkening, 2018; Erkal, Gangadharan and Koh, 2020; Schlag, Tremewan and Van der Weele, 2015; Schotter and Trevino, 2014; Trautmann and van de Kuilen, 2015). Laboratory studies suggest that incentivized methods using proper scoring rules perform better in predicting participants' own behavior than non-incentivized introspection. However, these studies suggest the methods do not differ in terms of the accuracy of the elicited beliefs, measured by comparing the belief with the external, objective truth (Schlag, Treme-

wan and Van der Weele, 2015; Schotter and Trevino, 2014; Trautmann and van de Kuilen, 2015). Moreover, most belief elicitation studies are conducted in-person in laboratory environments, where great care is taken to ensure participants are sufficiently motivated and lengthy procedures are used to promote understanding of task instructions and incentive compatibility of the mechanisms. However, samples of participants in such studies tend to be homogeneous and relatively educated (primarily undergraduate students), and it can be time-consuming and costly to achieve sufficiently-powered sample sizes.

Online studies provide a promising alternative because they can be conducted more quickly, less expensively, with greater numbers and diversity of participants, and they are increasingly being used in the behavioral and social sciences (Arechar, Gächter and Molleman, 2018; Berinsky, Huber and Lenz, 2012; Hergueux and Jacquemet, 2015; Horton, Rand and Zeckhauser, 2011; Mason and Suri, 2012; Paolacci, Chandler and Ipeirotis, 2010). Yet for any number of reasons, moving from highly controlled in-person laboratories to online settings may come at the cost of participants not being able to comprehend a complicated incentive mechanism. Subjects in online studies may be more time-constrained or aim to maximize their hourly rate of compensation and therefore try to complete a study as quickly as possible, a common problem in online survey research. They may be less attentive or distracted for other reasons, perhaps because they are surfing the web, watching television, or simultaneously engaged in other forms labor (Chandler, Mueller and Paolacci, 2014; Clifford and Jerit, 2014; Crump, McDonnell and Gureckis, 2013; Oppenheimer, Meyvis and Davidenko, 2009). Greater diversity of participants may also mean greater heterogeneity in cognitive ability and sophistication than a typical sample of undergraduates. So, there may be more participants online who struggle to understand the incentives than in a laboratory, where more complex belief elicitation mechanisms have already been shown to exhibit more sensitivity to differences in participants' probabilistic reasoning (Burfurd and Wilkening, 2021). How concerned should we be about the quality of belief data elicited in online studies? Which methods work better? How much is gained by implementing the elaborate

incentive compatibility mechanisms heavily relied on in the lab, versus simply asking for beliefs? Do incentives matter? How accurate are beliefs, and how much do cognitive factors such as attention, numeracy, or education matter given their greater heterogeneity in online samples?

To answer these questions, we study beliefs elicited from participants via an online interface and recruited from an online labor market. We compare adaptations of two popular elicitation mechanisms, the Binarized Scoring Rule (Hossain and Okui, 2013) and the stochastic Becker-deGroot-Marshak mechanism (Allen, 1987; DuCharme and Donnell, 1973; Grether, 1981; Holt, 2007; Karni, 2009), against a non-incentive-compatible flat rate payment. We elicit a variety of beliefs from each participant, including induced probabilities (which we can compare to objective benchmarks for accuracy), confidence in knowledge on a trivia quiz (subjective first-order beliefs), and beliefs about the accuracy of others' knowledge (subjective second-order beliefs). Hewing closely to standard laboratory practices, we provide complete descriptions of the incentive mechanisms and test participants about their comprehension. In addition, we ask participants for their subjective perceptions of the difficulty of the task. Importantly, we also collect a set of demographic and cognitive measures. Given the lack of previous evidence on the performance of different belief elicitation mechanisms in the online setting, our study constitutes an exploratory horse-race between the three methods we focused on.

We make several empirical contributions. First, our results suggest important procedural and perceptual differences across the different methods, such that complex methods require longer implementation time and are associated with higher comprehension costs (perceived difficulty and effort) than the flat rate payment. Second, our study highlights the importance of cognitive factors in affecting the quality of elicited beliefs. We find that both incentive-compatible methods improve the accuracy of elicited beliefs about induced probabilities, but only for less educated participants. Other cognitive measures such as attention and numeracy are positively related to the accuracy of these beliefs, uniformly across

elicitation methods. Third, when eliciting first and second order beliefs, our results show that the distribution of beliefs across methods differs, in that incentive-compatible methods lead people to report fewer beliefs at 50%. The Binarized Scoring Rule is the most successful in doing so. Moreover, these departures are more likely to occur in the direction of the underlying probability when beliefs are elicited using the Binarized Scoring Rule.

Our findings have several practical implications for belief elicitation research. First, if the quantities of interest are "easy" to compute or more familiar to a subject, if time constraints are a concern, or if the online platform does not allow for variable payments (such as with large-scale established social surveys), then the gain in accuracy may not be worth the effort of explaining and ensuring subjects comprehend a complicated elicitation mechanism. Second, for "hard" or less familiar quantities (such as second-order beliefs), relying on an incentive-compatible mechanism (and the Binarized Scoring Rule in particular) may be worthwhile because, by encouraging cognitive effort, subjects may be more likely to think about and formulate a belief other than 50%. Even if subjects are indeed completely uncertain, such beliefs are more meaningful if arrived at upon reflection rather than stated out of laziness. Finally, because a substantial amount of error in beliefs seems to be more likely related to innumeracy, and to a lesser extent inattention, studies of beliefs—online and offline—would benefit from regularly including such measures as covariates that can be used to condition the analysis.

The rest of the paper is structured as follows. Section 2 presents the experimental design and procedures, including a brief description of the mechanism behind each elicitation method. Section 3 details the results of our experiment, beginning with the analysis of the cognitive measures and moving on to the analysis of the three types of beliefs we elicited. Section 4 concludes with a discussion of the lessons we can draw from our study and provides recommendations for researchers interested in eliciting beliefs in online studies.

2 Experimental design and procedures

We recruited participants from the Amazon Mechanical Turk (MTurk) online labor market in December 2019, with 470 participants completing the study via the Qualtrics online platform. We required participants' location to be in the United States and to have a 95% HIT Approval Rate. The sample size was determined based on a power analysis designed to detect a medium sized difference in accuracy across conditions (Cohen's d = 0.5) with 80% power and 5% Type I error rate, suggesting a minimum sample size of 144 observations per condition, which we rounded up to 150 per condition.

Upon clicking on the study link in MTurk, subjects were randomly assigned to one of three incentive conditions (FLAT, BDM, or BSR, described in detail below) and then proceeded to give informed consent. Some amount of attrition is typical in online studies. Of the 522 total subjects who started our study, 90% completed it. We do not include any partial observations in our analysis of the results. The drop-out rates did not differ across conditions (12% in FLAT, 11% in BDM and 11% in BSR), and the final sample size was evenly distributed across conditions (n=155 in FLAT, n=157 in BDM, n=158 in BSR).

The order of procedures was the same in all conditions and is summarized in Table 1 along with the study measures and tasks. (See Online Appendix for the complete text of all survey items, instructions, and all experimental conditions.) For completing the study, participants received a fixed payment of $1.50 plus a variable bonus from the belief elicitation task that amounted to $1.17 on average. The average variable bonus per elicited belief across treatments was the following - FLAT: M= $0.20, SD=0.00; BDM: M= $0.24, SD=0.20; BSR: M= $0.27, SD=0.19. Due to the random nature of the incentive schemes, there was no reason to expect that the average bonus would be exactly the same across methods. Nevertheless, the payoffs were chosen to minimize actual and perceived potential differences. The average duration for completing the study was 14 minutes. The average hourly wage for participants in our study was therefore $11.44/hour, which is above the U.S. federal minimum wage of $7.25/hour at the time of our study. We thus follow the recommendation

for researchers to pay at least the minimum wage (Silberman et al., 2018; Williamson, 2016) to address potential ethical concerns about the level of compensation provided to participants in online labor markets (e.g., Fort, Adda and Cohen, 2011). Indeed, we chose a relatively high completion fee in order to ensure this is the case, while keeping the average belief elicitation incentives comparable to the incentives typically used for such objects in online studies (e.g. Coffman, Collis and Kulkarni, 2019; Hill, 2017; Roth and Wohlfart, 2020).

I. Consent	
II. Pre-Treatment Measures	
Demographics	
Political identification	
Attention check	
Factual knowledge	
Numeracy	
Belief scale comprehension	
III. Explanation of Incentive Scheme	
Incentives comprehension	
Subjective comprehension	
IV. Incentivized Belief Tasks	
Part 1. Physical probability events (induced probability beliefs)	
Part 2. Truth of factual statements (1^{st} order beliefs)	
Part 3. Accuracy of others' knowledge (2^{nd} order beliefs)	
Part 4. Accuracy of Democratic subgroup's knowledge (2^{nd} order beliefs)	} Order
Part 5. Accuracy of Republican subgroup's knowledge (2^{nd} order beliefs)	} randomized

Table 1: Order of procedures, measures, and tasks

We first elicited a variety of participant characteristics and pre-treatment measures before introducing the incentive schemes and belief tasks. Basic demographic characteristics included gender, age, ethnicity, and education. We then asked about partisan identification (using a branching format, generating a 7-point scale) and ideology (using a single 7-point item), followed by an attention check (Berinsky, Margolis and Sances, 2014). Next, participants took a brief (unincentivized) true-false *knowledge quiz* involving a set of political and historical facts (see Table 2) shown to subjects in random order. The items in the knowledge quiz provide the basis for the first- and second-order beliefs we elicit.

Before introducing the belief task, we asked participants six hypothetical probability

Item no.	Statement	True/False
1	More than half of unauthorized immigrants residing in the United States in 2016 had been living in the country for 10 years or more.	True
2	From 2009, when President Obama took office, to 2012, median household income adjusted for inflation in the United States fell by more than 4 percent.	True
3	More people in the United States work in the coal industry than in the solar industry.	False
4	West Virginia was part of the Confederacy during the American Civil War.	False

Table 2: Knowledge Quiz

questions. For example: "Imagine that we roll a fair, six-sided die with the numbers 1 through 6 on its sides. What is the likelihood that the die will come up even?" Asking about the likelihoods associated with physically randomized events (die rolls, coin tosses, ball draws, etc.) without incentives provides us with a measure of subjects' *numeracy*, which is "the ability to process basic probability and numerical concepts" (Peters et al., 2006, p. 407). We measured numeracy to control for the possibility that the quality of elicited beliefs might depend on participants' basic understanding of probability rather than on the properties of the incentive schemes themselves, or on the interaction between incentives and numeracy, and because we were concerned that participants recruited from online platforms might have lower levels of understanding of mathematical concepts than typical convenience samples of undergraduate students. The order in which these numeracy items were presented to participants was not randomized, and the full list of questions can be found in the Online Appendix (Table A1).

Following the numeracy questions, we then provided a general introduction to the belief elicitation task. We explained there were five different parts and that each part provided an opportunity to earn a bonus, and we then explained the process of reporting their beliefs and how different numbers should be interpreted in qualitative terms. We elicited beliefs about whether a statement (about an event or some fact) is true or false, and we

described their belief as a number B from 0 to 100 with meanings of the numbers described to participants exactly as in Table 3.

Your belief (B)	This means:
100	You think the statement is certainly TRUE, beyond any doubt
51-99	You think the statement is likely to be TRUE (higher numbers indicate greater certainty it is TRUE)
50	You are totally uncertain
1-49	You think the statement is likely to be FALSE (lower numbers indicate greater certainty it is FALSE)
0	You think the statement is certainly FALSE, beyond any doubt

Table 3: Explanations of the meaning of beliefs presented to participants

After reading the instructions, we asked participants to answer five comprehension questions about the belief scale. In this way, we ensured a common understanding of the meanings associated with subjects' reported beliefs. Subjects could make at most two errors in each of these questions, after which they were shown the correct answer and proceeded with the study. We imposed these limits for controlling the amount of time participants would spend on this section and to avoid frustration. The general task introduction, explanation of beliefs, and belief scale comprehension questions were identical across treatments and were completed prior to providing any treatment-specific information; hence, the belief scale comprehension questions are pre-treatment covariates.

Once participants completed the belief comprehension questions, we presented instructions about the incentive mechanism specific to their randomly assigned condition (FLAT, BDM, or BSR). We describe these in greater detail in the next section. At this point, we also emphasised that this is a "no deception" study to ensure that participants trusted that their bonuses were determined as described in the instructions. Following the explanation, we asked four comprehension questions about the incentive scheme. As with the belief scale comprehension questions, we allowed a maximum of two errors. After one er-

ror, we provided a reminder about the instructions before the second attempt, and after two errors subjects were presented with the correct answer and proceeded to the next question. Although we kept the reference numbers constant across conditions in these comprehension questions wherever possible, the content of the questions necessarily differed across conditions due to the differences in the incentive mechanisms themselves. Therefore, we cannot directly compare error rates in these comprehension questions across methods.

After the comprehension check, we then asked subjects two questions about how they *perceived* their understanding of the instructions. Specifically, we asked about the *difficulty* of understanding the incentive instructions and how much *effort* they felt this involved; both questions used a 5-point Likert scale, where higher numbers indicate higher difficulty or more effort, respectively.

The five belief tasks (described as "parts") were identical across conditions, save for minor differences reminding participants about their incentives. Participants reported their beliefs using a slider interface (depicted in Figure 1), which was identical across all three conditions.

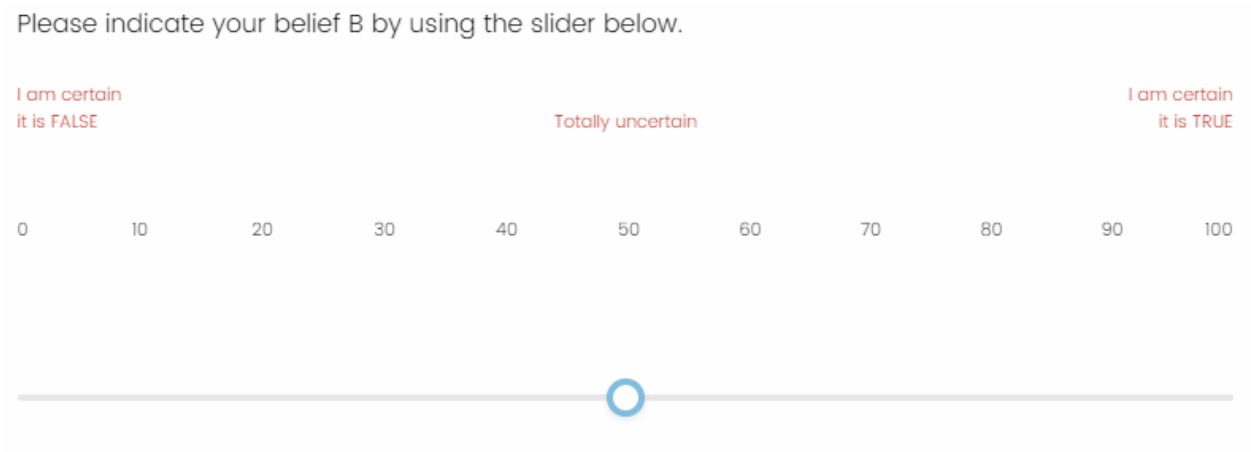


Figure 1: Interface for reporting beliefs

We implemented the belief task using a slider due to its intuitive presentation, ease of implementation, short duration, and comparability across our three elicitation methods. Another possibility would have been to use a text box where subjects could enter a number

from 0 to 100. We chose the slider format over the text box because we thought it could help subjects understand the task better due to its visual representation of the entire range of beliefs together with labels about the meaning of these beliefs. A slider interface has been successfully used in previous studies, for example by Hill (2017) and Mobius et al. (2011).

In Part 1, we asked about five physically randomized events similar to the kinds of events we asked in the (unincentivized) numeracy questions. These are shown in a fixed order because there is a natural sequence from easier to harder items. Instead of describing a hypothetical event, however, we described the event as occurring "behind the scenes" (and honestly followed through by performing each physical event prior to computing payments). For example:

> We will roll a fair, 6-sided die behind the scenes, with the numbers 1 through 6 on its sides. Now, consider the following statement:
>
> **"The outcome of the die roll is a number less than or equal to 6."**
>
> How likely do you think it is that the above statement is TRUE?

Note also that the belief we elicit is described as pertaining to the truth of a statement, which allows us to elicit participants' beliefs about it by asking an identical question ("How likely...?") for the different kinds of beliefs in all five parts of the experiment.

Part 1 elicits *induced* probability beliefs, because the statements are about clearly defined random, physical events. Consequently, each item has a corresponding correct objective probability that we can compare to the participants' responses (hence we also refer to these as "calibration" items). We included one statement that must be true (the one above about the roll of a six-sided die being less than or equal to six) and another that must be false (similar to the one above but where the die roll is equal to 0). We would expect participants who are paying attention to provide the corresponding objective probabilities (100 and 0, respectively) even if they had relatively poor numeracy. Second, three of the statements have direct counterparts to the numeracy items, so directly comparing them allows us to assess the effects of incentives, and whether they encourage accuracy or might

instead distort responses. When we analyze accuracy, we focus on these induced probability beliefs.

In Part 2, we elicited *first-order* beliefs about participants' own factual knowledge—their confidence about the truth of the four statements in the knowledge quiz. The subjects reported their belief for each statement one by one, and the order in which the statements were presented was randomized. We took precautions to guard against looking up answers. First, the statements were presented as pictures so that the text could not be copied and pasted quickly into a Google search. While this does not prevent participants from manually typing the text into a search, it does increase the time it takes to look up the answer. Second, other than the question about West Virginia's side in the Civil War, the answers were not immediately obvious from the search results page. Note that in contrast to Part 1 in which subjective beliefs could be compared to objective probabilities due to the nature of the event, there are no such objective benchmarks for assessing accuracy in Part 2.

