RATIONALITY
& COMPETITION
CRC TRR 190

A Horserace of Methods for Eliciting Induced Beliefs Online

Daniel Banko-Ferran (University of Pittsburgh)
Valeria Burdea (LMU Munich)
Jonathan Woon (University of Pittsburgh)

Discussion Paper No. 562

January 20, 2026

A Horserace of Methods for Eliciting Induced Beliefs Online

Daniel Banko-Ferran[*] Valeria Burdea[†] Jonathan Woon[‡]

January 20, 2026[§]

Abstract

This study evaluates the effectiveness of three widely used belief elicitation methods in an online setting: the binarized scoring rule (BSR), the stochastic Becker-DeGroot-Marschak mechanism (BDM), and unincentivized introspection. Despite the theoretical advantages of incentive-compatible methods (BSR and BDM), we find that they impose significantly higher cognitive costs on participants, requiring more time and effort to implement, without delivering clear improvements in belief accuracy. In fact, BSR systematically leads to greater errors in reported beliefs compared to introspection, while BDM also reduces accuracy, though to a lesser extent. Surprisingly, individual differences in probabilistic reasoning skills do not mitigate these errors for BSR but do help improve accuracy under BDM. Our findings suggest that simpler, unincentivized approaches may offer comparable or even superior accuracy at a lower cognitive cost. These results have broad implications for the design of experiments and the interpretation of belief data in behavioral and experimental economics.

Keywords: belief elicitation, induced beliefs, incentives, online experiment

JEL Classification: C81, C89, D83, D91

[*] daniel.bankoferran@gmail.com. Department of Economics, University of Pittsburgh

[†] valeria.burdea@econ.lmu.edu. Department of Economics, LMU Munich

[‡] woon@pitt.edu. Department of Political Science, Department of Economics (secondary), and Pittsburgh Experimental Economics Laboratory, University of Pittsburgh

[§] Financial support from the National Science Foundation Decision, Risk, and Management Science Program (Grant SES-2117694) and support by Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) are gratefully acknowledged. We pre-registered our analysis plan on the Open Science Framework (https://osf.io/d68f5). Data and analysis materials (R code) will be made available online upon publication. The study received approval by the University of Pittsburgh Human Research Protection Office (IRB) as an exempt study (STUDY21100109). The authors have no competing interests to declare.

1 Introduction

Understanding how individuals form beliefs about probabilistic outcomes, and what factors affect the accuracy of those assessments, is important for inferring people's preferences and learning how these factors influence social outcomes. For example, expectations about future inflation and employment can help us learn about the current state of the economy and anticipate a recession (Manski, 2004). Precise measurement of beliefs is also critical for interpreting polling results and evaluating policy changes (Brace et al., 2002).

However, eliciting accurate beliefs about probabilistic events is challenging. Factors such as cognitive biases, individual preferences, disinterest in the task or inability to form a belief pertinent to the task can lead participants to report beliefs in a noisy manner. Economists have developed several incentive-compatible methods with the theoretical properties necessary to address these issues. These properties ensure that rational participants are incentivized to provide the effort necessary to report the belief that maximizes their utility function.

In order to achieve this, though, these methods are complex to implement in an experimental setting. Such complexity can induce suboptimal decision-making, increase decision costs, and promote complexity aversion or avoidance which can lead to a reliance on simplistic heuristic procedures that reduce the accuracy of elicited beliefs (Oprea, 2020). This is because decision makers suffer cognitive costs from reasoning more deeply and this process can lead to more mistakes as the procedures become more complex (Banovetz and Oprea, 2023). A boundedly rational agent would have cognitive limitations that lead to some elicitation procedures being too costly for fully rational behavior (Plott, 1986). Indeed, studies conducted in controlled laboratory settings (Charness, Gneezy and Rasocha, 2021; Hao and Houser, 2012; Hollard, Massoni and Vergnaud, 2016; Hossain and Okui, 2013; Schlag, Tremewan and Van der Weele, 2015; Schotter and Trevino, 2014; Trautmann and van de Kuilen, 2014; Vespa and Wilson, 2016; Wang, 2011) suggest that the complexity of these incentive compatible mechanisms can confuse or distract participants, thus reducing the reliability of the beliefs that are elicited through these scoring rules. Moreover, some studies suggest the incentive-compatible methods do not differ from unincentivized methods in terms of the accuracy of the elicited beliefs, as measured by comparing the belief with an external objective truth (Schlag, Tremewan and Van der Weele, 2015; Schotter and Trevino, 2014), and may even perform worse (Danz, Vesterlund and Wilson, 2022).

Although these methods may be theoretically sound, their practical usefulness is limited if they fail to elicit consistently accurate beliefs and thus are not *behaviorally* incentive-compatible (Danz, Vesterlund and Wilson, 2022). Still, lab samples tend to be less diverse and more educated than the general population, which may impact the generalizability of these findings to other settings with more diverse participant pools, and costlier to achieve sufficiently-powered sample sizes (Matousek, Havranek and Irsova, 2022). Hence, determining the most effective method to incentivize participants to report accurate beliefs in non-lab settings remains an open issue.

To address this gap, we conducted an online experiment to investigate the performance of the most popular and state-of-the-art methods for eliciting participant beliefs about objective, induced probabilities. In particular, we compared the stochastic Becker-DeGroot-Marshak mechanism (BDM) (Allen, 1987; DuCharme and Donnell, 1973; Grether, 1981, 1992; Holt, 2007; Karni, 2009) and the Binarized Scoring Rule (BSR) (Hossain and Okui, 2013) to a benchmark consisting of a flat rate payment scheme relying on introspection (INTRO). We measured four performance outcomes: perceived difficulty, implementation time, rate of belief errors (i.e., deviations from objective probabilities), and frequency of "central-bias" in beliefs (a tendency to report moderate probabilities rather than more extreme values).

In addition, we examined differences in performance conditional on participants' probabilistic reasoning skills. The diversity of online participant pools—where individuals vary widely in terms of educational background, numerical proficiency, and familiarity with formal probability concepts—can make the assumption of *probabilistic sophistication* (Hossain and Okui, 2013, p. 989) especially questionable. Indeed, research shows that people's probabilistic reasoning skills can differ substantially (Burfurd and Wilkening, 2021; Primi et al., 2017), suggesting that a one-size-fits-all elicitation procedure could fail to yield accurate or reliable beliefs for a significant portion of respondents. Examining whether and how probabilistic sophistication influences the effectiveness of different elicitation methods is therefore a key contribution of our study, as well as a crucial consideration for researchers seeking to apply these methods in large-scale online settings.

We find that incentive compatible methods fare no better than introspection in terms of the accuracy of the elicited beliefs. In fact, often, both the BDM and the BSR methods increase the belief errors, with BSR doing so in a more systematic manner than BDM. The most important individual characteristic that influences how participants respond to these incentives is probabilistic reasoning. Specifically, individuals with high probabilistic reasoning skills are significantly less affected by the BDM method than those with low such skills, but no such positive effect is found for the BSR method. Nevertheless, even the high probabilistic reasoning individuals do not reduce their belief reporting error rate further than those facing the introspection method. Moreover, the incentive-compatible methods have additional implementation costs related to longer duration and higher perceived difficulty.

Through an exploratory survey of economists who use experiments in their research, we also find that although subject comprehension is a more important criterion than incentive-compatibility when choosing which method to use for belief elicitation, academics would rather use the BSR or the BDM method in their research as it is believed to have higher chances of publication.

To the best of our knowledge, this is the first study to systematically compare the performance of these three belief elicitation methods in an online setting, while accounting for individual differences in probabilistic reasoning (numeracy). By exploring the interaction between numeracy and belief accuracy, our findings contribute to the literature on belief formation and measurement, with implications for behavioral and experimental economics. In particular, our results shed light on the tradeoffs between accuracy and complexity in belief elicitation and provide guidance for future studies that aim to elicit beliefs in diverse and less controlled environments.