In Parts 3 through 5, we elicited three variations of *second-order* beliefs, which we also refer to as social beliefs. Specifically, we asked for beliefs about the accuracy of other participants' factual knowledge—if some other randomly selected participant correctly guessed whether the item in the pre-treatment knowledge quiz was true or false. In Part 3, the belief pertains to one other participant drawn from the entire sample of participants; hence, Part 3 elicits unconditional beliefs about others. To elicit such beliefs, participants were told that they will be randomly matched with another participant in the study, which will be referred to as "Person A". Then, for each of the four factual knowledge statements, we asked the following question: "How likely do you think it is that Person A was CORRECT in rating the following statement as true or false?".

In Parts 4 and 5, the beliefs pertain to one other participant drawn from an identifiable subset of participants, either the set of Democrats or Republicans, with each part corresponding to one partisan subgroup and the order of the subgroups randomized across participants. In particular, participants were told that in that part, they will be randomly

matched with another participant in the study who identifies with the Democratic (Republican) Party, and this person will be referred to as "Person D (R)". Hence, Parts 4 and 5 elicit beliefs that are conditional on knowing something about the other participant's subgroup. Then, to elicit beliefs regarding the accuracy of members from the two different parties, we asked the same question as in Part 3, but now referring to Person D and R, respectively.

Within each of these three parts, the order of statements is randomized. Note that there is an underlying objective benchmark in Parts 3 through 5 (the accuracy of guesses within each group in the sample). However, in contrast to the stated objective probabilities in Part 1, which can be deduced from the text of the statements themselves, participants have no direct access to information about the accuracy of others.

After all of the belief elicitation tasks were complete, we asked subjects a final survey question about their subjective *confidence* that they were maximizing their payoffs when reporting their beliefs (5-point Likert scale, with higher numbers representing higher confidence levels). Although this item makes little sense in FLAT, we included it for consistency across treatments.

Once all participants completed the study, we performed the physical randomization of the events for the calibration beliefs in Part 1 (once per event), and then performed the other randomizations electronically to compute each participant's total bonus.

2.1 Elicitation methods

Our experiment compares versions of two widely used belief elicitation mechanisms (BDM, BSR) against an unincentivized baseline (FLAT). The BDM incentive scheme is a version of the Becker, DeGroot and Marschak (1964) reservation-price elicitation mechanism adapted to elicit probabilities and using binary lotteries to determine payment. This mechanism was used by Grether (1981, 1992) and in several studies on beliefs and learning (Hao and Houser, 2012; Holt and Smith, 2009, 2016; Mobius et al., 2011). The BSR incentive scheme is a version of the widely-used Quadratic Scoring Rule (Brier et al., 1950; Harrison, Martínez-

Correa and Swarthout, 2013; Hossain and Okui, 2013; McKelvey and Page, 1990; Selten, 1998) that also uses binary lotteries for payment (also see Danz, Vesterlund and Wilson, 2020, for extensive references).

We chose these incentive schemes in part because of their popularity in the literature, but also because they have common features that ensure their comparison is on as even footing as possible. Indeed, both incentive schemes rely on lotteries with the same two prize values, which means neither mechanism's incentive compatibility depends on risk preferences. Both mechanisms have also been shown to be incentive compatible under less restrictive assumptions than expected utility maximization (Harrison, Martínez-Correa and Swarthout, 2013; Karni, 2009). In addition, both mechanisms can be explained using a set of rules that do not involve mathematical formulas, which is important in an online environment because, compared to a lab setting, we expect greater heterogeneity in numeracy in the participant pool and have less control over subject's attention as experimenters. This suggests that methods that are easier to comprehend should be preferable.

As is standard in laboratory experiments, we completely and truthfully explained the mechanics of the incentive schemes to participants. However, we did not provide explanations of their incentive compatibility. This was partly to put some limits on the time it would take to complete the study and to reduce cognitive load given that the rules for the incentive schemes may still seem complicated, even without mathematical formulas. One advantage of BDM over BSR (as typically implemented in laboratory settings) is that incentive compatibility of BDM can be explained more simply in terms of dominance arguments (Healy, 2018), whereas incentive compatibility of BSR requires the use of the scoring rule formula to explain the maximization of the objective function. In addition, the multiple price list format can be used for the BDM to make the underlying binary choice representation more concrete and easier to understand. In our design, we chose to hold constant the format of the direct elicitation while only varying the incentive mechanisms. Varying the format could also generate differences between BDM and BSR, although neither Burfurd and Wilkening

(2018) nor Holt and Smith (2016) find significant differences between BDM elicitations using direct versus list formats.

Despite the absence of the incentive compatibility explanation, the information provided to participants was sufficient to verify incentive compatibility for participants inclined to do so. Moreover, while we provided complete and accurate information about the incentive structure, we tried not to provide too much quantitative detail about the incentives in light of the findings by Danz, Vesterlund and Wilson (2020) that doing so increases discrepancies between elicited beliefs and the objective probabilities of a task that induce them. Instead of explaining incentive compatibility, we simply stated this fact in the BDM and BSR conditions, telling participants: "You will maximize your chance of earning the bonus for each statement if you report your beliefs as accurately as possible."Following the explanation of the incentives, we also added: "This procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible about the likelihood you think the statement is TRUE." For comparability, the statement in the FLAT condition was modified to read: "You should report your beliefs as accurately as possible." Then, in all conditions, these statements were followed by: "That is, there is nothing to gain by stating a number that differs from what you actually believe." Including such language explicitly stating incentive compatibility follows common practices in previous experiments eliciting subjective beliefs (e.g. Ambuehl, 2017; Armantier and Treich, 2013; Danz, Vesterlund and Wilson, 2020; Enke, Schwerter and Zimmermann, 2020; Hill, 2017; Holt and Smith, 2016). Moreover, note that any part of the instructions that is not directly related to the incentives mechanism is constant across conditions and so, any differences between treatments can only be due to differences in the incentive schemes.

2.1.1 Flat fee (FLAT)

In the FLAT condition, we simply paid subjects a constant \$0.20 for each belief elicitation part. It is possible that providing a flat payment encourages participants to exert some effort

on the task, akin to motives induced by a gift exchange (Akerlof, 1982; Fehr, Kirchsteiger and Riedl, 1993), compared to a scheme in which there are no payments at all (such as on a traditional survey). Nevertheless, the payment in FLAT is unrelated to the value of the belief, and so the accuracy of beliefs remains unincentivized. Hence, FLAT is our control condition and provides a baseline for accuracy in the absence of an incentive-compatible mechanism. Figure 2 includes the depiction of the instructions for the FLAT method, as presented to the participants in our study.

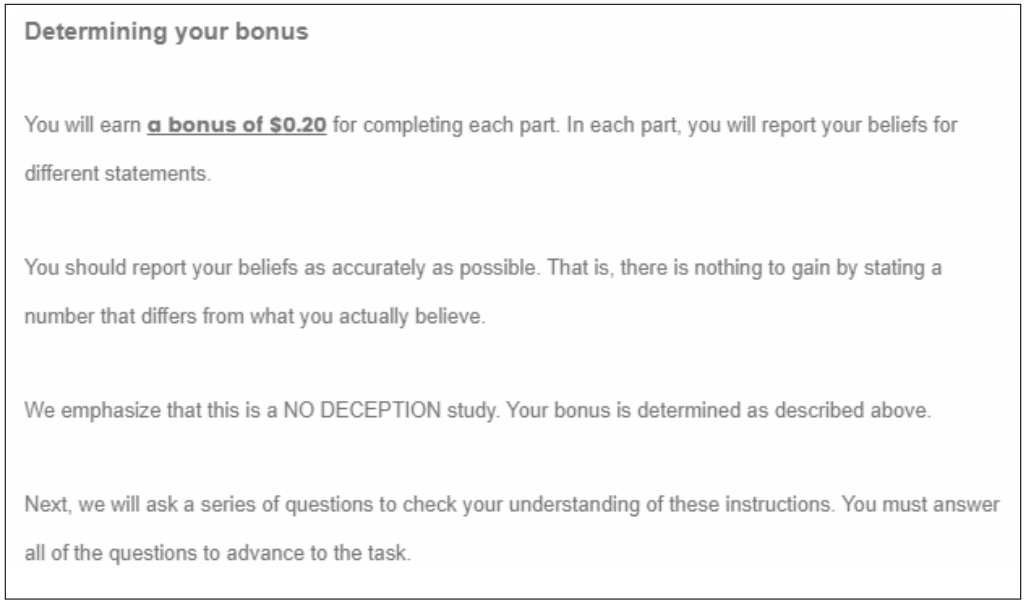
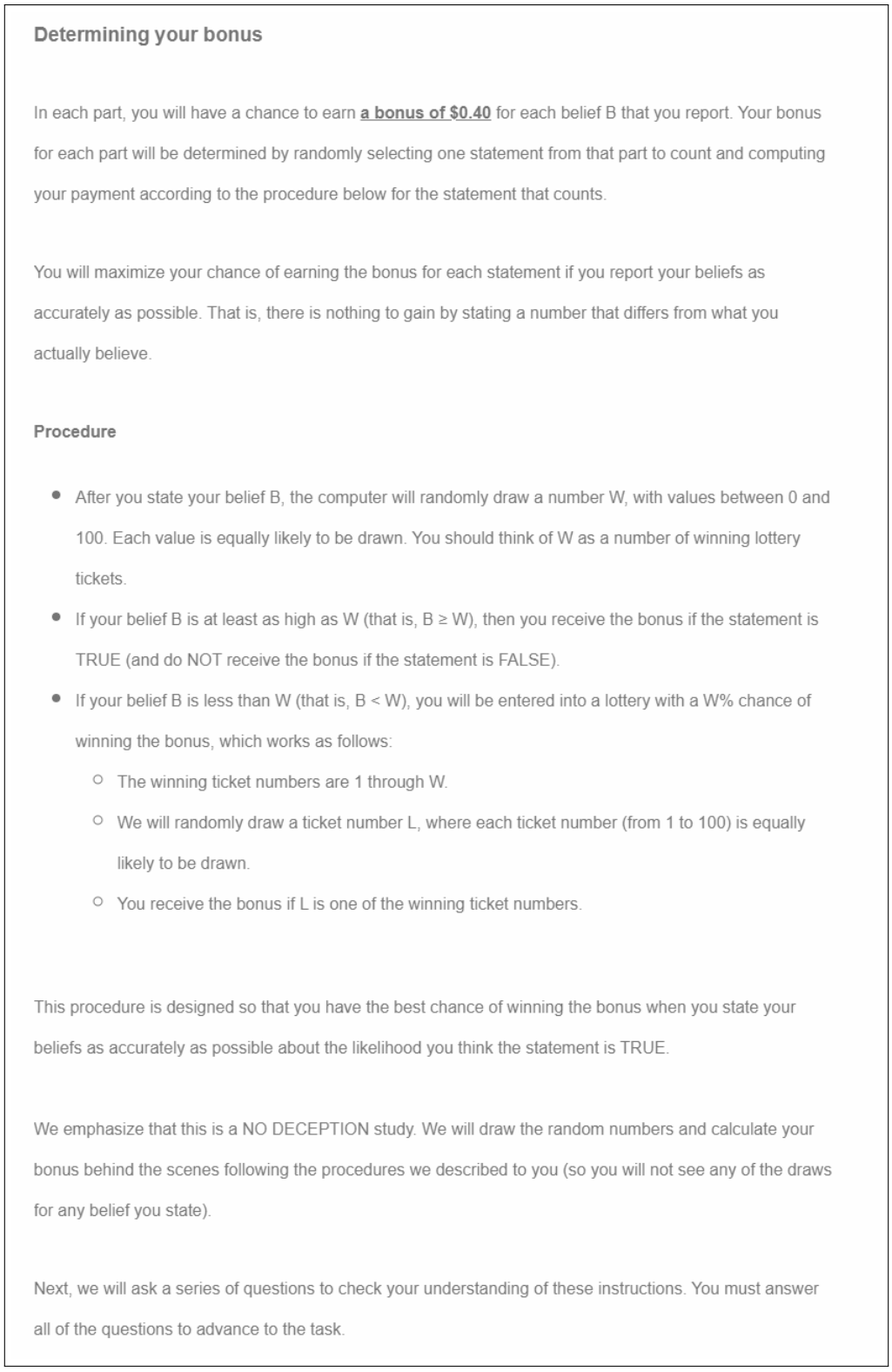


Figure 2: Instructions screen for FLAT mechanism as presented to study participants

2.1.2 Stochastic Becker-DeGroot-Marschak (BDM)

This mechanism elicits the value B for which the subject prefers to be paid based on the truth of the statement rather than any objective lottery of winning the prize with probability less than B. In other words, the value of B is the "crossover" or "switch point" between preferring a bonus that depends on one's subjective beliefs over a bonus that depends on objective lotteries. Our implementation is similar to Mobius et al. (2011) and closely follows Hill (2017). Figure 3 includes the depiction of the instructions for the BDM method, as presented to the participants in our study.



Determining your bonus

In each part, you will have a chance to earn **a bonus of $0.40** for each belief B that you report. Your bonus for each part will be determined by randomly selecting one statement from that part to count and computing your payment according to the procedure below for the statement that counts.

You will maximize your chance of earning the bonus for each statement if you report your beliefs as accurately as possible. That is, there is nothing to gain by stating a number that differs from what you actually believe.

Procedure

- After you state your belief B, the computer will randomly draw a number W, with values between 0 and 100. Each value is equally likely to be drawn. You should think of W as a number of winning lottery tickets.
- If your belief B is at least as high as W (that is, B ≥ W), then you receive the bonus if the statement is TRUE (and do NOT receive the bonus if the statement is FALSE).
- If your belief B is less than W (that is, B < W), you will be entered into a lottery with a W% chance of winning the bonus, which works as follows:
  - The winning ticket numbers are 1 through W.
  - We will randomly draw a ticket number L, where each ticket number (from 1 to 100) is equally likely to be drawn.
  - You receive the bonus if L is one of the winning ticket numbers.

This procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible about the likelihood you think the statement is TRUE.

We emphasize that this is a NO DECEPTION study. We will draw the random numbers and calculate your bonus behind the scenes following the procedures we described to you (so you will not see any of the draws for any belief you state).

Next, we will ask a series of questions to check your understanding of these instructions. You must answer all of the questions to advance to the task.

Figure 3: Instructions screen for BDM mechanism as presented to study participants

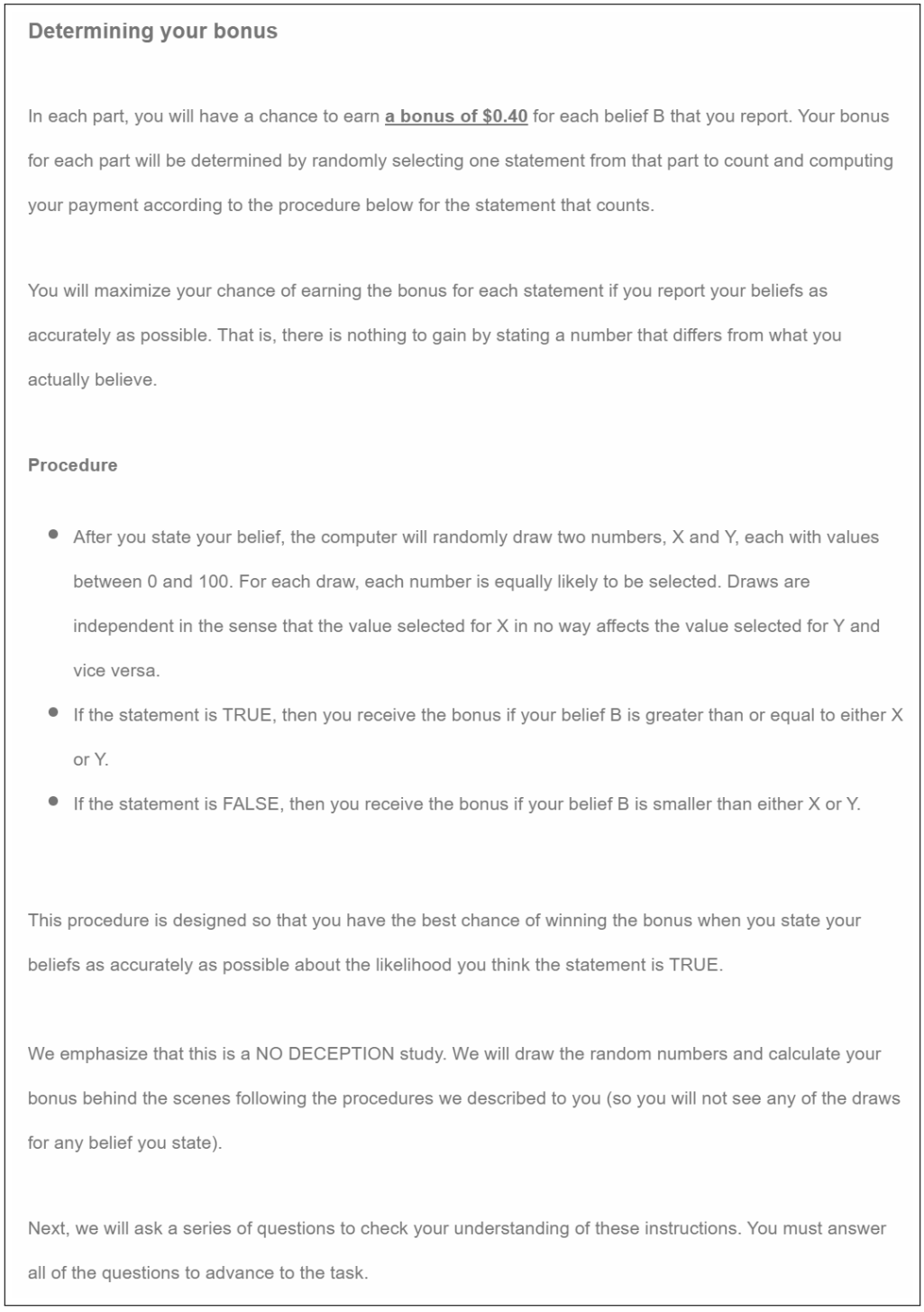
Subjects in BDM were informed that for each elicitation part, one of the reported beliefs will be randomly selected to count for payment. They were then told that if a given belief B counted for determining their bonus, we would randomly draw a whole number W from 0 to 100. If $B \geq W$, they would receive a bonus of \$0.40 if the statement were true and no bonus if the statement were false. If $B < W$, then they would be entered into a lottery with a $W\%$ chance of receiving the \$0.40 bonus. The lottery was described as involving 100 lottery tickets and winning the bonus if the randomly drawn ticket had a "winning" number (any ticket numbered 1 through W). The size of the bonus in BDM (and BSR described below) is in line with other online experiments eliciting subjective beliefs (e.g. Coffman, Collis and Kulkarni, 2019; Hill, 2017; Roth and Wohlfart, 2020), and is held constant across conditions enabling us to attribute any potential differences to the incentives mechanism.

2.1.3 Binarized Scoring Rule (BSR)

This method elicits B by offering a fixed prize with probability $1-(1-B)^2$ if the statement is true and with probability $1-B^2$ if the statement is false; otherwise, the prize is not received. The probabilities are equivalent to the variable prize values in the Quadratic Scoring Rule (Brier et al., 1950; Savage, 1971), and when they are converted to probabilities of winning a fixed prize, the procedure is known as the Binarized Scoring Rule. Like in the BDM condition, subjects were informed that one belief per part will be randomly selected to count for payment.

We rely on an implementation by Wilson and Vespa (2017), also used in Danz, Vesterlund and Wilson (2020), that cleverly describes the probabilities without the use of formulas. Specifically, subjects were told that after stating their belief B, the computer will randomly draw two numbers, X and Y, with each being a whole number from 0 and 100 that is equally likely and drawn independently. If the statement is true, the subject receives a bonus of \$0.40 if and only if B is greater than or equal to either X or Y. If the statement is false, they receive the bonus if and only if B is smaller than either X or Y. It is straightforward to

verify that this procedures generates the probabilities of $1-(1-B)^2$ and $1-B^2$ for true and false statements, respectively. Figure 4 includes the depiction of the instructions for the BSR method, as presented to the participants in our study.



Determining your bonus

In each part, you will have a chance to earn **a bonus of $0.40** for each belief B that you report. Your bonus for each part will be determined by randomly selecting one statement from that part to count and computing your payment according to the procedure below for the statement that counts.

You will maximize your chance of earning the bonus for each statement if you report your beliefs as accurately as possible. That is, there is nothing to gain by stating a number that differs from what you actually believe.

Procedure

- After you state your belief, the computer will randomly draw two numbers, X and Y, each with values between 0 and 100. For each draw, each number is equally likely to be selected. Draws are independent in the sense that the value selected for X in no way affects the value selected for Y and vice versa.
- If the statement is TRUE, then you receive the bonus if your belief B is greater than or equal to either X or Y.
- If the statement is FALSE, then you receive the bonus if your belief B is smaller than either X or Y.

This procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible about the likelihood you think the statement is TRUE.

We emphasize that this is a NO DECEPTION study. We will draw the random numbers and calculate your bonus behind the scenes following the procedures we described to you (so you will not see any of the draws for any belief you state).

Next, we will ask a series of questions to check your understanding of these instructions. You must answer all of the questions to advance to the task.

Figure 4: Instructions screen for BSR mechanism as presented to study participants

3 Results

3.1 Sample Demographics

Our online sample is more diverse in several respects than a typical laboratory sample of undergraduates, as we expected. Only 8% of our sample were college-aged (between 18 and 24 years old), with 49% between 25 and 34 years old and 10% who were at least 55 years old. In terms of educational attainment, the sample was more educated overall, while at the same time more diverse: 15% had no more than a high school diploma or equivalent, 44% had completed a 4-year undergraduate degree, and 8% had obtained a post-graduate degree. While our sample is more diverse than laboratory samples, it is still a convenience sample that is younger and more educated than the population as a whole. According to the U.S. Census Bureau's 2019 Current Population Survey, adults 55 years old or older constitute 38% of all adults over 18 years old in the U.S. population (compared to 10% in our sample), and 39% of the population completed no more than a high school education (compared to 15% of our sample).

Our sample also skewed white (83%), male (60%), and Democratic in party identification (49% Democrats, 26% Republican). By comparison, 76% of the U.S. population identifies as "white only" according to the Census Bureau, and 52% of the adult population is male. On the December 2019 Gallup Poll, 28% of respondents identified as Democrats and 28% identified as Republican. There were no significant differences in the distributions of any of these characteristics across treatments. See the Online Appendix (Table A3) for summary statistics across treatments.

3.2 Attention and Numeracy

Do respondents in online samples pay enough attention? Do they have enough cognitive sophistication to comprehend incentivized probability judgment tasks? We turn next to describing individual measures of attention and numeracy. For our attention check, we

used a single item with a 5-point response scale about government services and spending and, in the second sentence of the item, instructed participants to "select the numbers one and four no matter what your own views are." Our attention check therefore requires rather minimal attention and screens out participants who did not bother to read the second sentence. Overall, 88% of participants in our sample followed the instructions, while 12% did not pay close enough attention to do so (this distribution is not significantly different across treatments - see Table A4 in Online Appendix for details). While the level of attention necessarily varies between and within participants over the course of their participation in the study, this does suggest that we can expect a non-trivial amount of error due to inattention. Similar inattention rates have been documented in previous research (Goodman, Cryder and Cheema, 2013) but their size varies widely across studies (Paas and Morren, 2018).

More directly related to the belief elicitation task are the unincentivized probability questions that we included to measure numeracy. Participants gave responses to these questions using a slider, similar to the slider in the elicitation task, and we counted the number of correct answers given by each participant to generate a numeracy score. The average numeracy score was 3.1, with only 8% answering all six questions correctly and 18% answering none of them correctly. The results do not change much if we allow for error by counting answers within a range of the correct answer. For a range of ± 2, the average score is 3.4, and for a range of ± 5, it is 3.7.

We also find that our measure of numeracy is related to attention: 52% of participants who fail the attention check have a score of 0, while only 14% of those who pass the attention check do. Nevertheless, even if we excluded participants who failed the attention check, the average numeracy score increases only slightly to 3.3. Based on this, we categorize participants in two groups: high numeracy, if the average numeracy score was strictly greater than 3, and low numeracy otherwise. Overall, 54% of participants fit the low numeracy description and 46% the high numeracy one (this distribution is not significantly different across treatments - see Table A4 in the Online Appendix for details).

An additional pre-treatment measure related to attention and numeracy is participants' comprehension of the belief scale. Recall that we explained the meaning of beliefs to participants as in Table 3 and then asked a series of five multiple choice comprehension questions. For each question, participants had two opportunities to answer correctly. If a participant gave an incorrect answer, we showed them Table 3 again (which was not displayed during the first attempt) before asking them to answer the question again. If they answered incorrectly a second time, we then showed them the correct answer. Overall, 76% of questions were answered correctly on the first attempt. Participants gave an average of 1.9 incorrect answers (including second attempts) on the belief comprehension questions, with 43% of participants answering every question correctly on the first attempt and 66% answering every question correctly on either the first or second attempt. Note that a third of participants fail to answer at least one comprehension question correctly at all.

Overall, we find that a non-trivial proportion of participants fail the attention check, do not exhibit high levels of numeracy, or incorrectly answer comprehension checks about the meaning of the belief scale. These findings indicate that a sizeable share of participants recruited from online sources are either not devoting sufficient cognitive effort or may not have the requisite mathematical ability to understand the kinds of incentive mechanisms typically used in belief elicitation tasks administered in university laboratories. We suspect that this finding extends to samples more representative of local or national populations generally, although our data cannot speak to this possibility. Moreover, whether or not incentives might increase or decrease attention, comprehension, or accuracy in probabilistic reasoning is an empirical question that we cannot address directly (since we do not have a completely unincentivized baseline for comparison). To the extent that incentive mechanisms are more complicated than the comprehension checks or the numeracy questions, the individual cognitive measures suggest that anywhere between one-third to two-thirds of our sample could have difficulty comprehending the task, potentially affecting the quality of the data.

3.3 Task Comprehension

Result 1. *Participants have the most difficulty understanding BDM than either BSR or FLAT, with lowest comprehension rates and highest perceived difficulty for BDM.*

Participants answered four comprehension questions specific to the incentive mechanism in their condition. As with the belief scale comprehension questions, they had two opportunities to provide the correct answer for each question and were provided a reminder of the rules for the mechanism if they answered the first attempt incorrectly. Overall, 72% of comprehension questions were answered correctly on the first attempt in FLAT. The rate was similar in BSR, with 73% correct first attempts, but significantly lower in BDM, with 47% correct first attempts. We caution that the comprehension questions are not strictly comparable because they must be tailored to each mechanism. Although we asked similar questions across treatments (e.g. "Suppose you state a belief B equal to 28, which of the following is correct?"), the format of the multiple choices differed as it had to relate to the different random mechanisms that determined their bonus. As such, it is possible that these differences in error rates regarding the comprehension quiz may be due to differences in the difficulty of the questions rather than the underlying difficulty of the mechanisms.

Recall that we also asked participants for their self-reported perceptions of the difficulty of comprehending the instructions and the effort they exerted to understand the instructions, both on a 5-point scale. These subjective perceptions were elicited following the incentive comprehension questions but prior to beginning the belief task. We find that participants perceive BDM as the most difficult (mean of 3.60), followed by BSR (mean of 2.96), and FLAT perceived (naturally) as the least difficult (mean of 2.43). These differences are statistically significant ($p < .01$; see Table A5 in the Online Appendix for the regression table). Similarly, we find that participants rated BDM as requiring the most effort (mean of 3.76), followed also by BSR (mean of 3.43), and then FLAT (mean of 2.83). These differences are also statistically significant ($p < .01$; Table A5 in the Online Appendix). The perceived difficulty and effort differences between the two incentive compatible methods and FLAT are

in line with the ranking put forth by Charness, Gneezy and Rasocha (2021) where BDM and BSR are considered to be complex mechanisms. Therefore, participants exposed to these methods may incur additional complexity costs. However, their paper refers to the BSR rule as a "very complex mechanism" while BDM is considered to be only "more complex". Our findings suggest the opposite might be the case when we implement simplified instructions for these methods, as in our study. In terms of a third self-reported measure, elicited on a five-point scale at the end of the study, we did not find any differences between BSR (mean of 3.43) and BDM (mean of 3.43) in participants' confidence that they were reporting beliefs that maximized their earnings.

3.4 Duration

Result 2. *FLAT requires the least amount of time allocated to reading and understanding the incentives. BDM requires the least amount of time allocated to belief reporting.*

We also look at the average time subjects take to complete different individual parts of the study. Table 4 presents the overall duration in each treatment as well as the duration of individual parts of the study. Unsurprisingly, we find that subjects in the FLAT condition require significantly less time than those in BDM or in BSR for completing the study (see Table A6 in the Online Appendix for the linear regression results). As can be seen from Table 4 ('Incentives instructions reading' and 'Incentives comprehension quiz' lines), this difference is driven by the time allocated to reading the instructions about the incentive mechanism and answering the corresponding comprehension questions which in FLAT takes approximately half the time than in the other two treatments. There are no significant differences between BDM and BSR in overall completion times or in the time it takes participants to read the instructions regarding each payment mechanism and complete the corresponding comprehension quiz.

When focusing on the incentivized belief elicitation tasks (Part 1 - 5), we find that subjects in BDM are faster than in the BSR treatment ($p < .01$). Table A7 in the Online

Appendix presents the results from the mixed effects linear regression with part, item and subject random effects.

	FLAT	BDM	BSR
Overall	703	880	907
	(32.3)	(36.8)	(34.3)
Numeracy quiz	16.6	15.7	16.1
	(0.74)	(1.29)	(1.41)
Belief meaning (instructions reading + comprehension quiz)	115	107	119
	(7.30)	(7.46)	(13.9)
Incentives instructions reading	15.5	42.7	41.2
	(2.31)	(5.83)	(3.49)
Incentives comprehension quiz	50.4	113	120
	(2.45)	(6.24)	(8.48)
Part 1 (induced probability beliefs)	12.3	10.8	13.5
	(0.53)	(0.41)	(0.89)
Part 2 (first order beliefs)	8.03	6.94	8.77
	(0.55)	(0.29)	(0.69)
Part 3 (second order beliefs about all others)	7.25	7.09	7.82
	(0.36)	(0.32)	(0.61)
Part 4 (second order beliefs about Democrat others)	5.94	5.10	5.92
	(0.31)	(0.29)	(0.60)
Part 5 (second order beliefs about Republican others)	5.94	5.39	6.15
	(0.57)	(0.28)	(0.30)

Note: Standard errors in parentheses.

Table 4: Average duration in seconds for different parts of the study

3.5 Induced Probability Beliefs

For each calibration item, Figure 5 shows overlaid kernel density plots of the distributions of elicited beliefs for each incentive mechanism. With two exceptions, Anderson-Darling tests cannot reject the equality of any two distributions across the five items. The significant differences occur when comparing FLAT with BDM for Probability 3 ($p = 0.033$), and when comparing FLAT with BSR for Probability 4 ($p = 0.015$).

Next, we assess the quality of beliefs elicited on the calibration items. Importantly, we can measure the accuracy of these beliefs against the objective benchmarks corresponding to

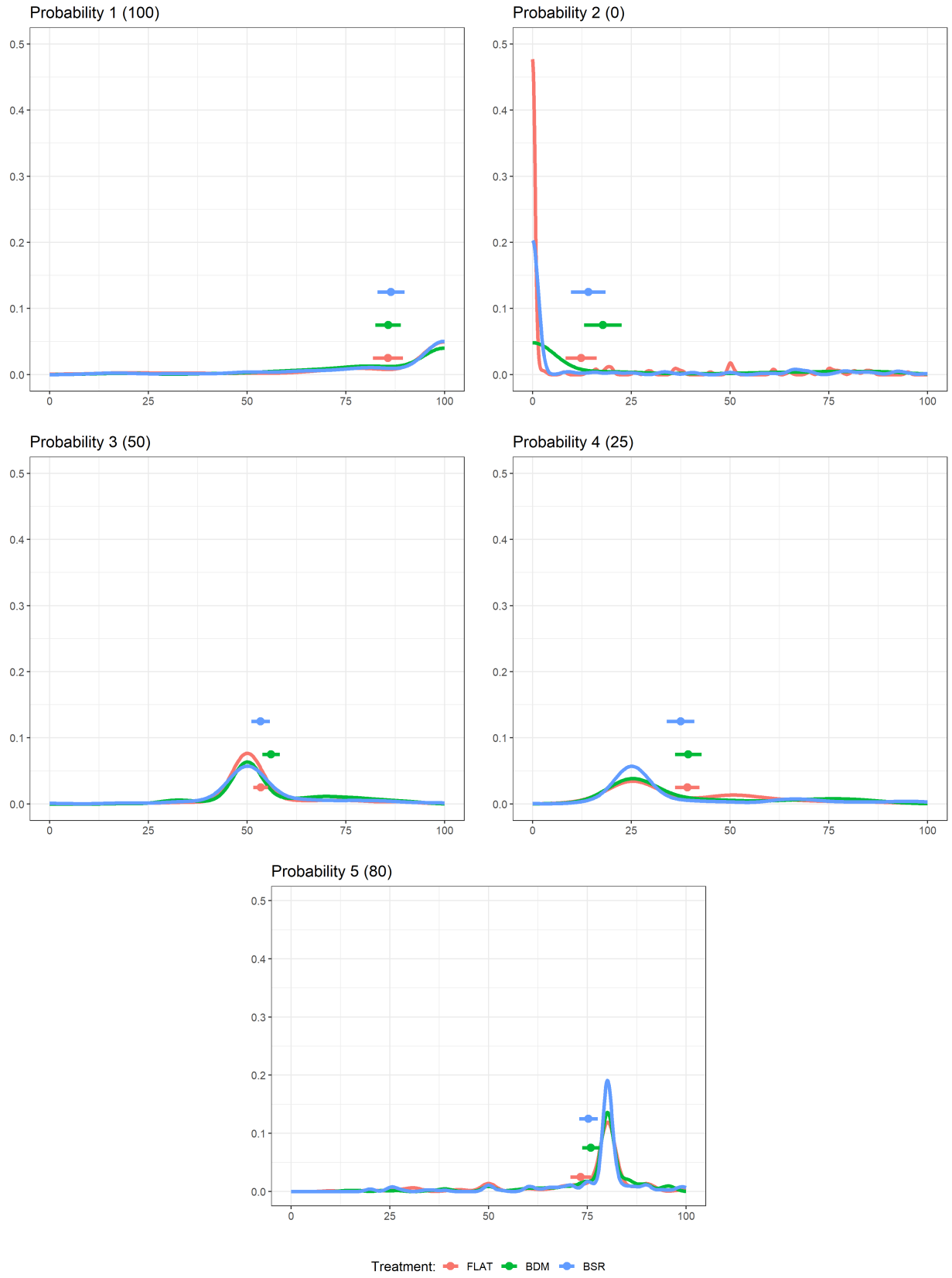


Figure 5: Kernel density plots of the distributions of induced beliefs for FLAT (red line), BDM (green line) and BSR (blue line). The corresponding means and 95% confidence intervals for each method are depicted with dots and horizontal bars.

the explicit descriptions of the probabilistic events. We measure the quality of these beliefs in two ways. First, we code a belief as *accurate* if it is equal to the objective probability corresponding to the event (and inaccurate otherwise). We acknowledge that although there is an objective probability that corresponds to each event, participants may have subjective beliefs that differ from the objective probability. Because such subjective beliefs are necessarily unobservable, we cannot directly assess how well the incentive mechanisms elicit subjective beliefs. That is, we cannot tell if subjects are "truthfully" reporting their beliefs. Hence, we deliberately use the term *accuracy* in reference to the objective benchmark. Overall, 59.5% of beliefs were accurate in FLAT, 53.6% in BDM, and 61.6% in BSR. We compare these values across treatments using mixed effects linear regressions with subject and item random effects (see Table 5 column (1)). Although the accuracy is lowest in BDM, we find no significant difference across treatments.