Our study is particularly relevant given the recent proliferation of online platforms for research such as Prolific and CloudResearch. There has been greater discussion of online studies as a promising complement to studies conducted in the lab. Online studies can be conducted more quickly, less expensively, and with greater numbers and diversity of participants. Yet moving from highly controlled in-person laboratories to online settings may come at the cost of subjects paying less attention, being more easily distracted, or being more time-constrained, due to the loss of experimenter control over the decision-making environment. Meta-analyses of experimental papers have found robust qualitative findings across populations, albeit with wide variability due to differences in design, context (lab vs. field), and subject pools (Matousek, Havranek and Irsova, 2022). We show that a concern about how the different available methods influence the quality of belief data elicited in online studies is justified and that theoretical benefits do not always translate to data quality.

The rest of the paper is organized as follows. In Section 2 we give an overview of the belief elicitation methods relevant for our study and the academics' views on these methods. Section 3 details our experimental design and corresponding hypotheses. This is followed by the presentation of the main and exploratory results in Section 4. Section 5 discusses our findings and concludes.

2 Background

2.1 Becker-DeGroot-Marshak mechanism

The Becker-DeGroot-Marshak (BDM) mechanism is one of the most widely used and empirically successful incentive-compatible methods for belief elicitation (Charness, Gneezy and Rasocha, 2021). The idea behind this mechanism was introduced by Savage (1971) and implemented for the first time by Grether (1981). Its more recent formalization as a belief elicitation mechanism is due to Karni (2009), Holt and Smith (2009, 2016) and Mobius et al. (2011). It is generally considered to be less complex than the Binarized Scoring Rule (BSR) (Hossain and Okui, 2013), although the complexity and effectiveness of BDM depends on the specific implementation. The simplest version of BDM is the multiple price list (MPL) format, where participants make a series of choices between a binary lottery and a certain monetary payoff in a tabular format, reporting their willingness to accept a certain payoff for uncertain outcomes (Baillon and Bleichrodt, 2015; Trautmann and van de Kuilen, 2014). The most thorough examination of BDM procedures was performed by Burfurd and Wilkening (2018), which compared three different procedures of BDM to eachother: the MPL format of Trautmann and van de Kuilen (2014), the detailed explanation of the mechanism as described in Holt and Smith (2009), and an "analogy-based" approach used in Hao and Houser (2012).[1] Although the MPL format is the simplest, Burfurd and Wilkening (2018) find that the analogy approach based on Hao and Houser (2012) is faster to implement, significantly reducing the amount of time it takes to run while still maintaining similar levels of accuracy and precision in the elicited beliefs.[2] In the implementation of the analogy-based approach of Burfurd and Wilkening (2018), after subjects report a belief r (from 0 to 100) about the probability of an event occurring, they are paid either according to the realization of the event or the draw of a black chip from a bag (with equal payoffs). Whether the participant is paid from the chip draw instead of the event realization is based on a random number (between 0 and 100) being larger than r.

Burfurd and Wilkening (2018) find that the experiment duration for participants in the analogy-based condition was 28% shorter, as measured by the total time it takes for a subject to go through the experiment. As Hao and Houser (2012) argue (and we concur), brevity and efficient communication in experiment instructions is important given the limited attention span of subjects. We view the analogy-based format as the most promising implementation of BDM as it is the shortest implementation procedure rigorously studied thus far, yet it helps subjects make quick decisions without compromising accuracy. This, together with a need for comparable instructions across treatments, is why we chose the analogy version for the implementation of the BDM method in our experiment.

2.2 Binarized Scoring Rule

The Binarized Scoring Rule (BSR) is a more recent incentive-compatible method for belief elicitation and is now more widely used than the previously popular Quadratic Scoring Rule (QSR)

[1] The "analogy-based" format explains the belief elicitation procedure by reframing the belief elicitation procedure as a drawing of a black chip from a bag filled with a known number (usually 10 or 100) of black and white chips.

[2] Accuracy and precision are measured by the mean of the absolute error of a subject's reports relative to an objective Bayesian posterior and the standard deviation of absolute errors for each individual, respectively.

due to its robustness to risk aversion. Hossain and Okui (2013) proposed and experimentally tested BSR to show that it outperforms QSR on a variety of metrics, but beyond this study, there are not many other studies which compare BSR to other belief elicitation methods so far. They find that BSR had improved accuracy in elicited beliefs for risk-averse agents (as measured by the squared difference between the reported and true probabilities) for less extreme probabilities. Two other studies found similar advantages of using BSR over QSR due to its better prediction of beliefs for risk-averse individuals Erkal, Gangadharan and Koh (2020); Harrison, Martínez-Correa and Swarthout (2014). However, Danz, Vesterlund and Wilson (2022) find deficiencies with the BSR method in that explaining the quantitative incentives creates similar errors in elicited beliefs to QSR. The no-information treatment in their comparison of BSR implementations fared better than the information (with or without the use of a calculator) treatments. This lends further evidence that less complex procedures may improve the quality of elicited beliefs.

2.3 Introspection

Simpler belief elicitation methods like introspection or flat-rate payments pose the risk that the mechanisms are not theoretically robust to concerns about payoff-maximization and incentive-compatibility. However, recent empirical evidence suggests these concerns may be unfounded. Despite the strong theoretical properties of BDM and BSR, many recent studies suggest that they fail to outperform introspection in terms of the accuracy of beliefs (Burfurd and Wilkening, 2021; Charness, Gneezy and Rasocha, 2021; Hollard, Massoni and Vergnaud, 2016; Schlag and Tremewan, 2021).

Indeed, often participants misreport their beliefs using BSR even when the probability is objectively known (Burfurd and Wilkening, 2018; Danz, Vesterlund and Wilson, 2022; Hao and Houser, 2012). Trautmann and van de Kuilen (2014) conducts an extensive examination of elicitation mechanisms (QSR, BDM, introspection, and outcome matching) in a simple two-person ultimatum game and found no significant difference in performance between incentivized methods and introspection. Introspection, in addition to being cost-effective, fared much better than incentivized elicitation methods in terms of additivity. This suggests that the complexity of the elicitation methods may hinder the theoretical advantages of the incentivized payment scheme.

2.4 Heterogeneity in probabilistic reasoning skills as source of error

Burfurd and Wilkening (2021) look at whether one source of difficulty might be an interaction between the belief elicitation mechanism and individual cognition skills. If heterogeneous decision-making varies systematically with cognitive ability, then complex mechanisms might create less reliable data among participants who score lower on tests of cognition. To check for this, they compare the performance of BDM and introspection in a belief-updating task and find smaller belief errors for participants who are consistent with probabilistic reasoning and those who are not (although they do not find differences in the average accuracy of beliefs elicited using the two methods overall). In a follow-up experiment, they use Raven's Progressive Matrices and a Cognitive Reflection Test to observe a negative relationship between cognitive performance scores and errors in reported beliefs (measured with an objective prior). This suggests that both the complexity of the mechanism and the probabilistic reasoning of participants are important considerations when eliciting and interpreting beliefs.

Burdea and Woon (2022) measure beliefs in an online setting about probabilities with non-induced priors (beliefs about the likelihood a factual statement about a real-world event is true) and compare them 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). They find that BSR and BDM require longer implementation time and are associated with higher comprehension costs (perceived difficulty and effort) than the flat rate payment. Their study also highlights the importance of cognitive factors in affecting the quality of elicited beliefs in that both incentive-compatible methods improved the accuracy of elicited beliefs about induced probabilities, but only for less educated participants. However, Burdea and Woon (2022) did not measure participants' probabilistic reasoning skills directly. When eliciting first and second order beliefs, they also find that the distribution of beliefs across methods differs, in that incentive-compatible methods lead people to report fewer beliefs at 50%. One limitation of their study is that it involved a limited number of items for eliciting participants' beliefs about induced probabilities. In our study, we compare these belief elicitation methods using beliefs about objective probabilities with induced priors in a commonly used belief updating task.