Result 3. *There are no significant differences across incentive mechanisms in the average accuracy and error size of the induced probability beliefs.*

Second, we compute the *size of the error* in terms of the absolute difference between the elicited belief and the benchmark objective probability. We do not find any significant differences in the average error size across the three incentive schemes (see Table 5 column (3)), with an average error of 11.3 percentage points in FLAT, 12.8 in BDM, and 11.3 in BSR.

When we directly compare beliefs elicited on the three calibration items with corresponding numeracy items with the same benchmark probabilities, we find that the calibration beliefs are *worse* than answers to the numeracy questions. Specifically, we find that elicited beliefs were 12.3 percentage points less accurate in FLAT than probabilities reported on the numeracy items, compared to 16.3 in BDM and 12.2 in BSR. Similarly, we find that error size is 3.43 higher in FLAT, 2.68 higher in BDM, and 3.00 higher in BSR (see Table 6 for the regression results). That all of these values are significantly different from 0 suggests that incentives might lead to lower quality responses than completely unincentivized survey

	Dependent variable:			
	Accurate response		Size of error	
	(1)	(2)	(3)	(4)
BDM	−0.059	−0.070	1.479	2.597
	(0.043)	(0.053)	(1.602)	(2.167)
× Failed attention check		−0.095		−0.754
		(0.095)		(3.864)
× Low numeracy		−0.102		4.518
		(0.062)		(2.515)
× No college degree		0.137		−6.359*
		(0.062)		(2.521)
BSR	0.022	−0.063	0.029	2.719
	(0.043)	(0.053)	(1.600)	(2.152)
× Failed attention check		0.039		−1.027
		(0.102)		(4.124)
× Low numeracy		−0.093		4.051
		(0.063)		(2.549)
× No college degree		0.162*		−6.038
		(0.063)		(2.555)
Failed attention check		−0.142		7.450*
		(0.066)		(2.685)
Low numeracy		−0.429***		13.478***
		(0.043)		(1.763)
No college degree		−0.098		2.216
		(0.044)		(1.795)
Constant	0.595***	0.874***	11.293***	2.579
	(0.030)	(0.039)	(1.137)	(1.585)
Observations	2,350	2,350	2,350	2,350
Log Likelihood	−1,313.059	−1,170.172	−10,236.500	−10,096.320

Note: *p<0.05; **p<0.01; ***p<0.001; p-values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure

Table 5: Incentives, cognition, and accuracy of induced probability beliefs

questions. We caution, however, that we cannot draw a clear causal inference that this is due solely to incentives, because these are within-subject differences that might also have been affected by other factors such as fatigue or experience. Nevertheless, we also note that none of the treatment differences are statistically significant.

	Dependent variable:	
	Accuracy difference	Size of error difference
	(1)	(2)
BDM	−0.041	−0.755
	(0.036)	(1.093)
BSR	0.0002	−0.439
	(0.036)	(1.092)
Constant	−0.123*	3.434*
	(0.062)	(1.344)
Observations	1,410	1,410
Log Likelihood	−859.275	−5,731.170

Note: Model includes item and subject random effects.
*p<0.05; **p<0.01; ***p<0.001

Table 6: Belief accuracy on incentivized probability items vs. unincentivized numeracy items

Overall accuracy rates in the range of 53-60% may not seem particularly assuring that online samples would yield reliable data for research eliciting beliefs to understand probability judgments, learning, or decisions under risk. Recall that a non-trivial proportion of our participants either failed the attention check or had difficulty comprehending at least some aspects of the task, whether it was the belief scale or the incentive schemes. We therefore investigate how attention and comprehension are related to accuracy, and we are particularly interested in whether the quality of beliefs elicited by each incentive mechanism might vary with cognitive effort or sophistication—that is, on their interaction.

Result 4. *The most attentive, numerate, and educated participants have the highest accuracy and the lowest average error size for induced beliefs, which do not vary by incentive*

mechanism. Attention, numeracy, and education are all positively related to accuracy.

Table 5 presents estimates from a mixed effects regression model that includes treatment indicators, measures of cognitive effort and sophistication, and their interactions (columns (2) and (4)). The measures of effort and sophistication include an indicator for whether the participant failed the attention check, an indicator for whether the participant scored low on numeracy, and an indicators for whether or not the respondent completed a college degree. In this specification, the excluded categories are such that the baseline participant passed the attention check, is high in numeracy, is more highly educated, and was assigned to the FLAT incentive condition. The constant term therefore reflects the accuracy of participants we would expect to be the most accurate in the absence of an incentive compatible mechanism. Indeed, we find that such respondents provide much more accurate beliefs than the average respondent. The estimate for the intercept in the second column (with accurate responses as the dependent variable) suggests that 87.4% of attentive, numerate, highly educated respondents give accurate responses. The estimate in the fourth column (with the size of error as the dependent variable) implies that the average error size for such respondents is around 2, and it is not significantly different from 0. Furthermore, we find that neither of the main treatment effect estimates for either BDM or BSR are statistically significant. Thus, attentive, numerate, highly educated participants report the most accurate beliefs and the quality of their beliefs does not depend on the incentive mechanism.

We find that the coefficients for each of the factors we would expect to be associated with lower accuracy are, in fact, all negative and statistically significant. The estimates imply that accuracy is 14.2 percentage points lower for participants failing the attention check and 9.8 percentage points lower for those with no college degree (compared to those passing the attention check and who completed a 4-year college degree). The largest decrease in accuracy corresponds to low numeracy, with the model suggesting a 42.9 percentage point decrease—that is, that low numeracy respondents are nearly half as accurate as high numeracy respondents. We therefore find that individual participants scoring lower on measures

associated with cognitive attention, effort, or ability report less accurate beliefs.

Does incentive compatibility induce some of these participants to exert greater attention or effort, thereby increasing their accuracy? That is, do any of these factors interact with the incentive mechanisms?

Result 5. *Incentives (BDM and BSR) increase accuracy and decrease error size of induced beliefs for less educated participants.*

In contrast with the interactions with failing the attention check or with low numeracy (none of which are significant), the interactions with education level are significant and positive. Thus, although we do not find that incentives increase accuracy for those failing the attention check or scoring low on numeracy, we do find that both the BDM and the BSR treatments increase accuracy among participants with no college education—enough so that the accuracy of beliefs of less educated participants elicited by incentive-compatible mechanisms is comparable to the most educated participants in FLAT. Of course, we cannot interpret these interactions as causally related to education, as highly educated online respondents may be quite different from less educated online respondents in unobserved ways, and education may be correlated with other unobservable characteristics such as income. Nevertheless, we say that incentives matter: we find that incentive compatible mechanisms increase the accuracy of beliefs for the least educated participants in our sample.

3.6 Factual Beliefs

Factual beliefs are first-order beliefs because they are beliefs about one's own knowledge. Such beliefs are inherently subjective, conveying an individual's degree of *confidence* that a given statement is true or false. Unlike with induced probabilities, we cannot compare them to an objective benchmark to gauge their accuracy. In this section, we instead examine whether incentives generate different distributions of beliefs and, more specifically, the extent to which the incentive scheme might affect the level of confidence in the beliefs we elicit.

Result 6. *Distributions of beliefs elicited with incentives (BSR and BDM) are distinguishable from beliefs elicited in FLAT.*

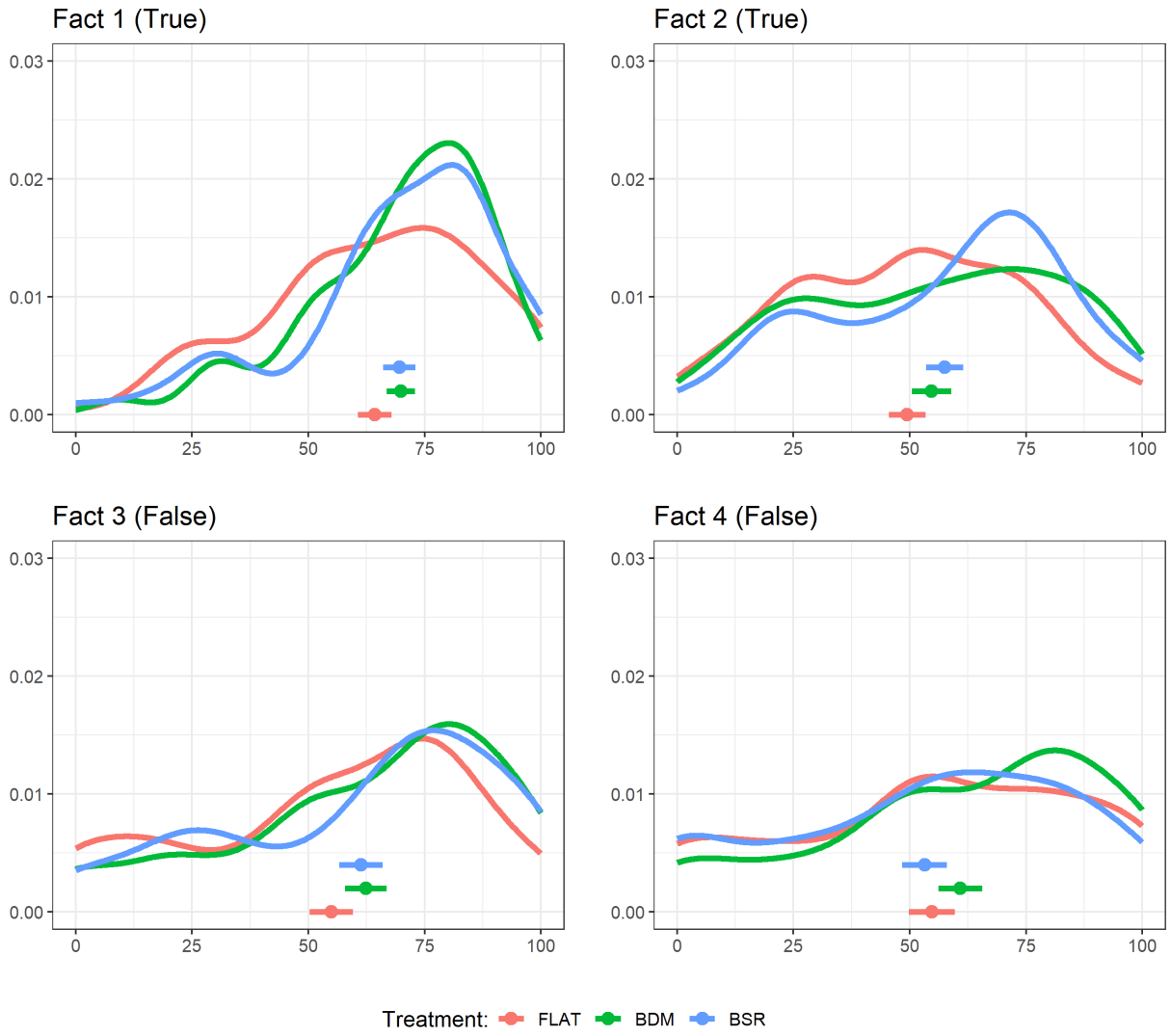


Figure 6: Kernel density plots of the distributions of factual beliefs for FLAT (red line), BDM (green line) and BSR (blue line). The corresponding means and 95% confidence intervals for each method are depicted with dots and horizontal bars.

For each factual item, Figure 6 shows overlaid kernel density plots of the distributions of elicited beliefs for each incentive mechanism. These plots suggest there are differences in the distributions associated with each incentive scheme, although it is difficult to discern any systematic patterns across the items from the densities alone. For Fact 1 (a true statement), both BDM and BSR have distributions with greater mass near their respective modes (beliefs

around 75%), whereas in FLAT there is more mass near complete uncertainty (50%) as well as on the opposite side of the belief scale (near 25%). Beliefs for Fact 2 (also true) appear to be much more heterogeneous than Fact 1 across mechanisms, with beliefs elicited by BSR appearing to have a greater mass of beliefs above 50% than either BDM or FLAT; similar to Fact 1, we also see greater mass for Fact 2 around 50% in FLAT. For both of these facts, using a 5% significance level, Anderson-Darling tests reject the equality of the BSR and FLAT distributions (Fact 1: $p = 0.015$; Fact 2: $p = 0.003$) , as well as the equality of the BDM and FLAT distributions (Fact 1: $p = 0.011$; Fact 2: $p = 0.042$), but the equality of the BSR and BDM distributions cannot be rejected (Fact 1: $p = 0.855$; Fact 2: $p = 0.377$).

For Facts 3 and 4 (both false statements), we note there appears to be more mass on the right-hand side of these distributions. This suggests a tendency for subjects to believe statements to be more likely true than false—a kind of credulity bias, although we cannot establish this with any generality beyond these two items. For the latter two facts, the distributions appear to be more similar across incentive schemes than for Facts 1 and 2. Nevertheless, using a 5% significance level, Anderson-Darling tests reject the equality of BSR and FLAT ($p = 0.028$) and the equality of BDM and FLAT ($p = 0.021$) for Fact 3; the equality of the distributions for Fact 4 cannot be rejected.

We further note participants' tendencies to select middle (50%) beliefs, and that this type of central tendency bias has been documented using other belief elicitation methods such as the QSR (Crosetto et al., 2020). To investigate more systematically the differences between the incentive schemes in the reporting of these middle beliefs, we estimate three linear mixed effects models reported in Table 7. The dependent variable for the model in the first column is the distance between the elicited belief and 50%, while the dependent variable for the second column is an indicator for whether the belief is equal to 50%. In the third column, the dependent variable is an indicator for whether the respondent's belief deviates from 50% in the "right direction", i.e. if the belief is greater than 50% when the fact is true, or lower than 50% when the fact is false. The specifications include indicators for

each incentive mechanism, pre-treatment cognitive measures (as in Table 5), fixed effects for knowledge items (not reported for presentation purposes), and participant random effects. We also estimated models with interactions between treatments and cognitive measures. None of the interactions were significant, so we only report the results of models without the interactions.

	Dependent variable:		
	Distance from 50%	Belief equal to 50%	Belief in right direction
	(1)	(2)	(3)
BDM	1.737	−0.031	0.029
	(1.180)	(0.023)	(0.028)
BSR	2.490*	−0.085***	0.088**
	(1.180)	(0.023)	(0.028)
Failed attention check	1.428	−0.049	0.004
	(1.560)	(0.030)	(0.037)
Low numeracy	1.055	−0.054**	−0.005
	(0.985)	(0.019)	(0.023)
No college degree	−1.388	0.060**	−0.046*
	(0.985)	(0.019)	(0.023)
Constant	23.071***	0.113***	0.760***
	(1.211)	(0.023)	(0.032)
Observations	1,880	1,880	1,880
Log Likelihood	−7,664.715	−288.698	−1,214.593
F Statistic (BDM=BSR)	0.420	5.786*	4.511*

Note: Model includes item fixed effects (not reported) and subject random effects.
*p<0.05; **p<0.01; ***p<0.001

Table 7: Centrality of first order beliefs

Our results suggest that BSR systematically pushes participants "off the fence," eliciting beliefs farther away from 50% (column 1), and to report 50% less often (column 2) compared to FLAT. An F-test for equality of coefficients suggests that the frequency of 50% beliefs in BSR is significantly lower than in BDM as well. Furthermore, the regression results in the third column indicate that this departure from 50% beliefs may be beneficial as participants in the BSR treatment are more likely than those in FLAT and BDM to report

beliefs that are in the right direction.

3.7 Social Beliefs

Our final set of results pertain to *social* beliefs. These are second-order beliefs because they are beliefs about others' beliefs. Recall that we elicited beliefs about all other participants (*others*) and beliefs about two distinct subgroups of participants (*Democrats*, *Republicans*). As with first-order beliefs, second-order beliefs are subjective, so we analyze the differences in the distributions in a similar manner (also with attention to 50% beliefs).

To streamline the exposition, we present density plots only for *others* in Figure 7. Similar plots for beliefs about Democrats and Republicans can be found in the Online Appendix. We observe prominent modes at 50% for FLAT across all items. We also observe greater mass at 50% in BDM for all items, although the modes are more pronounced in FLAT than BDM. In contrast, the only noticeable central mode for BSR is for item 2. According to pairwise Anderson-Darling tests for differences in distributions, the FLAT and BSR treatments lead to different distributions for items 1 and 2 ($p < .05$), but the differences between FLAT and BDM are not significant. We cannot reject pairwise equality between FLAT and either incentivized mechanism for item 4, and we cannot reject the equality of BSR and BDM for any of the items. Consistent with our findings for first-order beliefs, we find that both of the incentive mechanisms elicit different beliefs than flat-rate payments.

Result 7. *Incentivizing belief elicitation using BSR leads to significantly lower frequencies of 50% beliefs compared to FLAT and BDM, irrespective of the second-order belief type. The differences between BDM and FLAT are not consistently significant. Moreover, the departures from 50% beliefs are always more significantly likely to be in the direction of the underlying probability in BSR when compared to FLAT, and sometimes also when compared to BDM.*

Table 8 presents estimates of the effects of incentive mechanisms on 50% beliefs, controlling for attention, numeracy, and education. The first three columns show coefficients

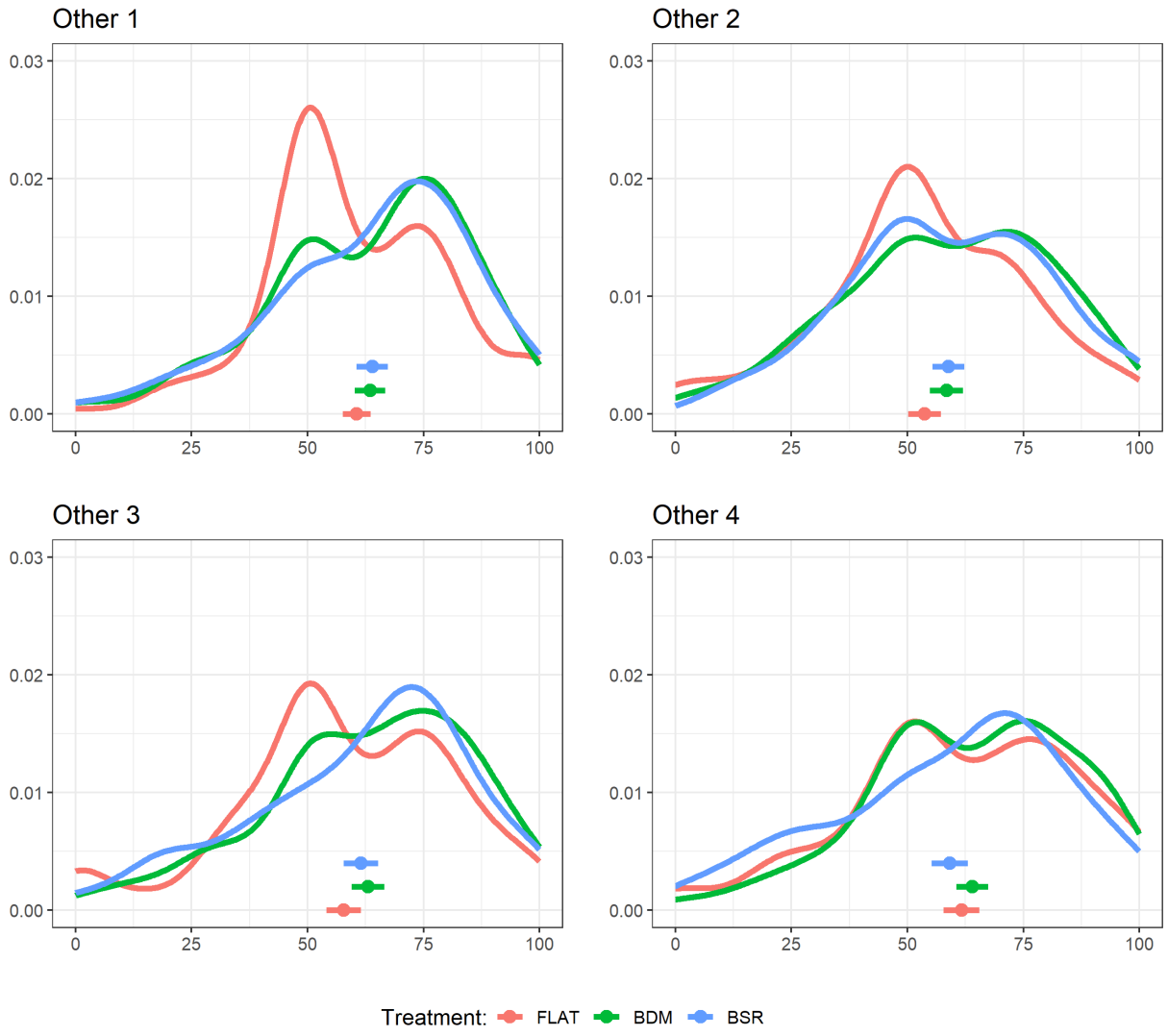


Figure 7: Kernel density plots of the distributions of social beliefs for FLAT (red line), BDM (green line) and BSR (blue line). The corresponding means and 95% confidence intervals for each method are depicted with dots and horizontal bars.

from mixed effects regressions with the absolute distance from 50% as the dependent variable for beliefs about all others (column 1), about Democrats (column 2), and about Republicans (column 3), using the same specifications as our analysis of first-order beliefs. In the second three columns, the dependent variable is an indicator for whether the respective belief is exactly equal to 50%. In the last three columns, the dependent variable is an indicator for whether the respondent's belief deviates from 50% in the "right direction", i.e. if the belief is greater (lower) than 50% when the underlying probability is also greater (lower) than 50%.

Three findings stand out from this regression analysis. First, when second-order beliefs pertain to all others (rather than a specific subgroup), the results suggest that the BDM and BSR mechanisms both move beliefs away from 50%, in terms of increasing distance (column 1) as well reducing the proportion of beliefs reported at exactly 50% (column 4). Second, BSR more consistently moves beliefs off the fence compared to BDM, as there are significantly fewer beliefs at exactly 50% for both partisan subgroups in BSR, whereas the coefficient for BDM is not significant for either subgroup. This difference is consistent also when comparing the coefficients of BSR and BDM (see F-test results). However, neither BSR nor BDM has any significant effect on the distance from 50%. Third, similar to the case of first-order beliefs, we find that the higher likelihood of departing from 50% reports in BSR may be beneficial as such beliefs seem to occur more often in the direction of the underlying probability in BSR than in FLAT or BDM, irrespective of the type of second-order belief. The differences between BSR and FLAT are always significant, while those between BSR and BDM are only significant in the case of beliefs about Democrats.

	Dependent variable:								
	Distance from 50%			Belief equal to 50%			Belief in right direction		
	Target group:								
	Other	Democrat	Republican	Other	Democrat	Republican	Other	Democrat	Republican
BDM	2.436	1.435	0.051	−0.083*	−0.017	−0.022	0.037	−0.023	0.052
	(1.275)	(1.271)	(1.306)	(0.032)	(0.029)	(0.028)	(0.027)	(0.027)	(0.027)
BSR	3.476**	2.125	0.931	−0.138***	−0.099***	−0.080**	0.090**	0.075**	0.076**
	(1.275)	(1.272)	(1.307)	(0.032)	(0.029)	(0.028)	(0.027)	(0.027)	(0.027)
Failed attention check	4.100*	4.578**	2.898	−0.105*	−0.075	−0.051	0.042	0.026	0.073*
	(1.685)	(1.681)	(1.727)	(0.043)	(0.038)	(0.037)	(0.036)	(0.035)	(0.036)
Low numeracy	4.601***	3.165**	2.035	−0.120***	−0.089***	−0.065**	0.058*	0.002	0.085***
	(1.065)	(1.062)	(1.091)	(0.027)	(0.024)	(0.024)	(0.023)	(0.022)	(0.023)
No college degree	−0.521	−0.121	−0.019	0.032	0.014	0.031	−0.014	0.014	−0.048*
	(1.064)	(1.061)	(1.090)	(0.027)	(0.024)	(0.024)	(0.023)	(0.022)	(0.023)
Constant	15.031***	18.645***	21.745***	0.313***	0.212***	0.162***	0.571***	0.646***	0.458***
	(1.262)	(1.265)	(1.295)	(0.032)	(0.029)	(0.028)	(0.031)	(0.031)	(0.031)
Observations	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880	1,880
Log Likelihood	−7,447.380	−7,499.461	−7,506.250	−610.111	−385.293	−348.952	−1,191.074	−1,170.409	−1,244.678
F Statistic (BDM=BSR)	0.671	0.297	0.457	2.927	8.168**	4.236*	3.817	13.464***	0.848

Note: Model includes item fixed effects (not reported) and subject random effects. The 'Other' columns refer to beliefs elicited in Part 3 of the experiment, where participants were asked about the accuracy of a randomly selected other participant in the study, without giving any further details regarding personal characteristics of this person. The 'Democrat' and 'Republican' columns refer to beliefs elicited in Parts 4 and 5 of the experiment, where participants were asked about the accuracy of a randomly selected other participant in the study that identifies with the Democratic or the Republican Party, respectively. *p<0.05; **p<0.01; ***p<0.001

Table 8: Centrality of second order beliefs

Finally, we find an in-group bias regarding judgments of other group's accuracy that persists across our treatments (see Figure 8). In particular, Democrats believe that Republicans are less accurate than fellow Democrats while Republicans believe that Democrats are less accurate than fellow Republicans (both differences are statistically significant irrespective of the treatment group). Neither of these beliefs are justified given that we did not observe any significant difference in accuracy across members of different parties: Democrats were 54% likely to accurately rate a statement as true while Republicans 52% ($p = 0.439$, Chi-squared test).

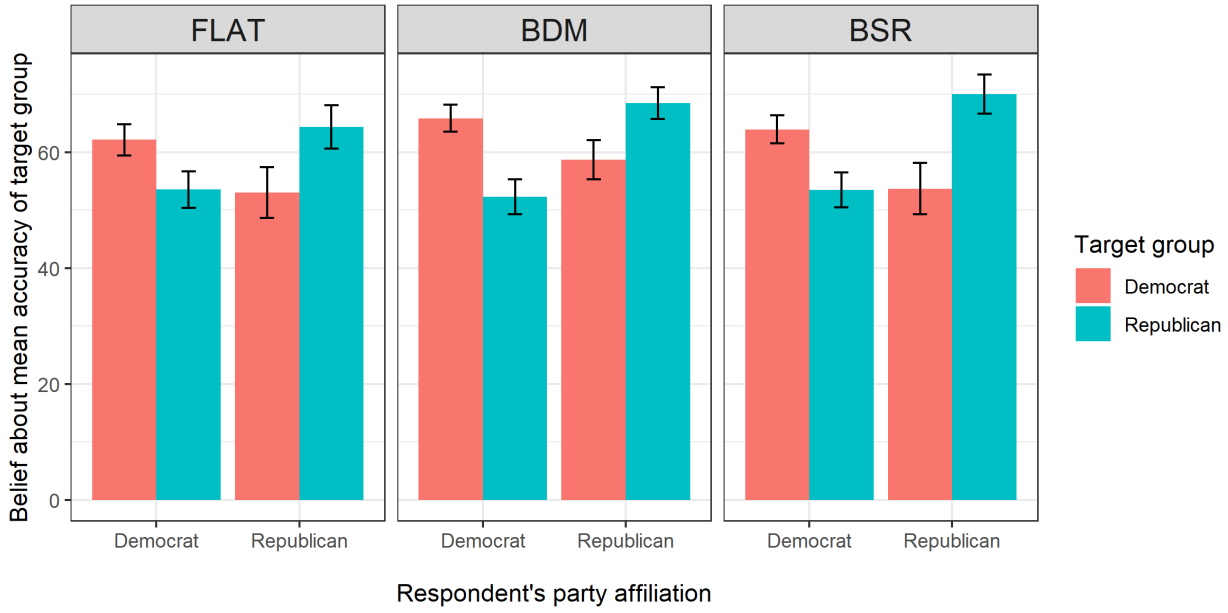


Figure 8: Average belief about the accuracy of Democrats and Republicans by respondent's party affiliation and across treatments. Error bars represent 95% confidence intervals.

4 Discussion

Our study complements the existing literature regarding belief elicitation mechanisms in several ways. First, whereas previous related evidence comes exclusively from controlled, lab environments with undergraduate student samples, we focus on the online research market which has seen an increasing usage. There are many variables in the online setting that differ

from the lab one which makes the generalization of conclusions drawn from lab studies problematic, a view echoed also in Charness, Gneezy and Rasocha (2021). A few studies have looked at how to incentivize participants for various tasks in online studies. The findings suggest that performance-based financial incentives manage to increase the quality of data obtained from online samples (Ho et al., 2015; Shaw, Horton and Chen, 2011) and that this difference can be mitigated by controlling for participants' intrinsic motivation (Vinogradov and Shadrina, 2013). These studies, however, focus on effort tasks with an objective performance benchmark. Although objective probabilities may exist for belief elicitation tasks as well, the goal of a belief elicitation incentives scheme is to elicit the subjective target irrespective of how this relates to the objective one. Nevertheless, intrinsic motivation is still relevant and we find that the incentive compatible mechanisms may actually increase participants' motivation to "discover" their beliefs - an idea we elaborate upon later in this section.

Second, we perform a horse-race between three popular elicitation mechanisms that have not yet been compared in the same setting. While both the BSR and the BDM methods have each been compared with other methods, we are not aware of any head-to-head comparisons as of yet. The BSR mechanism has only been compared with the Quadratic Scoring Rule (QSR), as it was a direct response to some of the theoretical caveats of the QSR. The results of those studies suggest that, for non-extreme events, the BSR manages to elicit beliefs closer to the true probabilities than the QSR does (Erkal, Gangadharan and Koh, 2020; Hossain and Okui, 2013). When comparing it with BDM and FLAT, we do not find the BSR significantly outperforming in terms of overall average accuracy in our setting. In fact, there is no significant accuracy difference between any of the three methods. As far as we know, only Massoni, Gajdos and Vergnaud (2014) have found that the BDM method significantly improves the accuracy of elicited beliefs when compared to an equivalent of our FLAT treatment (which they call introspection). Nevertheless, in that study, participants report their confidence about their own likelihood of having made the correct

decision, a type of belief we do not elicit in our study. Burfurd and Wilkening (2021) also compare BDM with unincentivized introspection and find no difference in accuracy between the two methods when pooling across all subjects within a treatment. However, they find that the BDM mechanism is less sensitive to the difficulty of the task than introspection, while cognitive ability is not a significant moderator of differences between measures. Their cognitive measure, however, consists of a type of Cognitive Reflection Test which may be only a subset of probabilistic reasoning skills captured by our education measure that we find to significantly moderate the difference between BDM and FLAT.

One limitation of our study is, however that it involved a particular and limited number of items for eliciting participants' beliefs. We chose to focus on items that are more likely to appear in online surveys deployed by practitioners. Therefore, despite selecting a relatively diverse set of possible questions (about abstract and real events varying in the level of familiarity and scope for motivated beliefs) these do not cover the universe of possible objects researchers may be interested in eliciting beliefs about. Consequently, our results may be less informative for items significantly different than those involved in this study, e.g., items regarding people's behavior in games, or items regarding which people's knowledge may vary more and subsequently, also their confidence in the corresponding beliefs.

With this in mind, we offer three main practical recommendations for researchers eliciting beliefs online. First, we suggest that the use of complicated belief elicitation mechanisms may not provide many added benefits if the elicited beliefs are about objective or familiar events (i.e., induced beliefs or events that people are likely to have thought about before), for which social incentives for belief misrepresentation is small. If participants have little uncertainty about these beliefs they will likely report them irrespective of the incentive scheme. Therefore, the gain in accuracy from using complex elicitation methods in this case, which we conjecture stems from an increased effort to discover the particular belief, may not outweigh the cost of additional time and effort necessary to properly implement such incentive compatible elicitation methods. This recommendation is also made by Manski (2004),

who provides a detailed and insightful analysis of the different practices across the social sciences (e.g. economics, psychology, sociology) and concludes that unincentivized belief elicitation for topics with personal significance can provide informative data. It is unclear though whether this would still be the case when strongly motivated beliefs are at play.

Second, if researchers are interested in beliefs that are more inherently subjective or in novel situations about which people may have given less sustained thought, we generally recommend using the Binarized Scoring Rule. Although we find that no method uniformly outperforms the others, the BSR emerges as the best overall elicitation method for our online environment. This surprised us, as our prior beliefs favored using the BDM in laboratory settings (Burdea and Woon, 2021; Woon and Kanthak, 2019). In terms of benefits, we note that both incentivized methods, BSR and BDM, do better in several respects than FLAT. Thus, incentives matter. While induced beliefs are equally accurate across all three methods for the most educated subjects, both the BSR and BDM increase accuracy for subjects without a college degree. Similarly, while both BSR and BDM elicit different distributions of first-order and second-order beliefs, we find that BSR more consistently elicits beliefs different from 50% than either BDM or FLAT—and in the right direction. Thus, while both BSR and BDM do better than FLAT in terms of accuracy, BSR does better in terms of pulling beliefs "off the fence" (more on this below). In terms of costs, both BSR and BDM take more time to implement and take more effort for subjects to understand than FLAT, but BSR is easier for subjects to understand than BDM and therefore has a clear advantage in terms of task comprehension.

Third, we recommend researchers measure covariates related to attention and cognition. This is because we find that accuracy, irrespective of the elicitation method, is increasing in attention, education, and numeracy. Of the covariates in our study, numeracy is most important, as it has the strongest relationship with the accuracy of induced beliefs. Indeed, the accuracy of low numeracy subjects is half that of high numeracy subjects. Similar to attentiveness in survey experiments (Berinsky, Margolis and Sances, 2014), numeracy

can serve as a useful control or conditioning variable. Future research could explore whether other measures of cognition, such as the Cognitive Reflection Test, are also related to the characteristics of elicited beliefs. The laboratory study of Burfurd and Wilkening (2021) suggests a significant correlation between subjects' performance on such tasks and the quality of elicited beliefs, but no significant differences between the BDM and the introspection mechanisms. It would be worthwhile exploring whether these results extend to online samples and to the BSR mechanism. Furthermore, if a study's sample is likely to be composed of a majority of low-numeracy individuals, then various techniques such as greater reliance on visual aids can be used to enhance comprehension (Delavande and Rohwedder, 2008). Equally relevant is the question of how best to measure such covariates online given the results of Wolff (2019) suggesting that different methods (e.g. incentives' size, timing and framing) significantly influence the quality of this data in laboratory experiments.

It would also be worthwhile for future research to investigate the marginal costs and benefits of further adapting instructions and procedures for online environments. In our study, we focused on comparing three methods of online belief elicitation while implementing and holding constant important procedural features that are standard in laboratory studies. Specifically, we provided complete and transparent instructions, including full descriptions of the incentive mechanisms, and we checked and encouraged comprehension using instruction quizzes with feedback. Would additional training in interpreting probabilities or proving incentive compatibility be worth the effort? Could the procedures be streamlined without sacrificing accuracy and comprehension? On the one hand, Burfurd and Wilkening (2018) find that eliminating the comprehension quiz leads to worse performance of the BDM mechanism. On the other hand, Danz, Vesterlund and Wilson (2020) show that simplifying the instructions by eliminating information about marginal incentives improves the accuracy of the BSR method, reducing the frequency of 50% beliefs (the "pull-to-center" effect). However, we did not directly compare online and in-person laboratory environments, and such comparisons would also be worthwhile endeavors for future research.