2.5 Exploratory expert survey

As of late, there is no evidence regarding how different researchers perceive the differences and similarities between methods of eliciting beliefs and whether certain preferences are due to publication bias or genuine belief in the superiority of a certain method. We conducted an exploratory survey of the experimental community of researchers to fill this gap. We collected data on what methods researchers prefer and their reasoning behind their preferences and attitudes across various methods. Participants were recruited through email and selected based on having a history of publications in experimental economics journals which involve the use of belief elicitation procedures. They were asked to fill out an online Qualtrics survey and informed that participation is completely voluntary. No compensation was offered to participants for completing the survey. The survey took approximately 5-7 minutes to complete and was conducted over 2 weeks.

The survey was distributed in April 2022 to 151 academic scholars. 25 accepted our invitation and completed the survey. 70% of these are faculty with tenure while 20% are also faculty but without tenure. Hence, even though our sample is small, the respondents have accumulated significant experience in the field and this should render their opinions informative.

Participants were first presented with a brief description of five major belief elicitation methods (BDM, Allen (1987); DuCharme and Donnell (1973); Grether (1981); Holt (2007); Karni (2009) BSR, Hossain and Okui (2013); QSR, Brie (1950); frequency/interval method; introspection). Respondents then were asked questions related to their preferred method for their own research and when refereeing papers of other researchers. We then asked participants to imagine they are reviewing a paper for a journal where beliefs were elicited from participants in an experiment and these were a primary outcome variable. We asked how likely they would be to reject the paper or require the experiment to be run with a different method if the belief elicitation used the given method. We asked the same question for the scenario where beliefs are secondary outcome variables, and for the case where they are reviewing a paper at a field journal (and beliefs are either primary or secondary outcome variables). Then, we asked participants how important incentive compatibility, comprehension, duration, and scope of belief (primary/secondary measure) were in their decision. Participants were also asked a set of demographic questions: year of PhD completion, gender, country where current university/institution is located, and position at the current university/institution.

We present the results related to the three methods compared in our experiment and discuss the remaining two in the relevant sections. The corresponding tabulation of answers related to the QSR and the frequency methods are presented in the Appendix. The majority of respondents reported being extremely familiar with all methods (Table 1). Additionally, when beliefs were used as a primary outcome, no respondents reported using introspection often as a belief elicitation method. More common was BDM and BSR, where about a third reported using one of the two "often". Introspection was more commonly used when beliefs were elicited as a secondary outcome measure.

Table 1: Familiarity with belief elicitation method

		BDM	**BSR**	**INTRO**
Familiarity	Not at all	0.00	0.04	0.00
	Slightly – Moderately	0.08	0.12	0.00
	Extremely – Very	0.92	0.84	1.00
Use as primary outcome	Never	0.28	0.36	0.32
	Rarely – Sometimes	0.4	0.4	0.68
	Often	0.32	0.24	0
Use as secondary outcome	Never	0.16	0.44	0.12
	Rarely – Sometimes	0.4	0.4	0.68
	Often	0.32	0.24	0

Note: Tabulation of responses to question: "How familiar are you with the XX method for belief elicitation?" and "How often do you use the XX method for belief elicitation when beliefs are a primary (secondary) outcome variable in your work? N = 25. Numbers represent frequency of answer. BDM = Becker-DeGroot- Marshak mechanism, BSR = Binarized Scoring Rule, INTRO = Introspection.

This pattern seems to be at least partly driven by publication likelihood reasons since introspection is reported to be notably more likely to be rejected for publication in a top 5 journal when beliefs are used as a primary outcome when compared to BDM and BSR, though this difference shrinks when it is used to elicit beliefs as a secondary outcome (Table 2). This difference, though shrinking slightly, persists even when considering lower ranked (top field) journals. While the experts were very familiar with the QSR method as well (88% reporting being extremely or very familiar with it), its attractiveness in terms of meeting the publication requirements is lower than of the other popular complex methods with 28% of participants stating that it would be very/somewhat likely for their paper to be rejected at a top 5 journal if they used it for their primary outcome.

Table 2: Likelihood to reject based on belief elicitation method

		BDM	**BSR**	**INTRO**
Top 5	Very/somewhat unlikely	0.88	0.8	0.36
likelihood to reject (primary outcome)	Neither unlikely nor likely	0.12	0.16	0.2
	Very/somewhat likely	0	0.04	0.44
likelihood to reject (secondary outcome)	Very/somewhat unlikely	0.96	0.84	0.64
	Neither unlikely nor likely	0.04	0.12	0.28
	Very/somewhat likely	0	0.04	0.08
Non-top 5 (top field)	Very/somewhat unlikely	0.84	0.84	0.44
... (primary or secondary outcome)	Neither unlikely nor likely	0.12	0.16	0.32
	Very/somewhat likely	0.04	0	0.24

Note: Tabulation of responses to question: “Suppose you were reviewing a paper, for a top 5 economics journal [field journal (not a top 5)], that elicited beliefs from participants in an experiment. These beliefs are a primary [secondary] (primary or secondary) outcome variable. How likely would you be to reject the paper or require the experiment be re-run with a different method if the belief elicitation process used this method?” Online survey conducted in October 2022. N = 25. Numbers represent frequency of answer. BDM = Becker-DeGroot-Marshak mechanism, BSR = Binarized Scoring Rule, INTRO = Introspection.

When asked which method economists would use in a new study of their own, the BDM method is preferred by a large majority of participants (Table 3). Perhaps most interesting in light of the findings of our experiment is that subject comprehension is ranked more important than incentive compatibility when it comes to choosing how to elicit beliefs in an experiment (Table 4). This exposes a paradox in experimental research: although experimenters highly value subject comprehension and recognize that incentive-compatible methods are more challenging for participants, they still favor these methods in practice—largely driven by the fear of journal rejection.

Table 3: Preferred method for new study

Elicitation Procedure	BDM	QSR	BSR	FREQ	INTRO	OTHER
Prefer to use (0/1)	0.76	0.6	0.48	0.36	0.24	0.24

Note: Tabulation of responses to question: “If you were conducting a new study now, which method(s) would you use to elicit beliefs? Check all that you would be willing to implement in your own research”. Online survey conducted in October 2022. N = 25. Numbers represent frequency of answer. BDM = Becker-DeGroot-Marshak mechanism, QSR = Quadratic Scoring Rule, BSR = Binarized Scoring Rule, FREQ = Frequency Method, INTRO = Introspection.

Table 4: Importance of belief elicitation criteria

	Not at all	Slightly – Moderately	Extremely – Very
Subject comprehension	0.04	0.12	0.84
Incentive compatibility	0	0.36	0.64
Length of time	0.04	0.44	0.52
Scope of Belief	0.2	0.6	0.2

Note: Tabulation of responses to question: ”How important is each of the following criteria when choosing how to elicit beliefs in an experiment?” Online survey conducted in October 2022. N = 25. Numbers represent frequency of answer.