We conclude with thoughts about our study's implications for belief elicitation in general. Much of the literature on belief elicitation emphasizes the problem of ensuring that agents report their beliefs "truthfully" (e.g., Offerman et al., 2009). In other words, this literature tends to conceive of belief elicitation in purely mechanism design terms, in which the goal is to ensure that agents, solving a maximization problem, report their private information.

Our results suggest that designing effective and reliable methods for eliciting beliefs should also take into account "cognitive production" (Camerer and Hogarth, 1999; Rydval and Ortmann, 2004). Indeed, that cognitive measures such as numeracy are strongly related to belief accuracy is a reminder that cognitive sophistication ("cognitive capital") is necessary for complex incentive-compatible procedures to work. Even if agents are sufficiently sophisticated, a related problem is whether they are sufficiently motivated to put in the cognitive effort ("cognitive labor") to solve the maximization problem (even implicitly or unconsciously), or to form quantitative judgments of uncertainty in the first place.

We suspect that if each mechanism induces a different amount of cognitive effort, this could account for the differences we observe between mechanisms in the proportions of 50% beliefs. Importantly, to make sense of such an explanation, we first acknowledge that a 50% belief (i.e. reporting that one is "totally uncertain" about a statement) might mean something to subjects other than "it is equally likely that the statement is true or false" (the probabilistic interpretation of the midpoint between "the statement is likely to be true" and "the statement is likely to be false"). Instead, they might think of the 50% belief as conveying "I have absolutely no idea how to answer this question" or "I don't know *because I haven't thought about it*" (e.g., Fischhoff and Bruine De Bruin, 1999), which is related to Enke and Graeber's (2021) notion of cognitive uncertainty.

We conjecture that these results might be explained by a cognitive model in which subjects don't have access to their own beliefs unless they exert a sufficient amount of cognitive effort, and if they are able to access their beliefs that they report them accurately.

This can explain why we observe the most 50% beliefs in FLAT, where there is no monetary gain from spending additional cognitive effort trying to figure out what one believes. Such a cognitive model can also explain why there are fewer 50% beliefs in BDM than in FLAT, and the fewest in BSR. This is because subjects perceive it to be worthwhile spending cognitive effort in return for the increased monetary reward from accessing and reporting their beliefs. And because BSR is easier to comprehend than BDM, the perceived marginal cost of effort is lower in BSR, thereby inducing the most cognitive effort and the lowest frequency of 50% beliefs. It is plausible that under this model, larger incentives would move more subjects' beliefs off the fence. However, this may depend on individual characteristics. For example, Armantier and Treich (2013) find in the case of the Quadratic Scoring Rule that steeper incentives reduce the likelihood that beliefs are centrally biased but only for people exhibiting decreasing relative risk aversion.

Alternatively, though perhaps less straightforwardly, one could increase the cognitive sophistication of subjects to achieve a similar outcome. Even though this is fixed in the short-term, as Camerer and Hogarth (1999) and Rydval and Ortmann (2004) suggest, it can increase through learning. Since the latter option is more time-intensive, it may be more appropriate for a laboratory environment than an online one. Understanding the channels through which (different) incentives can affect the quality of elicited beliefs and how these interact with the elicitation environment and individual characteristics is worthy of further investigation.

References

Akerlof, George A. 1982. "Labor contracts as partial gift exchange." *The Quarterly Journal of Economics* 97(4):543–569.

Allen, Franklin. 1987. "Discovering personal probabilities when utility functions are unknown." *Management Science* 33(4):542–544.

Ambuehl, Sandro. 2017. "An offer you can't refuse? Incentives change how we inform ourselves and what we believe." *CESifo Working Papers* (6296).

Arechar, Antonio A, Simon Gächter and Lucas Molleman. 2018. "Conducting interactive experiments online." *Experimental Economics* 21(1):99–131.

Armantier, Olivier, Giorgio Topa, Wilbert Van der Klaauw and Basit Zafar. 2017. "An overview of the survey of consumer expectations." *Economic Policy Review* (23-2):51–72.

Armantier, Olivier and Nicolas Treich. 2013. "Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging." *European Economic Review* 62:17–40.

Baillon, Aurélien and Han Bleichrodt. 2015. "Testing ambiguity models through the measurement of probabilities for gains and losses." *American Economic Journal: Microeconomics* 7(2):77–100.

Becker, Gordon M, Morris H DeGroot and Jacob Marschak. 1964. "Measuring utility by a single-response sequential method." *Behavioral Science* 9(3):226–232.

Belfield, Chris, Teodora Boneva, Christopher Rauh and Jonathan Shaw. 2020. "What drives enrolment gaps in further education? the role of beliefs in sequential schooling decisions." *Economica* 87(346):490–529.

Berinsky, Adam J, Gregory A Huber and Gabriel S Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Berinsky, Adam J, Michele F Margolis and Michael 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.

Brier, Glenn W et al. 1950. "Verification of forecasts expressed in terms of probability." *Monthly weather review* 78(1):1–3.

Burdea, Valeria and Jonathan Woon. 2021. "Getting it Right: Communication, Voting, and Collective Truth Finding." Working paper.

Burfurd, Ingrid and Tom Wilkening. 2018. "Experimental guidance for eliciting beliefs with the Stochastic Becker–DeGroot–Marschak mechanism." *Journal of the Economic Science Association* 4(1):15–28.

Burfurd, Ingrid and Tom Wilkening. 2021. "Cognitive heterogeneity and complex belief elicitation." *Experimental Economics* pp. 1–36.

Camerer, Colin F and Robin M Hogarth. 1999. "The effects of financial incentives in experiments: A review and capital-labor-production framework." *Journal of Risk and Uncertainty* 19(1-3):7–42.

Chandler, Jesse, Pam Mueller and Gabriele Paolacci. 2014. "Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers." *Behavior Research Methods* 46(1):112–130.

Charness, Gary, Uri Gneezy and Vlastimil Rasocha. 2021. "Experimental methods: Eliciting beliefs." *Journal of Economic Behavior & Organization* 189:234–256.

Clifford, Scott and Jennifer Jerit. 2014. "Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies." *Journal of Experimental Political Science* 1(2):120–131.

Coffman, Katherine B, Christine L Exley and Muriel Niederle. 2021. "The role of beliefs in driving gender discrimination." *Management Science* .

Coffman, Katherine, Manuela Collis and Leena Kulkarni. 2019. "Stereotypes and belief updating." *Harvard Business School, Working Paper Series* (19-068).

Coutts, Alexander. 2019. "Testing models of belief bias: An experiment." *Games and Economic Behavior* 113:549–565.

Crosetto, Paolo, Antonio Filippin, Peter Katuščák and John Smith. 2020. "Central tendency bias in belief elicitation." *Journal of Economic Psychology* p. 102273.

Crump, Matthew JC, John V McDonnell and Todd M Gureckis. 2013. "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research." *PloS ONE* 8(3):e57410.

Danz, David, Lise Vesterlund and Alistair J Wilson. 2020. "Belief elicitation: Limiting truth telling with information on incentives." *National Bureau of Economic Research, Working Paper Series* (27327).

Delavande, Adeline and Susann Rohwedder. 2008. "Eliciting subjective probabilities in Internet surveys." *Public Opinion Quarterly* 72(5):866–891.

DuCharme, Wesley M and Michael L Donnell. 1973. "Intrasubject comparison of four response modes for "subjective probability" assessment." *Organizational Behavior and Human Performance* 10(1):108–117.

Enke, Benjamin, Frederik Schwerter and Florian Zimmermann. 2020. "Associative Memory and Belief Formation." *National Bureau of Economic Research, Working Paper Series* (26664).

Enke, Benjamin and Thomas Graeber. 2021. "Cognitive uncertainty." *National Bureau of Economic Research, Working Paper Series* (26518).

Erkal, Nisvan, Lata Gangadharan and Boon Han Koh. 2020. "Replication: Belief elicitation with quadratic and binarized scoring rules." *Journal of Economic Psychology* 81:102315.

Fehr, Ernst, Georg Kirchsteiger and Arno Riedl. 1993. "Does fairness prevent market clearing? An experimental investigation." *The Quarterly Journal of Economics* 108(2):437–459.

Fischhoff, Baruch and Wändi Bruine De Bruin. 1999. "Fifty–fifty= 50%?" *Journal of Behavioral Decision Making* 12(2):149–163.

Fort, Karën, Gilles Adda and K Bretonnel Cohen. 2011. "Amazon mechanical turk: Gold mine or coal mine?" *Computational Linguistics* 37(2):413–420.

Goodman, Joseph K, Cynthia E Cryder and Amar Cheema. 2013. "Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples." *Journal of Behavioral Decision Making* 26(3):213–224.

Grether, David M. 1981. "Financial incentive effects and individual decision-making." California Institute of Technology(Working Paper 401).

Grether, David M. 1992. "Testing Bayes rule and the representativeness heuristic: Some experimental evidence." *Journal of Economic Behavior & Organization* 17(1):31–57.

Hao, Li and Daniel Houser. 2012. "Belief elicitation in the presence of naïve respondents: An experimental study." *Journal of Risk and Uncertainty* 44(2):161–180.

Harrison, Glenn W, Jimmy Martínez-Correa and J Todd Swarthout. 2013. "Inducing risk neutral preferences with binary lotteries: A reconsideration." *Journal of Economic Behavior & Organization* 94:145–159.

Healy, Paul J. 2018. "Explaining the BDM—Or any random binary choice elicitation mechanism—To Subjects." Working Paper.

Hergueux, Jérôme and Nicolas Jacquemet. 2015. "Social preferences in the online laboratory: a randomized experiment." *Experimental Economics* 18(2):251–283.

Hill, Seth J. 2017. "Learning together slowly: Bayesian learning about political facts." *The Journal of Politics* 79(4):1403–1418.

Ho, Chien-Ju, Aleksandrs Slivkins, Siddharth Suri and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. pp. 419–429.

Holt, Charles A. 2007. *Markets, games, & strategic behavior*. Pearson Addison Wesley Boston.

Holt, Charles A and Angela M Smith. 2009. "An update on Bayesian updating." *Journal of Economic Behavior & Organization* 69(2):125–134.

Holt, Charles A and Angela M Smith. 2016. "Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes." *American Economic Journal: Microeconomics* 8(1):110–39.

Horton, John J, David G Rand and Richard J Zeckhauser. 2011. "The online laboratory: Conducting experiments in a real labor market." *Experimental Economics* 14(3):399–425.

Hossain, Tanjim and Ryo Okui. 2013. "The binarized scoring rule." *Review of Economic Studies* 80(3):984–1001.

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

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

Martinangeli, Andrea FM. 2021. "Do what (you think) the rich will do: Inequality and belief heterogeneity in public good provision." *Journal of Economic Psychology* 83:102364.

Mason, Winter and Siddharth Suri. 2012. "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior Research Methods* 44(1):1–23.

Massoni, Sébastien, Thibault Gajdos and Jean-Christophe Vergnaud. 2014. "Confidence measurement in the light of signal detection theory." *Frontiers in Psychology* 5:1455.

McKelvey, Richard D and Talbot Page. 1990. "Public and private information: An experimental study of information pooling." *Econometrica* pp. 1321–1339.

Mobius, Markus M, Muriel Niederle, Paul Niehaus and Tanya S Rosenblat. 2011. "Managing self-confidence: Theory and experimental evidence." *National Bureau of Economic Research, Working Paper Series* (17014).

Offerman, Theo, Joep Sonnemans, Gijs Van de Kuilen and Peter P Wakker. 2009. "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes." *The Review of Economic Studies* 76(4):1461–1489.

Oppenheimer, Daniel M, Tom Meyvis and Nicolas Davidenko. 2009. "Instructional manipulation checks: Detecting satisficing to increase statistical power." *Journal of Experimental Social Psychology* 45(4):867–872.

Paas, Leonard J and Meike Morren. 2018. "Please do not answer if you are reading this: Respondent attention in online panels." *Marketing Letters* 29(1):13–21.

Paolacci, Gabriele, Jesse Chandler and Panagiotis G Ipeirotis. 2010. "Running experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5(5):411–419.

Peeters, Ronald and Marc Vorsatz. 2021. "Simple guilt and cooperation." *Journal of Economic Psychology* 82:102347.

Peters, Ellen, Daniel Västfjäll, Paul Slovic, CK Mertz, Ketti Mazzocco and Stephan Dickert. 2006. "Numeracy and decision making." *Psychological Science* 17(5):407–413.

Roth, Christopher and Johannes Wohlfart. 2020. "How do expectations about the macroeconomy affect personal expectations and behavior?" *Review of Economics and Statistics* 102(4):731–748.

Rydval, Ondrej and Andreas Ortmann. 2004. "How financial incentives and cognitive abilities affect task performance in laboratory settings: An illustration." *Economics Letters* 85(3):315–320.

Savage, Leonard J. 1971. "Elicitation of personal probabilities and expectations." *Journal of the American Statistical Association* 66(336):783–801.

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

Schotter, Andrew and Isabel Trevino. 2014. "Belief elicitation in the laboratory." *Annual Review of Economics* 6(1):103–128.

Selten, Reinhard. 1998. "Axiomatic characterization of the quadratic scoring rule." *Experimental Economics* 1(1):43–61.

Shaw, Aaron D, John J Horton and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. pp. 275–284.

Silberman, M Six, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani and Andrew Zaldivar. 2018. "Responsible research with crowds: pay crowdworkers at least minimum wage." *Communications of the ACM* 61(3):39–41.

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

Vinogradov, Dmitri and Elena Shadrina. 2013. "Non-monetary incentives in online experiments." *Economics Letters* 119(3):306–310.

Williamson, Vanessa. 2016. "On the ethics of crowdsourced research." *PS: Political Science & Politics* 49(1):77–81.

Wilson, Alistair and Emanuel Vespa. 2017. "Paired-uniform scoring: Implementing a binarized scoring rule with non-mathematical language." Working paper.

Wolff, Irenaeus. 2019. "The reliability of questionnaires in laboratory experiments: What can we do?" *Journal of Economic Psychology* 74:102197.

Woon, Jonathan and Kristin Kanthak. 2019. "Elections, ability, and candidate honesty." *Journal of Economic Behavior & Organization* 157:735–753.

Online Belief Elicitation Methods

Valeria Burdea Jonathan Woon

Online Appendix

Appendix A Additional analysis

Item no.	Question	Correct answer
1	Imagine that we roll a fair, six-sided die with the numbers 1 through 6 on its sides. What is the likelihood that the die will come up even?	50%
2	Consider a standard deck of 52 cards (with 13 hearts, 13 diamonds, 13 spades, and 13 clubs). Imagine that we shuffle this deck of cards and draw one card. What is the likelihood that the card is a spade?	25%
3	Imagine that we put 100 balls (20 red and 80 green) in a bag and draw one without looking. What is the likelihood that the ball we draw is green?	80%
4	Imagine that we put 33 pink balls and 17 yellow balls in a bag and draw one without looking. What is the likelihood that the ball we draw is yellow?	34%
5	Imagine that we flip a fair coin twice. The coin has heads on one side and tails on the other. What is the likelihood that at least one coin flip results in heads?	75%
6	Imagine you are flipping a fair coin (with heads on one side and tails on the other) and after eight flips you observe the following result: tails - tails - tails - heads - tails - heads - heads - heads. What is the likelihood that the next flip is tails?	50%

Table A1: Numeracy questions

Item no.	Statement	True probability
1	We will roll a fair, 6-sided die behind the scenes, with the numbers 1 through 6 on its sides. Now, consider the following statement: "The outcome of the die roll is a number less than or equal to 6."	100%
2	We will roll another fair, 6-sided die, separately, behind the scenes. The sides of this die are also numbered from 1 to 6. Now, consider the following statement: "The outcome of the die roll is a number equal to 0."	0%
3	We will roll another fair, 6-sided die, separately, behind the scenes. The sides of this die are also numbered from 1 to 6. Now, consider the following statement: "The outcome of the die roll is an odd number"	50%
4	We will shuffle a deck of cards behind the scenes and draw the top card. This deck is a standard deck of 52 cards (with 13 hearts, 13 diamonds, 13 spades, and 13 clubs). Now consider the following statement: "The suit of the card that was drawn is hearts."	25%
5	We will put 10 poker chips in a bag and draw one without looking. The bag has 8 white chips and 2 red chips. Now, consider the following statement: "The color of the chip that was drawn is white."	80%

Table A2: List of statements for induced probabilities (calibration)