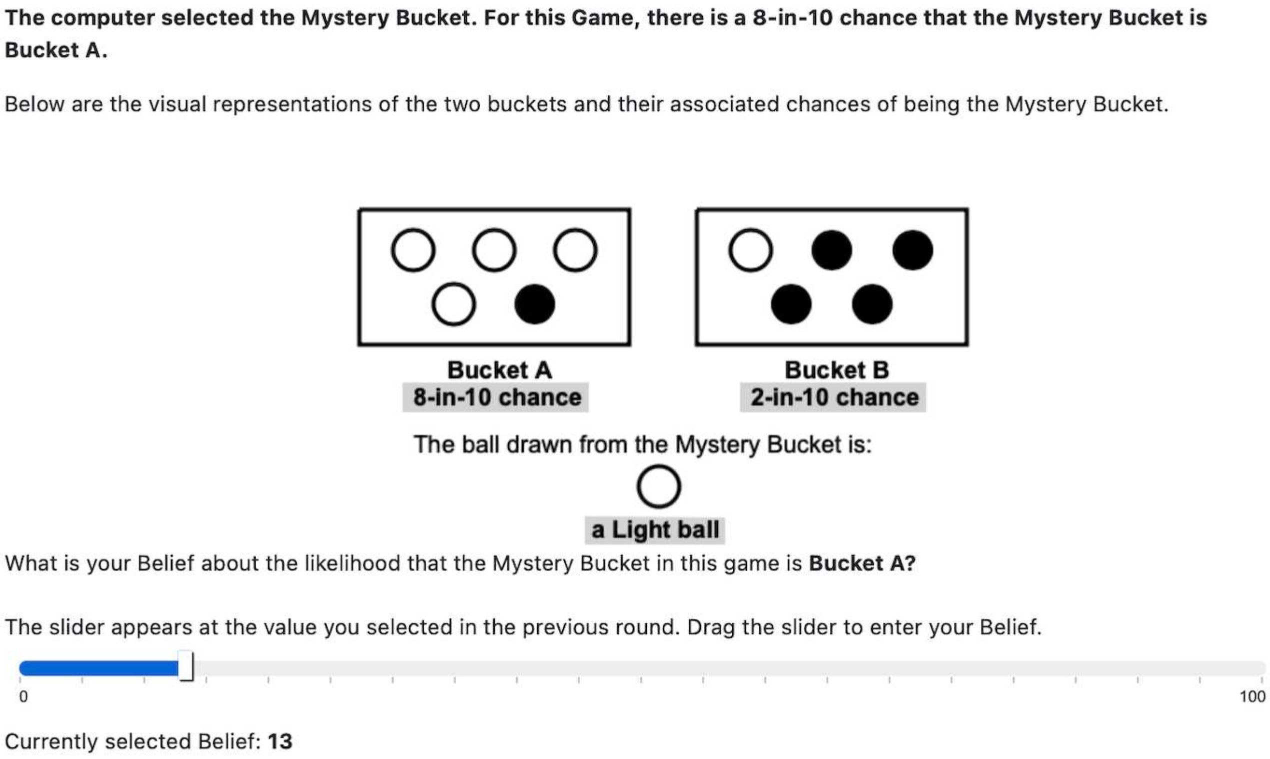
3 The experiment

Our design, hypotheses, and analysis plan were pre-registered on the Open Science Framework and can be accessed at: `https://osf.io/d68f5`. We compared methods for eliciting participants' probabilistic beliefs about an uncertain event using a between-subject design with three conditions, described below, that differ in the elicitation method used. The exact instructions used in the experiment are in the Appendix.

3.1 Design

We used a standard belief updating task known as the Bucket Game adapted from Burfurd and Wilkening (2018) with a pre-specified objective probability known to the participants. In this game there are two computerized buckets, A and B, and a pre-specified chance ($X - in - 10$) of the computer selecting bucket A as the Mystery Bucket. Each bucket is filled with 5 dark and light balls and the composition of dark and light balls varies between buckets. Participants know these compositions as well as the chance of the computer selecting Bucket A. Given this information, participants are first asked about their belief that the selected bucket (the Mystery Bucket) is Bucket A. Next, participants observe the draw of a ball and are asked again about their belief that the selected bucket is Bucket A, given the observation of the drawn ball being dark or light. Thus, participants are asked to state their belief twice every period: once before the ball is randomly drawn-our measure of prior beliefs-and once afterwards-our measure of posterior beliefs. Figure 1 shows a screenshot of the instructions page for the elicitation of posterior beliefs (after observing a signal).

Figure 1: Screenshot where X = 8 (after signal)



The computer selected the Mystery Bucket. For this Game, there is a 8-in-10 chance that the Mystery Bucket is Bucket A.

Below are the visual representations of the two buckets and their associated chances of being the Mystery Bucket.

Bucket A
8-in-10 chance

Bucket B
2-in-10 chance

The ball drawn from the Mystery Bucket is:

a Light ball

What is your Belief about the likelihood that the Mystery Bucket in this game is **Bucket A?**

The slider appears at the value you selected in the previous round. Drag the slider to enter your Belief.

Currently selected Belief: **13**

Participants played the Bucket Game for 10 periods. In each period, only the value of *X* changed. Participants were presented with the following values of *X*: 2, 3, 5, 7, and 8. The prior probabilities are thus 20%, 30%, 50%, 70%, and 80%-the center value, two lower probabilities and their corresponding higher counterparts, like in Danz, Vesterlund and Wilson (2022). Participants first observe all 5 values in random order (first elicitation), and then the same five values in random order again (second elicitation). That is, they observe the same value twice, but the randomization of the order is conducted in blocks of 5 periods. Participants were not informed about the possibility that the values of X may be encountered multiple times, but they did know that they will play this game 10 times. After reading the description of the Bucket Game, participants were provided with a summary and then asked to complete 4 related comprehension questions. They could only proceed after answering these questions correctly.

In all treatments, participants were told that their belief represents their best guess about the chance that the computer selected Bucket A given X. After the explanation of the Bucket Game, we also provided detailed instructions for what each number (between 0 and 100) they might report as their belief in every period meant in the context of the Bucket Game. We then asked 4 related comprehension questions that participants had to answer correctly before proceeding, in order to make sure that everyone understood the mapping between the reported beliefs and the percent chance that the Mystery Bucket is Bucket A.

Participants were then provided with instructions for the elicitation mechanism corresponding to their experimental condition. They were randomly assigned to one of the following belief elicitation conditions: the stochastic Becker-deGroot-Marschak mechanism (BDM), the Binarized Scoring Rule (BSR), or introspection (INTRO). Thus the incentives for the elicitation of beliefs in the Bucket Game differed across conditions.

In BDM, we used an analogy-based approach similar to (Hao and Houser, 2012) and Burfurd and Wilkening (2018). Participants are informed that their belief about the likelihood that the selected bucket is Bucket A is compared to a randomly drawn number W (from 0 to 100) that determines the number of winning tickets in a lottery. If the elicited belief is greater than or equal to W, then the participant receives an \$8 bonus if the statement is true. If not, then the chance of winning the bonus is based on a lottery with a $W\%$ probability of winning.

In BSR, subjects were asked to state a belief between 0 and 100 that Bucket A was selected, as in Vespa and Wilson (2016). Then, two numbers between 0 and 100 are randomly and independently drawn. The subject receives an $8 bonus if Bucket A was selected and their belief is higher than either of the two draws, or if Bucket B was selected and their belief is lower than either of the two draws.

Finally, in INTRO, subjects were asked to state their belief that Bucket A was selected but informed that their payment was not tied to their report. In order to control for any differences arising simply from the uncertainty regarding receiving the bonus, we introduced a lottery for the bonus also in INTRO. Specifically, participants in this treatment received an $8 bonus if the outcome of a fair coin flip was heads, irrespective of their reported belief.

After reading the payment rule description, participants were asked two questions regarding perceived difficulty and complexity of the mechanism. Perceived difficulty was measured on a 5-point scale ranging from “extremely easy” to “extremely difficult”. Complexity was also measured on a 5-point scale ranging from “extremely simple” to “extremely complex”. Next, participants were asked another set of comprehension questions about the process by which the bonus for their beliefs is determined. All questions were multiple-choice type with four possible answers. For all of these as well as previous comprehension question sets, participants had two trials to answer correctly, after which they were provided with the correct answer and could proceed. The comprehension quiz following the elicitation mechanism instructions included four questions specifically tailored to the incentive mechanism, and two that were common across mechanisms. Even though these questions are implemented to ensure understanding rather than to test it, we introduced the two common questions to check whether participants are less likely to answer these questions correctly from the first try, while holding the wording of the question constant. We measured response time for all questions in the study.

Before the belief elicitation game, participants, in all conditions, completed a survey that collected basic demographic information and evaluated their probabilistic reasoning, a component of numeracy skills. To measure probabilistic reasoning, we use the Probabilistic Reasoning Scale introduced by Primi et al. (2017) and shown to be positively correlated with subjective numeracy, the cognitive reflection test, and the conditional probability task (Primi et al., 2019). This scale involves a 9-item questionnaire, where each item has three possible answers presented to the participant, but only one correct answer. Participants have only one chance to answer each question correctly and are not provided with any feedback on their answers. A participant's probabilistic reasoning score is the number of correct answers in this questionnaire.

3.2 Procedures

517 participants were recruited from CloudResearch in late July 2022. Participants were screened so that only CloudResearch-approved participants (passed attention and engagement measures) who reside in the United States and had completed 100 approved HITs (worker tasks) and had an approval rating of at least 95 percent could enroll in the experiment. Participants received a $4 completion fee plus an additional bonus of up to $8 for completing the experiment. The average payout was $10 for approximately 15-20 minutes of work, which is a comparably competitive rate for other tasks on CloudResearch. This is clearly stated in the description of the study that a worker sees in CloudResearch and in the consent for the potential participant needs to accept before proceeding with the study. Participants could terminate the study at any point.

Given the variables measured with this design, we can define the prior error as the absolute difference between a subject's reported belief and the objective prior probability of Bucket A being chosen in the first stage of each period, before the observation of a random ball draw. Similarly, the posterior error is the absolute difference between the subject's reported belief and the objective Bayesian posterior belief in the second stage of each period after a random draw is observed. We also compute the rate at which subjects make a central report of 50 when reporting the prior or posterior probability of Bucket A in either the first or second stage of each period.

Moreover, we use the reported subjective difficulty of the incentive mechanism to categorize participants into two groups based on the median split: participants that perceive the mechanism as *highly difficult* if their report is greater or equal to the median value, and *lowly difficult* otherwise. Similarly, we use their probabilistic reasoning score to categorize participants into two groups based on the median split: participants are categorized as having *high probabilistic reasoning* if their probabilistic reasoning score is greater than or equal to the median value, and as having *low probabilistic reasoning* otherwise.

3.3 Hypotheses

We use the following criteria to compare the belief elicitation methods: (1) how easy/complex participants found it to understand the methods' instructions for determining their bonus (perceived difficulty); (2) how long it took to elicit participants' beliefs under each method (implementation duration); and (3) qualitative characteristics of the elicited beliefs including the accuracy of elicited beliefs and the frequency of central reports (beliefs equal to 50%). We also investigate if any differences in accuracy are due to differences in the perceived difficulty of the method and in participants' probabilistic reasoning skills. Our study is designed to test six hypotheses regarding these comparisons, as detailed below. Our main hypotheses and analysis focus on prior beliefs since almost all previous data related to belief elicitation methods focuses on these. However, as exploratory analysis, we also analyze the posterior belief data we elicited and this is presented in a later section.

Hypothesis 1 (H1)**.** *Perceived Difficulty: Participants will report the highest perceived difficulty for the BDM mechanism, followed by BSR and then introspection.*

Our first hypothesis is motivated by the finding in Burdea and Woon (2022) that BDM is rated as more difficult than BSR. Moreover, previous literature has shown that incentivized methods are perceived as more difficult than introspection (Charness, Gneezy and Rasocha, 2021).

Hypothesis 2 (H2)**.** *Implementation Speed: The implementation of the belief elicitation task will be fastest under introspection, followed by BSR and BDM.*

Our second hypothesis follows from the first as more difficult instructions should require participants more time to process them, translating to longer implementation speeds. Implementation speed is measured by the time (in seconds) it takes to read the instructions, answer the comprehension questions, and complete the belief elicitation task.

Hypothesis 3 (H3)**.** *Belief Accuracy: Incentive compatible methods (BSR and BDM) are associated with lower belief error than introspection.*[3]

The third hypothesis relies on the general belief in experimental economics, that without monetary incentives, subjects will act more careless, and make more errors in reporting beliefs (see for e.g., Charness, Gneezy and Rasocha, 2021; Smith and Walker, 1993). We use this intuition for our online setting that has not been investigated so far, despite lab evidence suggesting the opposite might be the case for the BSR method (Danz, Vesterlund and Wilson, 2022).

Hypothesis 4 (H4)**.** *Central-Biased Beliefs: The frequency of central (50%) beliefs will be lowest under BSR, followed by BDM and then introspection.*

The above hypothesis is motivated by empirical findings in Burdea and Woon (2022) showing that incentive compatible methods incentivize participants to report less uncertainty associated with reporting a belief at 50%.

The next hypotheses are independent of the general effect of incentive-compatible methods that we describe in H3. In what follows, we describe how people may respond to incentive compatible methods not necessarily because of their incentive compatibility property, but rather due to participants' subjective perception of these methods or their psychological characteristics. In particular, we hypothesize that incentive-compatible measures (BSR and BDM) will interact positively with participants' perceived difficulty (leading to increased belief error) and negatively with their probabilistic skill.

Hypothesis 5 (H5)**.** *Belief Accuracy Sensitivity to Perceived Difficulty: Higher perceived difficulty of understanding the method's instructions increases the observed average belief error. Therefore,*

[3] The pre-registration erroneously stated that this hypothesis refers to both prior and posterior beliefs. However, the Exploratory Analysis section clarifies that the corresponding test for posterior beliefs is not part of the main analysis. We test this hypothesis for posteriors as well, in any case, in a later section.

the difference in belief error between introspection and incentive compatible methods (BSR and BDM) will be higher for participants who report a higher perceived difficulty.

According to this hypothesis, we expect that participants who perceive the BDM and BSR mechanisms to be more difficult, to over-strategize and report beliefs with higher error than under introspection compared to those participants who perceive the methods to be less difficult.

Hypothesis 6 (H6)**.** *Belief Accuracy Sensitivity to Probabilistic Reasoning: Higher probabilistic reasoning will lead to a decrease in the observed average belief error for incentive compatible methods. Therefore, the difference in belief error between introspection and incentive compatible methods (BSR and BDM) will be lower for participants who are categorized as having high probabilistic reasoning skills, compared to those categorized as having low such skills.*

Our last hypothesis is motivated by the higher mathematical complexity of the incentive compatible methods and the fact that people differ in their probabilistic reasoning skills (Benjamin, 2019). Given this, we expect that individuals with higher probabilistic reasoning skills will be better able to understand the incentive-compatible properties of the BSR and BDM methods and respond positively to them by reporting more accurate beliefs.

4 Results

We structure the presentation of our results into three sections: preliminaries, main analysis focusing on prior beliefs, and secondary analysis focusing on posterior beliefs. All non-preregistered analyses related to the main findings will be denoted as such.

4.1 Preliminaries

4.1.1 Sample description

Of the 517 subjects who started our study, 493 (95%) completed it. We do not include any partial observations in our analysis of the results. Attrition rates did not differ across conditions (4% in

INTRO, 3% in BDM and 6% in BSR), and the final sample size was evenly distributed across conditions (n = 164 in INTRO, n = 167 in BDM, n = 162 in BSR).

As expected, our online sample is more diverse than a typical laboratory sample of undergraduates. 4% of our sample is college-aged (between 18 and 24 years old), with 34% between 25 and 34 years old and 11.58% who were at least 55 years old. In terms of education, 16% had no more than a high school diploma or equivalent, 42% had completed a four-year undergraduate degree, and 10% had obtained a post-graduate degree. Though more diverse than laboratory samples, it is still younger and more educated than the general U.S. population according to the U.S. Census Bureau's 2019 Current Population Survey. Adults 55 years or older constitute 38% of all adults over 18 years old in the U.S population and 39% of the population completed no more than a high school education. Relative to the general population, our sample also skews slightly more white (80%) and male (58%). By comparison, 76% of the US. population identifies as "white only" according to the Census Bureau, and 52% of the adult population is male. Results are shown in Table A3 in the Appendix and stratified by treatment in Table A4 in the Appendix.

Comprehension of the instructions, as measured by the likelihood to answer the comprehension questions correctly on the first or second attempt (without the answer being revealed to them), was high at about 90% overall. We discuss the small differences across treatments that arose, in the "Perceived difficulty" section of the main analysis.