	FLAT (N=155)	BDM (N=157)	BSR (N=158)	Total (N=470)	p value
Age					0.771
18 - 24	15 (9.7%)	13 (8.3%)	10 (6.3%)	38 (8.1%)	
25 - 34	76 (49.0%)	79 (50.3%)	77 (48.7%)	232 (49.4%)	
35 - 44	32 (20.6%)	37 (23.6%)	34 (21.5%)	103 (21.9%)	
45 - 54	15 (9.7%)	14 (8.9%)	21 (13.3%)	50 (10.6%)	
55 - 64	16 (10.3%)	11 (7.0%)	11 (7.0%)	38 (8.1%)	
65 - 74	1 (0.6%)	3 (1.9%)	4 (2.5%)	8 (1.7%)	
85 or older	0 (0.0%)	0 (0.0%)	1 (0.6%)	1 (0.2%)	
Gender					0.254
Female	70 (45.2%)	57 (36.3%)	61 (38.6%)	188 (40.0%)	
Male	85 (54.8%)	100 (63.7%)	97 (61.4%)	282 (60.0%)	
Race: White/Caucasian					0.154
FALSE	30 (19.4%)	19 (12.1%)	30 (19.0%)	79 (16.8%)	
TRUE	125 (80.6%)	138 (87.9%)	128 (81.0%)	391 (83.2%)	
**Race: African American **					0.354
FALSE	138 (89.0%)	147 (93.6%)	144 (91.1%)	429 (91.3%)	
TRUE	17 (11.0%)	10 (6.4%)	14 (8.9%)	41 (8.7%)	
Race: Hispanic					0.598
FALSE	144 (92.9%)	150 (95.5%)	148 (93.7%)	442 (94.0%)	
TRUE	11 (7.1%)	7 (4.5%)	10 (6.3%)	28 (6.0%)	
Race: Asian or Pacific Islander					0.692
FALSE	148 (95.5%)	150 (95.5%)	148 (93.7%)	446 (94.9%)	
TRUE	7 (4.5%)	7 (4.5%)	10 (6.3%)	24 (5.1%)	
Race: Native American					0.367
FALSE	154 (99.4%)	153 (97.5%)	156 (98.7%)	463 (98.5%)	
TRUE	1 (0.6%)	4 (2.5%)	2 (1.3%)	7 (1.5%)	
Race: Other					0.361
FALSE	154 (99.4%)	157 (100.0%)	158 (100.0%)	469 (99.8%)	
TRUE	1 (0.6%)	0 (0.0%)	0 (0.0%)	1 (0.2%)	
Education					0.407
2 year college degree (Associate)	26 (16.8%)	18 (11.5%)	15 (9.5%)	59 (12.6%)	
4 year college degree (Bachelor)	68 (43.9%)	72 (45.9%)	65 (41.1%)	205 (43.6%)	
High School / GED	19 (12.3%)	22 (14.0%)	27 (17.1%)	68 (14.5%)	
Less than High School	0 (0.0%)	1 (0.6%)	1 (0.6%)	2 (0.4%)	
Post-graduate degree (Professional, Masters, Doctorate)	7 (4.5%)	12 (7.6%)	17 (10.8%)	36 (7.7%)	
Some college	35 (22.6%)	32 (20.4%)	33 (20.9%)	100 (21.3%)	
Party affiliation					0.173
Democrat	77 (49.7%)	79 (50.3%)	75 (47.5%)	231 (49.1%)	
Republican	32 (20.6%)	49 (31.2%)	39 (24.7%)	120 (25.5%)	
Independent	44 (28.4%)	27 (17.2%)	40 (25.3%)	111 (23.6%)	
Other (please specify)	2 (1.3%)	2 (1.3%)	4 (2.5%)	8 (1.7%)	

Note: The p-values are from Chi-square tests for equality of distributions across treatments.

Table A3: Sample summary statistics

	FLAT (N=155)	BDM (N=157)	BSR (N=158)	Total (N=470)	p value
Passed attention check					0.733
FALSE	20 (12.9%)	19 (12.1%)	16 (10.1%)	55 (11.7%)	
TRUE	135 (87.1%)	138 (87.9%)	142 (89.9%)	415 (88.3%)	
Low numeracy					0.296
FALSE	79 (51.0%)	81 (51.6%)	93 (58.9%)	253 (53.8%)	
TRUE	76 (49.0%)	76 (48.4%)	65 (41.1%)	217 (46.2%)	

Note: The p-values are from Chi-square tests for equality of distributions across treatments.

Table A4: Attention and numeracy - summary statistics

	Dependent variable:		
	Difficulty	Effort	Confidence
	(1)	(2)	(3)
BDM	1.173***	0.926***	0.259
	(0.130)	(0.129)	(0.139)
BSR	0.530***	0.598***	0.263*
	(0.134)	(0.132)	(0.137)
Constant	2.426***	2.832***	3.168***
	(0.094)	(0.099)	(0.109)
Observations	470	470	470
R^2	0.145	0.104	0.011
Adjusted R^2	0.141	0.100	0.007
Residual Std. Error (df = 467)	1.167	1.127	1.168
F Statistic (BDM=BSR)	24.215***	7.007**	0.0009

Note: Robust standard errors in parentheses.

*p<0.05; **p<0.01; ***p<0.001

Table A5: Task comprehension measures

	Dependent variable:
	Time (seconds)
BDM	86.428
	(48.816)
BSR	113.183*
	(48.931)
Constant	793.368***
	(32.175)
Observations	470
R^2	0.012
Adjusted R^2	0.008
Residual Std. Error	432.203 (df = 467)
F Statistic (BDM=BSR)	0.281

Note: Robust standard errors in parentheses. *p<0.05; **p<0.01; ***p<0.001

Table A6: Total study duration

	Dependent variable:
	Time (seconds)
BDM	−0.857
	(0.518)
BSR	0.573
	(0.517)
Constant	9.142***
	(1.369)
Observations	9,870
Log Likelihood	−39,327.590
F Statistic (BDM=BSR)	7.706**

Note: Model includes part, item and subject random effects.
*p<0.05; **p<0.01; ***p<0.001

Table A7: Belief elicitation duration

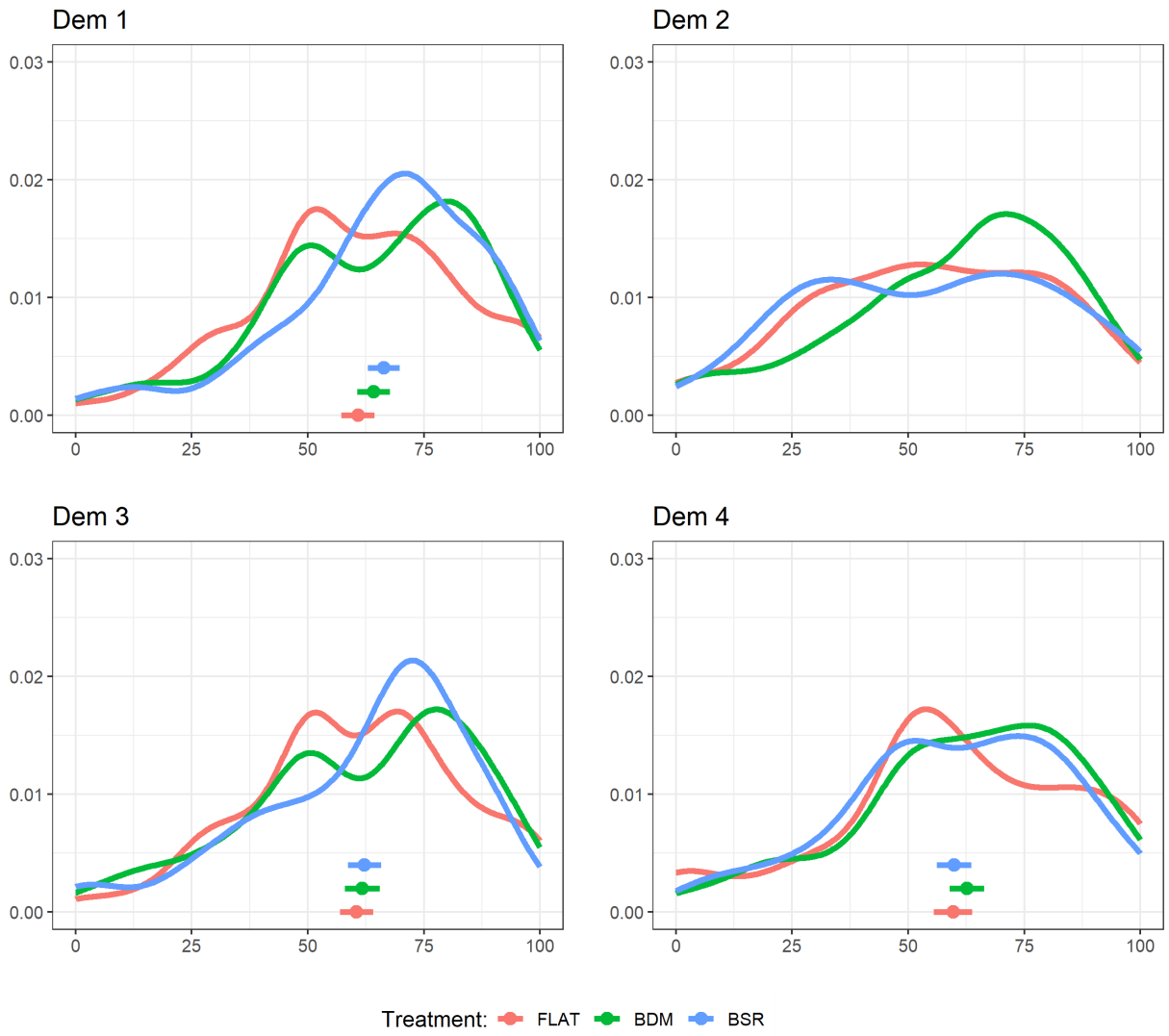


Figure A1: Kernel density plots of the distributions of beliefs about the accuracy of Democrats for FLAT (red line), BDM (green line) and BSR (blue line). The corresponding means and 95% confidence intervals for each method are depicted with dots and horizontal bars.

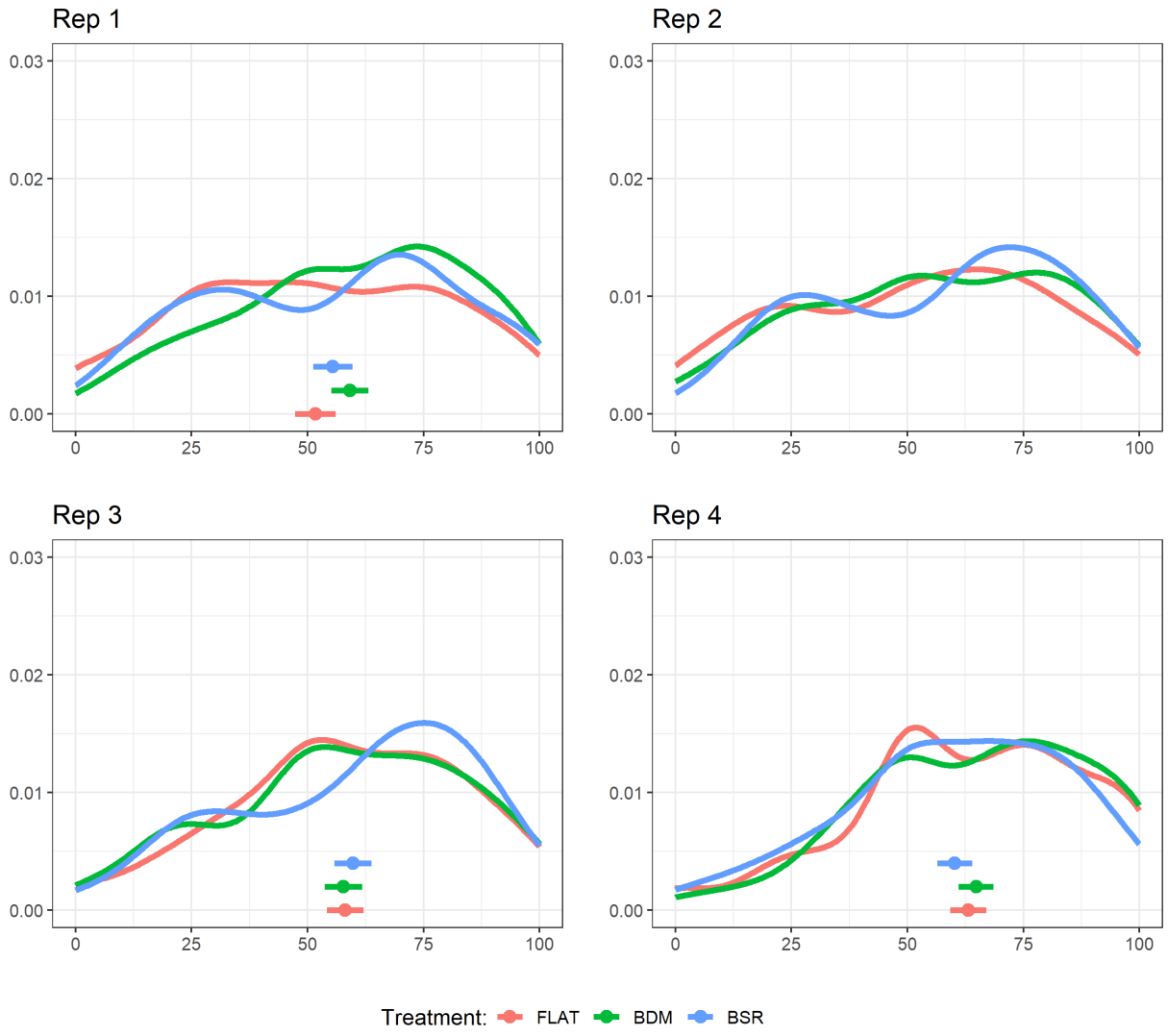


Figure A2: Kernel density plots of the distributions of beliefs about the accuracy of Republicans for FLAT (red line), BDM (green line) and BSR (blue line). The corresponding means and 95% confidence intervals for each method are depicted with dots and horizontal bars.

Appendix B Experimental instructions

Note: Section headings that are in bold and underlined are not shown to participants.

CONSENT

Welcome to this study! This study is part of a research project about beliefs. It is expected to take approximately 15-20 minutes to complete. All participants must be 18 years of age or older and live in the United States. There are no foreseeable risks associated with this project, nor are there any direct benefits to you. Your participation is completely voluntary. During the study we will ask you some questions about your background and there may also be questions to check that you are paying attention. You will earn $1.50 for successfully completing the study. It is also possible for some participants to earn an additional bonus of up to $2.00. We will not ask for your name or any other personally identifiable information. All responses are confidential and the confidentiality of your records will be maintained by using only codes to identify your responses. Your participation is voluntary, and you may stop completing the survey at any time. However, we are only able to pay you if you complete the survey. The study is being conducted by Jonathan Woon and his research associates in the Department of Political Science at the University of Pittsburgh. If you have any questions about the study, you may send an email to woon@pitt.edu.

- ○ I have read the above and consent to take part in this study
- ○ I do not wish to participate

INDIVIDUAL CHARACTERISTICS

What is your gender?
[Male; Female; Other]

What is your age?
[Under 18; 18 – 24; 25 – 34; 35 – 44; 45 – 54; 55 – 64; 65 – 74; 75 – 84; 85 or older]

What is your race/ethnicity? (check all that apply)
[White/Caucasian; African American; Hispanic; Asian or Pacific Islander; Native American; Other]

What is the highest level of education you have completed?
[Less than High School; High School / GED; Some college; 2 year college degree (Associate); 4 year college degree (Bachelor); Post-graduate degree (Professional, Masters, Doctorate)]

Generally speaking, do you usually think of yourself as a Republican, Democrat, or Independent?
[Republican; Democrat; Independent; Other]

(If Independent or Other) Do you usually think of yourself as closer to the Republican Party or the Democratic Party?
[Republican Party; Democratic Party; Neither; Not sure]

(If Republican) Would you call yourself a strong Republican or not so strong Republican?
[Strong Republican; Not so strong Republican]

(If Democrat) Would you call yourself a strong Democrat or not so strong Democrat?

[Strong Democrat; Not so strong Democrat]

Thinking of politics these days, how would you describe your own political viewpoint?
[Very Liberal; Liberal; Slightly Liberal; Moderate / Middle of the Road; Slightly Conservative; Conservative; Very Conservative]

Some people think the government should provide fewer services, even in areas such as health and education, in order to reduce spending. To demonstrate that you've read this much, just go ahead and select the numbers one and four no matter what your own views are.

Where would you place YOURSELF on this scale?

[1 - Fewer services; 2; 3; 4; More services (5)]

Before you continue, please click below to indicate that you are not a robot.

KNOWLEDGE QUIZ

First, we are going to ask for your opinion about whether various statements about the world are TRUE or FALSE. We will ask you about 4 statements. Try to be as accurate as possible.

Quiz1. Consider the following statement:

> More than half of unauthorized immigrants residing in the United States in 2016 had been living in the country for 10 years or more.

Do you think this statement is TRUE or FALSE?

Quiz2. Consider the following statement:

> From 2009, when President Obama took office, to 2012, median household income adjusted for inflation in the United States fell by more than 4 percent.

Do you think this statement is TRUE or FALSE?

Quiz3. Consider the following statement:

> More people in the United States work in the coal industry than in the solar industry.

Do you think this statement is TRUE or FALSE?

Quiz4. Consider the following statement:

> West Virginia was part of the Confederacy during the American Civil War.

Do you think this statement is TRUE or FALSE?

NUMERACY

Next, we will ask you a series of questions about chance events. Again, please try to be as accurate as possible.

Probability1. Imagine that we roll a fair, six-sided die with the numbers 1 through 6 on its sides. What is the likelihood that the die will come up even?

Drag the slider to indicate the percentage chance that the die will come up even. When you drag the slider, a number will appear that indicates your answer.

Probability2. Consider a standard deck of 52 cards (with 13 hearts, 13 diamonds, 13 spades, and 13 clubs). Imagine that we shuffle this deck of cards and draw one card. What is the likelihood that the card is a spade?

Drag the slider to indicate the percentage chance that the card would be a spade.

Probability3. Imagine that we put 100 balls (20 red and 80 green) in a bag and draw one without looking. What is the likelihood that the ball we draw is green?

Drag the slider to indicate the percentage chance that the ball is green.

Probability4. Imagine that we put 33 pink balls and 17 yellow balls in a bag and draw one without looking. What is the likelihood that the ball we draw is yellow?

Drag the slider to indicate the percentage chance that the ball is yellow.

Probability5. Imagine that we flip a fair coin twice. The coin has heads on one side and tails on the other. What is the likelihood that at least one coin flip results in heads?

Drag the slider to indicate the percentage chance that at least one coin flip results in heads.