4.1.2 Differences between first and second elicitation

If the distribution of the elicited beliefs is different between the first and second elicitation for game periods with the same prior, this would be indicative of participants' behavior changing across rounds. If this is the case, we cannot pool the two elicitations for the main analysis. We check this by first comparing, using paired t-tests, the means of the elicited beliefs in the first five rounds of the Bucket game to the last five rounds (10 rounds in total). Then we compare the variances of the two sets of beliefs using Pitman's test of equality of variances for paired samples. Finally we compare the frequency of 50% beliefs using McNemar's Chi-Square test. We run these tests for each value of the prior likelihood that Bucket A was selected (0.2, 0.3, 0.5, 0.7, 0.8). The results of these tests are found in the Additional Tables section of Appendix A (Tables A5-A7). Our tests show that for some objective priors, there are significant differences between the first and second elicitation of beliefs. Thus, for our main analysis, we present our results separately for the first and second elicitation and discuss potential learning effects in the exploratory analysis section.

4.2 Main analysis

The main results, highlighted via numbers (from 1 to 6), are based on tests for each of our pre-registered hypotheses. Any deviations from the pre-registered analysis plan are documented and motivated. All the alternative, related analyses are exploratory.

4.2.1 Subjective difficulty

Recall that our pre-registered analysis regarding differences in comprehension difficulty across mechanisms utilizes a subjective 5-point measure of perceived difficulty ranging from 1 ("Extremely easy") to 5 ("Extremely difficult"). In line with H1, participants reported that the instructions for the two incentive-compatible methods (BDM and BSR) were significantly more difficult than the introspection one. Participants reported an average difficulty of 2.5 in INTRO and 2.9 in both BSR and BDM (different from our H1 prediction where we state that BDM would lead to higher average difficulty than BSR). Table 5 presents the regression results comparing these averages. Participants reported to experience significantly more difficulty understanding BSR and BDM than INTRO but no significant difference in understanding between BSR and BDM, even when controlling for high probabilistic skills (which do not seem to have a significantly negative effect on perceived difficulty; not pre-registered analysis).

Result 1. *Participants reported to have significantly more difficulty understanding BSR and BDM than INTRO but no significant difference in understanding between BSR and BDM was observed.*

Table 5: Average perceived difficulty of instructions across elicitation methods

	Pre-registered	Controlling for Numeracy
(Intercept)	1.88^{***}	1.90^{***}
	(0.22)	(0.22)
BSR	0.43^{***}	0.42^{***}
	(0.12)	(0.12)
BDM	0.41^{***}	0.41^{**}
	(0.12)	(0.12)
High Numeracy		-0.04
		(0.10)
Num. obs.	494	494
Includes covariates?	Yes	Yes

Note: The dependent variable is measured on a 1-5 Likert scale. The covariates include gender, age and education level. The regression model controlling for numeracy in column 2 is not pre-registered. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

A similar pattern emerges when we use alternative measures of difficulty (not pre-registered). First, we consider the likelihood of answering correctly any of the 6 comprehension questions related to the incentive mechanism. Table 6 shows that participants are significantly less likely to answer a question correctly in the first attempt in the BSR and BDM treatments compared to INTRO (with no significant difference between BSR and BDM). This difference weakens when considering correct responses provided in the second attempt though the direction is similar.

Table 6: Average accuracy in elicitation instructions quiz

	First attempt	First or second attempt
(Intercept)	0.69^{***}	0.85^{***}
	(0.05)	(0.03)
BSR	-0.06^{*}	-0.01
	(0.03)	(0.02)
BDM	-0.11^{**}	-0.03^{*}
	(0.03)	(0.02)
Num. obs.	494	494
Includes covariates?	Yes	Yes

Note: BSR vs BDM: Column 1 - $\chi^2(1,494) = 2.345, p = 0.126$; Column 2 - $\chi^2(1,494) = 1.295, p = 0.255$.The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Second, we consider participants' answers to the perceived complexity question which was measured on a similar 5-point Likert scale as perceived difficulty. Table 7 shows almost identical results to those for perceived difficulty. These results are in line with Charness, Gneezy and Rasocha (2021) who rank BDM and BSR as complex mechanisms and INTRO as less complex.

Table 7: Average perceived complexity of instructions across elicitation methods

	Simple	Controlling for Numeracy
(Intercept)	2.01^{***}	2.04^{***}
	(0.22)	(0.22)
BSR	0.43^{***}	0.43^{***}
	(0.12)	(0.12)
BDM	0.59^{***}	0.59^{***}
	(0.12)	(0.12)
High Numeracy		-0.08
		(0.10)
Num. obs.	494	494
Includes covariates?	Yes	Yes

Note: The dependent variable is measured on a 1-5 Likert scale. The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Comprehension. We further explore alternative difficulty measures and additional differences in the potential costs of implementation across methods by comparing the rates of correctly answering the various comprehension quizzes we implemented. Table 8 shows participant comprehension as measured by pass rates for each quiz and treatment condition. Participants answered three quizzes related to the instructions of the Bucket Game, general belief reporting, and the incentivization procedure. The quizzes are identical for the Bucket Game and general belief reporting. In the incentives quiz, four questions are unique and specific to each incentives method and two additional questions are common across the three conditions.

Participant comprehension in the Bucket Game quiz was high, as shown by Column 1. Across all three conditions, the rate of passing each question was above 2/3 (67%) after one attempt. This rate increases to almost 90% after the second attempt. Participant comprehension was even higher on the belief quiz, with a pass rate of over 85% on the first attempt. There is no significant difference in the pass rate for these quizzes across treatment conditions. This is expected since there are no differences across conditions up until this point,

The likelihood of answering correctly from the first try any of the questions related to the

Table 8: Comprehension Pass Rate (%) By Treatment and Quiz Type

	Instructions Quiz				
	Bucket	Belief	Incentives		
			Overall	Common	Specific
Treatment Condition					
INTRO					
Pass Likelihood (One Trial)	71.80	89.63	79.07	73.17	82.01
Pass Likelihood (Two Trials)	85.37	99.70	91.06	89.63	91.77
BDM					
Pass Likelihood (One Trial)	75.30	89.67	68.46	74.25	65.57
Pass Likelihood (Two Trials)	87.87	100.00	87.92	86.83	88.47
BSR					
Pass Likelihood (One Trial)	68.83	85.96	72.63	72.84	72.53
Pass Likelihood (Two Trials)	86.27	99.85	89.71	88.58	90.28

incentives method was significantly lower for the incentive-compatible procedures (BDM and BSR) compared to INTRO. This pass rate is 68.46% for BDM and 72.63% for BSR, compared to 79.07% for INTRO (BSR vs. INTRO: $t(324) = 2.26, p = 0.025$; BDM vs. INTRO: $t(329) = 3.65, p = 0.000$; BSR vs. BDM: $t(327) = -1.44, p = 0.151$). Breaking this down further, the same pass rate for common elicitation questions is not different across conditions (BSR vs. INTRO: $t(324) = 0.08, p = 0.934$; BDM vs. INTRO: $t(329) = 0.283, p = 0.777$; BSR vs. BDM: $t(327) = -0.38, p = 0.704$). However, for the elicitation questions that were uniquely specific to each treatment condition the first try pass rate is lowest for BDM with a value of 65.57%, followed by 72.53% for BSR and 82.01% for INTRO (BSR vs. INTRO: $t(324) = 3.08, p = 0.002$; BDM vs. INTRO: $t(332) = 5.13, p = 0.000$; BSR vs. BDM: $t(329) = -2.14, p = 0.033$). This suggests that participants had more difficulty in understanding the BDM and BSR methods than compared to INTRO, as expected given our pre-registered analysis.

4.2.2 Implementation speed

Table 9 presents the comparison of the total time (in seconds) it took participants to complete the experiment across the three treatments. This includes reading instructions, answering comprehension questions, and completing the belief elicitation task. The BDM treatment condition took over three additional minutes longer, on average, to complete the experiment compared to INTRO, and BSR took approximately two extra minutes on average to complete the experiment. The difference between BDM and BSR is not significant ($\chi^2(1,494) = 1.518, p = 0.218$). We therefore find partial support for H2 since only BDM takes significantly longer to implement than INTRO by an average of 269 seconds.

Table 9: Average time to complete experiment (in seconds) across elicitation methods

	Pre-registered	Controlling for Numeracy
(Intercept)	915.11***	874.46***
	(172.471)	(176.40)
BSR	124.98	128.78
	(96.710)	(96.75)
BDM	269.06*	270.72*
	(95.834)	(95.83)
High Numeracy		89.31
		(81.64)
Num. obs.	494	494
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. The regression model controlling for numeracy in column 2 is not pre-registered. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Result 2. *The overall implementation duration of BDM is significantly larger than that of INTRO. BSR is also taking longer to implement than INTRO, but the difference is not significant.*

We also investigate whether any timing differences arise for different sub-parts of the experiment. This analysis was not pre-registered and the corresponding statistical test values are presented in Table A8 in the Appendix. When separating the total duration between time allocated to reading the instructions and answering the comprehension questions, and time allocated to reporting beliefs, we find that the difference is largely due to longer time spent reading the instructions and completing comprehension questions, with BDM taking over three minutes longer and BSR taking less than two minutes longer to complete the comprehension section compared to INTRO. When omitting questions that are common across treatment conditions from the analysis, we find that completing the BDM and BSR comprehension questions is more time consuming than INTRO comprehension questions, taking approximately 40-60 seconds longer to complete. For answering questions that are common across treatment conditions, we find no significant difference in duration between BDM and BSR (155 to 153 seconds respectively), but somewhat of a decline for implementation of INTRO (137.29 seconds). This may suggest that participants are less cognitively burdened and thus able to answer comprehension questions more quickly because there is less information to learn.

When considering the time taken to elicit a single belief in the bucket game, we do not find significant differences between treatments. On average, it takes participants 13.28 seconds to complete the elicitation task in the incentivized treatments (BDM and BSR), compared to 10.98 seconds in the unincentivized treatment (INTRO). There is no significant difference in average task duration between BDM and BSR (11.26 seconds compared to 11.16 seconds respectively). Each belief elicitation (prior or posterior) takes, on average, 12.5 seconds to complete. However, this average response time hides meaningful variation across conditions and participant characteristics as can be seen in Table 10.

Table 10: One Belief Elicitation Duration (in seconds)

	No	Yes	t-test p-value
Second Elicitation?	13.50	11.70	.380
High Probabilistic Reasoning?	9.98	15.63	.006
Posterior Belief?	10.59	14.08	.088
High Subjective Difficulty?	13.40	11.06	.269

Notably, the average time does not decrease significantly for the second elicitation within a task, decreasing by only 1.8 seconds on average, suggesting participants become only slightly more efficient as they gain familiarity with the elicitation process. However, for participants identified with higher levels of probabilistic reasoning skills, the average elicitation time increases significantly to 15.6 seconds compared to 9.98 seconds, likely reflecting the additional cognitive effort invested in carefully evaluating probabilities. When comparing across types of elicitation, posterior belief elicitation tends to take longer amounts of time than prior belief elicitations, with an average response time of 14 seconds compared to 10.6 seconds, indicating the additional complexity and cognitive demand associated with updating beliefs after a signal is drawn. Finally, for participants who report higher difficulty with the instructions, we see less time on average devoted on each page, for an average of 11.1 seconds compared to 13.4 for participants who reported low difficulty with understanding the instructions, suggesting the use of cognitive heuristics as a result of the increased perceived complexity.

4.2.3 Belief accuracy

So far, we have established that incentive-compatible methods come with higher implementation costs, both in terms of potentially higher participant frustration and confusion due to the more complex instructions, and in terms of longer implementation speed which should be compensated for via higher monetary payments. This may all be worth it if the data quality obtained is superior. Unfortunately, this is not what we find.

Figure 2 presents the average error size of the elicited priors across treatments and elicitations, for each objective priors. We notice that the incentive compatible methods *never* lead to lower average errors than INTRO, and most often, the errors of the priors elicited under BSR or BDM are higher than in INTRO.

Indeed, using the pre-registered regression analysis to compare the average absolute error size across treatments, we find that participants were less accurate using the incentivized methods (BDM and BSR). Table 11 shows that participants in BSR, had significantly and systematically larger absolute errors in prior beliefs compared to INTRO, both in the first and in the second elicitation. The larger errors in BDM appear less robust though, being significantly higher than in INTRO only in the second elicitation. As can be seen in figure 2, this is mainly because the average error size for INTRO and BSR decreases in the second elicitation compared to the first, while

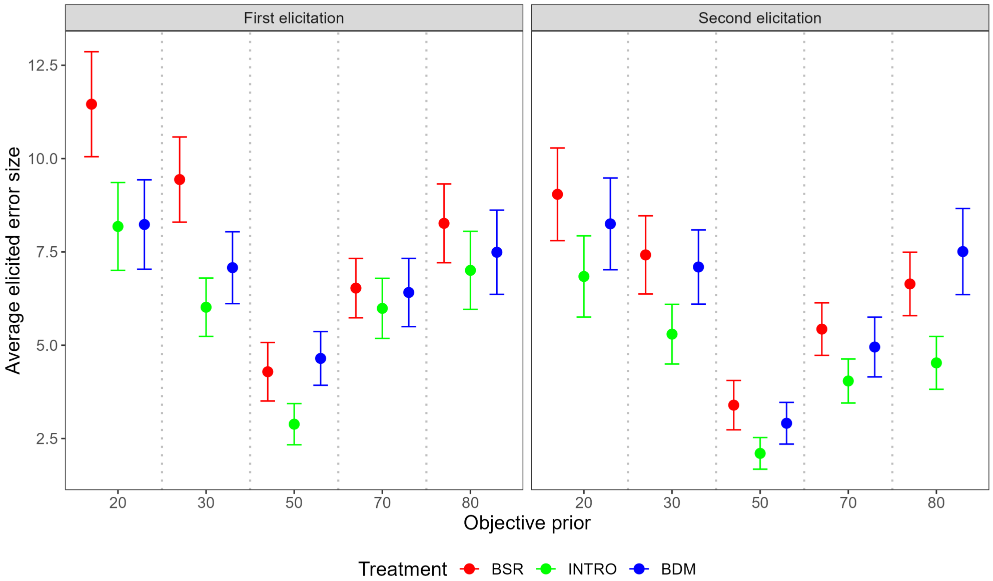


Figure 2: *Average error size of elicited prior beliefs across treatments and elicitations.* Vertical lines represent standard error bars.

for BDM it remains largely unaffected by experience. Regression analysis confirms this (see table A10 in the Appendix).

Table 11: Average prior belief error size across elicitation methods

	First elicitation	Second elicitation
(Intercept)	7.67^{***}	7.27^{***}
	(1.23)	(1.11)
BSR	1.91^{**}	1.77^{***}
	(0.63)	(0.57)
BDM	0.78	1.63^{**}
	(0.62)	(0.56)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Result 3. *The average absolute error of the elicited priors is significantly and systematically higher in BSR compared to INTRO. BDM also leads to larger absolute errors compared to INTRO but not in a systematic way.*

This is contrary to our third hypothesis predicting that incentive-compatible methods should lower the size of errors and increase the accuracy of reporting the objective (induced) priors. To better understand whether this finding is due to a few people reporting large errors or to a more general tendency to deviate from the objective prior, we investigate the frequency of inaccurate elicited priors. Figure 3 presents these results across treatments and elicitations for each objective prior. We notice almost identical frequencies between BDM and INTRO, supporting the observations that increased deviations in BDM are less systematic. BSR, however, leads to significantly higher frequencies (of about 6-7 percentage points) of inaccurate elicited priors compared to INTRO (see Table A9 in Appendix).