Probability6. Imagine you are flipping a fair coin (with heads on one side and tails on the other) and after eight flips you observe the following result: tails - tails - tails - heads - tails - heads - heads - heads. What is the likelihood that the next flip is tails?

Drag the slider to indicate the percentage chance that the next flip is tails.

GENERAL INSTRUCTIONS

General description of parts and bonus payment

The rest of the study has 5 parts, and you will have an opportunity to earn a bonus for each part. After you complete the study, we will add the bonuses together from all parts to determine your final bonus for the study. These calculations will take place behind the scenes and you will not find out the amount of the bonus until it is paid.

Next, we will provide detailed instructions about your task in each part. Pay attention and follow the instructions closely as these will describe how you will earn money and how your earnings will depend on the choices that you make.

Reporting beliefs

In each part, we will ask you for your beliefs about the likelihood that various statements, facts, or events are TRUE. We will ask you about different kinds of statements, but the way that we will ask for your beliefs will be the same.

Specifically, we would like you to report your beliefs in terms of a number B from 0 to 100, which you will indicate by dragging a slider on the screen. You should think of the number B as representing the **percentage chance that the statement is TRUE**. The following table summarizes what each reported number indicates about your belief:

Your belief (B)	**This means:**
100	You think the statement is certainly TRUE, beyond any doubt
51-99	You think the statement is likely to be TRUE (higher numbers indicate greater certainty it is TRUE)
50	You are totally uncertain
1-49	You think the statement is likely to be FALSE (lower numbers indicate greater certainty it is FALSE)
0	You think the statement is certainly FALSE, beyond any doubt

Next, we will ask a series of questions to check that you understand what different values of B mean.

Belief_check1. If you select the number 100, this means:

[You think the statement is certainly TRUE; You think the statement is likely to be TRUE; You are totally uncertain; You think the statement is likely to be FALSE; You think the statement is certainly FALSE]

Belief_check2. If you select the number 50, this means:
[You think the statement is certainly TRUE; You think the statement is likely to be TRUE; You are totally uncertain; You think the statement is likely to be FALSE; You think the statement is certainly FALSE]

Belief_check3. If you select the number 0, this means:

[You think the statement is certainly TRUE; You think the statement is likely to be TRUE; You are totally uncertain; You think the statement is likely to be FALSE; You think the statement is certainly FALSE]

Belief_check4. Consider two different beliefs, 88 and 54. Select ALL of the meanings of these beliefs that are correct:
[88 means greater certainty the statement is TRUE than 54; 88 means less certainty the statement is TRUE than 54; 88 means the statement is likely to be TRUE; 54 means the statement is likely to be FALSE; 88 means the statement is likely to be FALSE]

Belief_check5. Consider two different beliefs, 19 and 74. Select ALL of the meanings of these beliefs that are correct:

[74 means greater certainty the statement is TRUE than 19; 74 means less certainty the statement is TRUE than 19; 74 means the statement is likely to be TRUE; 19 means the statement is likely to be FALSE; 19 means the statement is likely to be TRUE]

FLAT INSTRUCTIONS

Determining your bonus

You will earn a bonus of $0.20 for completing each part. In each part, you will report your beliefs for different statements.

You should report your beliefs as accurately as possible. That is, there is nothing to gain by stating a number that differs from what you actually believe.

We emphasize that this is a NO DECEPTION study. Your bonus is determined as described above.

Next, we will ask a series of questions to check your understanding of these instructions. You must answer all of the questions to advance to the task.

Flat_quiz1. If you report a belief B equal to 100, which of the following is correct?
[You win the bonus only if the statement is indeed TRUE.; You win the bonus only if the statement is indeed FALSE.; You will be entered into a lottery if the statement is indeed TRUE.; You win the bonus irrespective of the truth of the statement.]

Flat_quiz2. Suppose you state a belief B equal to 28, which of the following is correct?
[You win the bonus only if the statement is indeed FALSE.; You win the bonus only if the statement is indeed TRUE.; You will be entered into a lottery if the statement is indeed TRUE.; You win the bonus irrespective of the truth of the statement.]

Flat_quiz3. Suppose you state a belief B equal to 67, which of the following is correct? [You will be entered into a lottery if the statement is indeed TRUE.; You win the bonus irrespective of the truth of the statement.; You win the bonus only if the statement is indeed TRUE.; You win the bonus only if the statement is indeed FALSE.]

Flat_quiz4. Suppose you state a belief B equal to 42, which of the following is correct? [You win the bonus irrespective of the truth of the statement.; You will be entered into a lottery if the statement is indeed TRUE.; You win the bonus only if the statement is indeed TRUE.; You win the bonus only if the statement is indeed FALSE.]

BDM INSTRUCTIONS

Determining your bonus

In each part, you will have a chance to earn **a bonus of $0.40** for each belief B that you report. Your bonus for each part will be determined by randomly selecting one statement from that part to count and computing your payment according to the procedure below for the statement that counts.

You will maximize your chance of earning the bonus for each statement if you report your beliefs as accurately as possible. That is, there is nothing to gain by stating a number that differs from what you actually believe.

Procedure

- After you state your belief B, the computer will randomly draw a number W, with values between 0 and 100. Each value is equally likely to be drawn. You should think of W as a number of winning lottery tickets.
- If your belief B is at least as high as W (that is, $B \geq W$), then you receive the bonus if the statement is TRUE (and do NOT receive the bonus if the statement is FALSE).
- If your belief B is less than W (that is, $B < W$), you will be entered into a lottery with a W% chance of winning the bonus, which works as follows:
    - The winning ticket numbers are 1 through W.
    - We will randomly draw a ticket number L, where each ticket number (from 1 to 100) is equally likely to be drawn.
    - You receive the bonus if L is one of the winning ticket numbers.

This procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible about the likelihood you think the statement is TRUE.

We emphasize that this is a NO DECEPTION study. We will draw the random numbers and calculate your bonus behind the scenes following the procedures we described to you (so you

will not see any of the draws for any belief you state).

Next, we will ask a series of questions to check your understanding of these instructions. You must answer all of the questions to advance to the task.

BDM_quiz1. If you report a belief B equal to 100, which of the following is correct?
[You win the bonus only if the statement is indeed TRUE.; You win the bonus only if the statement is indeed FALSE.; You will be entered into a lottery if the statement is indeed TRUE.; You may be entered into a lottery, depending on the value of the randomly drawn number W.]

BDM_quiz2. Suppose you state a belief B equal to 28, which of the following is correct?
[You win the bonus only if the statement is indeed FALSE.; You win the bonus only if the statement is indeed TRUE.; You will be entered into a lottery if the statement is indeed TRUE.; You may be entered into a lottery, depending on the value of the randomly drawn number W.]

BDM_quiz3. Suppose you state a belief B equal to 67, and the randomly drawn value of W is 82, which of the following is correct?
[You have a 67% chance of winning the bonus.; You have an 82% chance of winning the bonus.; You win the bonus only if the statement is indeed TRUE.; You win the bonus only if the statement is indeed FALSE.]

BDM_quiz4. Suppose you state a belief B equal to 42, and the randomly drawn value of W is 37, which of the following is correct?

[You have a 55% chance of winning the bonus.; You have a 37% chance of winning the bonus.; You win the bonus only if the statement is indeed TRUE.; You win the bonus only if the statement is indeed FALSE.]

BSR INSTRUCTIONS

Determining your bonus

In each part, you will have a chance to earn a bonus of $0.40 for each belief B that you report. Your bonus for each part will be determined by randomly selecting one statement from that part to count and computing your payment according to the procedure below for the statement that counts.

You will maximize your chance of earning the bonus for each statement if you report your beliefs as accurately as possible. That is, there is nothing to gain by stating a number that differs from what you actually believe.

Procedure

- After you state your belief, the computer will randomly draw two numbers, X and Y, each with values between 0 and 100. For each draw, each number is equally likely to be selected. Draws are independent in the sense that the value selected for X in no way affects the value selected for Y and vice versa.
- If the statement is TRUE, then you receive the bonus if your belief B is greater than or equal to either X or Y.
- If the statement is FALSE, then you receive the bonus if your belief B is smaller than either X or Y.

This procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible about the likelihood you think the statement is TRUE.

We emphasize that this is a NO DECEPTION study. We will draw the random numbers and calculate your bonus behind the scenes following the procedures we described to you (so you will not see any of the draws for any belief you state).

Next, we will ask a series of questions to check your understanding of these instructions. You must answer all of the questions to advance to the task.

BSR_quiz1. If you report a belief B equal to 100, which of the following is correct?

[You win the bonus if the statement is TRUE and B is larger than or equal to either X or Y.; You win the bonus only if the statement is FALSE.; You win the bonus if the statement is FALSE and B is larger than or equal to either X or Y.; You win the bonus if the statement is TRUE and B is smaller than both X and Y.]

BSR_quiz2. Suppose you state a belief B equal to 28, which of the following is correct?

[You win the bonus only if the statement is FALSE.; You win the bonus only if the statement is TRUE.; You win the bonus if the statement is FALSE and B is smaller than either X or Y.; You win the bonus if the statement is TRUE and B is smaller than both X and Y.]

BSR_quiz3. Suppose you state a belief B equal to 67 and the randomly drawn numbers are X = 60 and Y = 2, which of the following is correct?
[You have a 100% chance of winning the bonus.; You have a 60% chance of winning the bonus.; You win the bonus only if the statement is FALSE.; You win the bonus only if the statement is TRUE.]

BSR_quiz4. Suppose you state a belief B equal to 42 and the randomly drawn numbers are X = 88 and Y = 49, which of the following is correct?
[You have a 100% chance of winning the bonus.; You have a 49% chance of winning the bonus.; You win the bonus only if the statement is FALSE.; You win the bonus only if the statement is TRUE.]

SUBJECTIVE COMPREHENSION

Difficulty. Did you find it easy or difficult to understand the instructions for determining the bonus?
[Extremely easy; Somewhat easy; Neither easy nor difficult; Somewhat difficult; Extremely difficult]

Effort. How much effort did it take you to understand the instructions for determining the bonus?
[No effort at all; A little effort; A moderate amount of effort; A lot of effort; A great deal of effort]

PART 1

(Flat) In this part, we will ask you for your belief (the number B, between 0 and 100) about five different statements about chance events. Your bonus for this part is determined as described previously (you will earn a $0.20 bonus for completing this part).

Please state your beliefs as accurately as possible.

(BDM/BSR) In this part, we will ask you for your belief (the number B, between 0 and 100) about five different statements about chance events. Your bonus for this part is determined as described previously (one statement will be randomly selected to count for a $0.40 bonus).

Remember that the procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible.

Induced1. We will roll a fair, 6-sided die behind the scenes, with the numbers 1 through 6 on its sides. Now, consider the following statement:

"The outcome of the die roll is a number less than or equal to 6."

How likely do you think it is that the above statement is TRUE?

Induced2. We will roll another fair, 6-sided die, separately, behind the scenes. The sides of this die are also numbered from 1 to 6. Now, consider the following statement:

"The outcome of the die roll is a number equal to 0."

How likely do you think it is that the above statement is TRUE?

Induced3. We will roll another fair, 6-sided die, separately, behind the scenes. The sides of this die are also numbered from 1 to 6. Now, consider the following statement:

"The outcome of the die roll is an odd number."

How likely do you think it is that the above statement is TRUE?

Induced4. We will shuffle a deck of cards behind the scenes and draw the top card. This deck is a standard deck of 52 cards (with 13 hearts, 13 diamonds, 13 spades, and 13 clubs). Now, consider the following statement:

"The suit of the card that was drawn is hearts."

How likely do you think it is that the above statement is TRUE?

Induced5. We will put 10 poker chips in a bag and draw one without looking. The bag has 8 white chips and 2 red chips. Now, consider the following statement:

"The color of the chip that was drawn is white."

How likely do you think it is that the above statement is TRUE?

PART 2

Part 2 - your beliefs about factual statements

(flat) In this part, we will ask you for your beliefs B about each of the four factual statements you rated earlier in the study. Your bonus for this part is determined as described previously (you will earn a $0.40 bonus for completing this part).

Please state your beliefs as accurately as possible.

(BDM/BSR) In this part, we will ask you for your beliefs B about each of the four factual statements you rated earlier in the study. Your bonus for this part is determined as described previously (one statement will be randomly selected to count for a $${e://Field/bonus_bdm_bsr} bonus).

Remember that the procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible.

(order of Own1-Own 4 randomized)

Own1. Consider the following statement:

More than half of unauthorized immigrants residing in the United States in 2016 had been living in the country for 10 years or more.

How likely do you think it is that the above statement is TRUE?

Own2. Consider the following statement:

From 2009, when President Obama took office, to 2012, median household income adjusted for inflation in the United States fell by more than 4 percent.

How likely do you think it is that the above statement is TRUE?

Own3. Consider the following statement:

More people in the United States work in the coal industry than in the solar industry.

How likely do you think it is that the above statement is TRUE?

Own4. Consider the following statement:

West Virginia was part of the Confederacy during the American Civil War.

How likely do you think it is that the above statement is TRUE?

PART 3

Part 3 - your beliefs about others

In this part, we are interested in your beliefs about other participants' accuracy.

Specifically, we are going to randomly match you with another participant in the study. Let's call this individual "Person A".

Recall that earlier in the study, you rated the set of factual statements as either true or false. We will ask you for your beliefs B about how likely it is that **Person A was CORRECT** in determining if each statement is true or false.

The rules for determining the bonus are basically the same as in previous parts with the only difference that B = 100 means "I am certain Person A was CORRECT" while B = 0 means "I am certain Person A was INCORRECT".

Equivalently, you can think of B as the percentage of other respondents (besides yourself) who correctly determined whether the statement was true or false.

(flat) Please state your beliefs as accurately as possible.

(BDM/BSR) Remember that the procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible.

(order of Other1-Other4 randomized)

Other1. How likely do you think it is that **Person A was CORRECT** in rating the following statement as true or false?

More than half of unauthorized immigrants residing in the United States in 2016 had been living in the country for 10 years or more.

Other2. How likely do you think it is that **Person A was CORRECT** in rating the following statement as true or false?

> From 2009, when President Obama took office, to 2012, median household income adjusted for inflation in the United States fell by more than 4 percent.

Other3. How likely do you think it is that **Person A was CORRECT** in rating the following statement as true or false?

> More people in the United States work in the coal industry than in the solar industry.

Other4. How likely do you think it is that **Person A was CORRECT** in rating the following statement as true or false?

> West Virginia was part of the Confederacy during the American Civil War.

GENERAL INTRO TO PARTS 4 AND 5

Parts 4 and 5 - your beliefs about particular groups

The next two parts are similar to the last part, only now we are interested in your beliefs about the accuracy of specific groups of participants. We will ask you for your beliefs about one group before asking for your beliefs about a second group.

INTRO TO DEMOCRATS

(part 4 or 5 depending on order)

Part 4/5 - your beliefs about Democrats

In this part, we are going to randomly match you with another participant in the study who identifies with the Democratic Party. Let's call this individual "Person D".

We will ask you for your beliefs B about how likely it is that Person D was CORRECT in determining if each statement is true or false. Recall that B = 100 means "I am certain Person D was CORRECT" while B = 0 means "I am certain Person D was INCORRECT".

Equivalently, you can think of B as the percentage of Democrats in the study who correctly determined whether the statement was true or false.

(flat) Please state your beliefs as accurately as possible.

(BDM/BSR) Remember that the procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible.

(order of Dem1-Dem4 randomized)

Dem1. How likely do you think it is that Person D (a Democrat) was CORRECT in rating the following statement as true or false?

> More than half of unauthorized immigrants residing in the United States in 2016 had been living in the country for 10 years or more.

Dem2. How likely do you think it is that Person D (a Democrat) was CORRECT in rating the following statement as true or false?

> From 2009, when President Obama took office, to 2012, median household income adjusted for inflation in the United States fell by more than 4 percent.

Dem3. How likely do you think it is that Person D (a Democrat) was CORRECT in rating the following statement as true or false?

> More people in the United States work in the coal industry than in the solar industry.

Dem4. How likely do you think it is that Person D (a Democrat) was CORRECT in rating the following statement as true or false?

> West Virginia was part of the Confederacy during the American Civil War.

INTRO TO REPUBLICANS

(part 4 or 5 depending on order)

Part 5 - your beliefs about Republicans

In this part, we are going to randomly match you with another participant in the study who identifies with the Republican Party. Let's call this individual "Person R".

We will ask you for your beliefs B about how likely it is that Person R was CORRECT in determining if each statement is true or false. Recall that B = 100 means "I am certain Person R

was CORRECT" while B = 0 means "I am certain Person R was INCORRECT".

Equivalently, you can think of B as the percentage of Republicans in the study who correctly determined whether the statement was true or false.

(flat) Please state your beliefs as accurately as possible.

(BDM/BSR) Remember that the procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible.

(order of Rep1-Rep4 randomized)

Rep1. How likely do you think it is that Person R (a Republican) was CORRECT in rating the following statement as true or false?

> More than half of unauthorized immigrants residing in the United States in 2016 had been living in the country for 10 years or more.

Rep2. How likely do you think it is that Person R (a Republican) was CORRECT in rating the following statement as true or false?

> From 2009, when President Obama took office, to 2012, median household income adjusted for inflation in the United States fell by more than 4 percent.

Rep3. How likely do you think it is that Person R (a Republican) was CORRECT in rating the following statement as true or false?

> More people in the United States work in the coal industry than in the solar industry.

Rep4. How likely do you think it is that Person R (a Republican) was CORRECT in rating the following statement as true or false?

> West Virginia was part of the Confederacy during the American Civil War

CONFIDENCE

How confident are you that whenever you selected a belief B during this study, you were maximizing your chances of earning the bonus by precisely reporting what you actually believed?

[Not at all confident; Not very confident; Moderately confident; Very confident; Extremely confident]

Do you have any other comments for us? (optional)