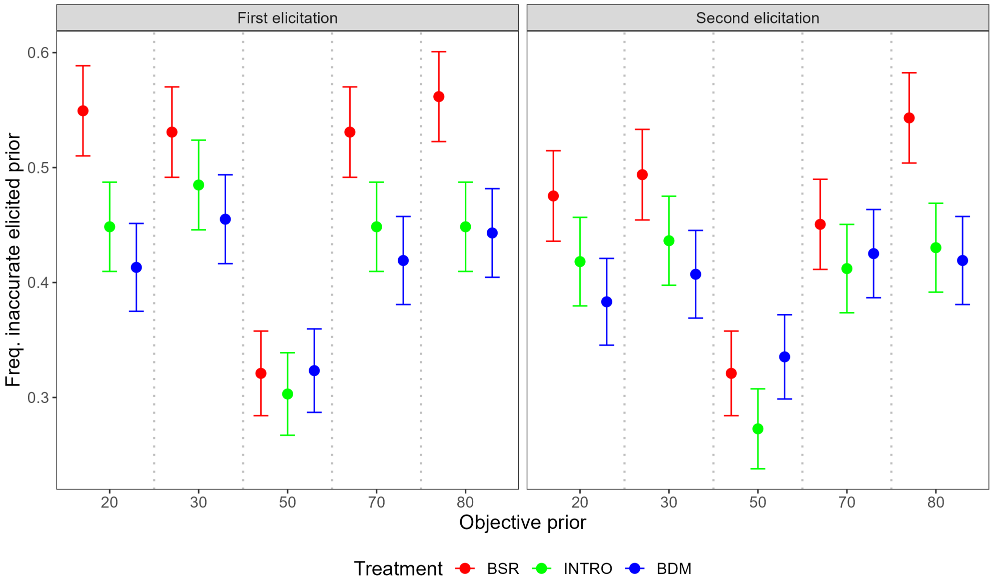


Figure 3: *Frequency of inaccurate elicited prior beliefs across treatments and elicitations.* Vertical lines represent standard error bars.

What also stands out from both Figure 2 and Figure 3 is that the average error size and the frequency of inaccurate elicited priors is lowest when the objective prior is 0.5. This could be due to participants reporting central beliefs more often when deviating from the objective prior as previously observed in Danz, Vesterlund and Wilson (2022) for the BSR method, which would mechanically lower the frequency of inaccurate reports when the objective prior is 0.5. We analyze the validity of this explanation in the next section.

4.2.4 Central-biased beliefs

Burdea and Woon (2022) found that incentive-compatible methods lead to a lower frequency of 50% beliefs compared to a flat incentive scheme, with BSR doing so in a more systematic way than BDM. There, unlike in this study, the objective probabilities participants were asked about were either unknown or less obvious to them. Nevertheless, given the cognitive nature of this effect, we based our hypothesis 4 on previous results rather than theoretical properties. Hence, in the present study we expected to find an effect of incentive compatible methods on beliefs about induced objective probabilities similar to that in Burdea and Woon (2022).

Opposite to our expectations, we find that participants were equally likely to report 50% beliefs in all treatments, for both the first and the second elicitation (see Table 12).

Table 12: Marginal effects from probit regression of prior belief being equal to 50%

	First elicitation	Second elicitation
BSR	0.01	0.02
	(0.01)	(0.01)
BDM	−0.00	0.00
	(0.01)	(0.01)
Num. obs.	1976	1976
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Result 4. *The frequency of central (50%) beliefs is not significantly different across the three elicitation methods (BSR, BDM and INTRO).*

The frequency of precisely centered beliefs is actually rather low and equal to approximately 3% across all treatments when we consider those objective priors different than 0.5. So, the higher frequency of reporting errors when the prior is different than 0.5 observed in figure 3 is not due to people simply reporting 50% when in doubt. While reporting a 50% belief has a natural interpretation of reporting uncertainty when the objective prior is not obvious as in Burdea and Woon (2022), this interpretation fits less well in the current study. Moreover, this measure may be too coarse to identify any pull-to-center effect of information on the incentives documented by Danz, Vesterlund and Wilson (2022) for the BSR method. Following that paper, to better understand the fact that objective priors different than 0.5 are associated with higher reporting errors and the corresponding treatment differences, we check for the distribution of beliefs that are different than the objective prior (π_0) within three categories: (a) centered: in between the objective prior and 0.5; (b) near extreme: in between the nearest extreme (1 or 0) and the objective prior; and (c) distant extreme: in between the most distant extreme and the objective prior. The results are presented in Table 13. The first three columns include a numerical summary of Figure 3, where we only distinguish between cases where the objective prior is equal to 0.5 or not. The last three columns depict the distribution of the error in the reported beliefs among the three mutually-exclusive categories.

Table 13: Frequency of inaccurate prior beliefs across treatments

		Inaccurate beliefs frequency			Inaccurate belief type ($\pi_0 \neq 0.5$)		
Elicitation	Treatment	All Priors	By Prior		Center	Near	Distant
			$\pi_0 = 0.5$	$\pi_0 \neq 0.5$		Extreme	Extreme
	INTRO	0.427	0.303	0.458	0.153	0.226	0.079
First	BSR	0.499	0.321	0.543	0.213	0.230	0.100
	BDM	0.411	0.323	0.433	0.169	0.168	0.096
	INTRO	0.394	0.273	0.424	0.150	0.218	0.056
Second	BSR	0.457	0.321	0.491	0.218	0.205	0.068
	BDM	0.394	0.335	0.409	0.172	0.142	0.094

We notice that the large proportion of errors are distributed in the vicinity of the true prior rather than at the distant extreme. Comparing the distribution between the center and the near extreme categories across treatments we observe that in INTRO, participants' errors are more likely to occur in the near extreme direction than towards the center. This is a similar result to the No Information treatment of Danz, Vesterlund and Wilson (2022) that is closest to our INTRO treatment. For the BSR and BDM treatments, the two types of errors tend to occur in almost equal proportion. Hence, we do not find evidence for the pull-to-center effect that was observed in the treatment in Danz, Vesterlund and Wilson (2022) where participants received information on the BSR mechanism, which would be the most comparable to our BSR treatment.

Consequently, while incentive compatible methods seem to increase the average error size in the reported beliefs compared to a flat incentive scheme, they do not seem to lead to a particular bias towards certain types of erroneous reports in our online environment. Next we analyze how individual differences in cognition interact with the effect of incentive-compatibility.

4.2.5 Perceived difficulty and belief accuracy

One reason why incentive compatible methods may increase reporting errors is because people may misunderstand or be confused by their mathematical structure. We proxy this by the reported perceived difficulty, assuming that participants who report higher difficulty (higher than the median) in understanding the incentive scheme are more likely to be confused by it, and hence, report beliefs with higher errors. The results of this analysis are presented in Table 14.[4]

Table 14: Prior belief error size and perceived difficulty (median-split)

	First elicitation	Second elicitation
(Intercept)	7.47^{***}	7.25^{***}
	(1.28)	(1.15)
BSR	2.31^{*}	1.31
	(0.89)	(0.80)
BDM	1.45	2.83^{***}
	(0.88)	(0.79)
High Difficulty	0.26	0.03
	(0.92)	(0.83)
BSR x High Difficulty	-0.78	0.78
	(1.29)	(1.16)
BDM x High Difficulty	-1.25	-2.08
	(1.29)	(1.16)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

We do not find support for our fifth hypothesis, neither in the first nor the second elicitation. If

[4] We pre-registered to run this analysis using the continuous version of the difficulty measure. This approach does not accurately control for the differences in average difficulty across conditions. Therefore, for the main analysis we prefer using the median split. The results using the continuous version are presented in the Appendix and are largely similar to those in Table 14.

anything, higher perceived difficulty leads to a *decrease* in the error size associated with incentive-compatible methods as the interaction terms are negative. The largest reduction in prior error size is observed in the second elicitation of the BDM method. This effect is significantly different than the one associated with the interaction between the BSR method and difficulty ($\chi^2(1,2470) = 5.283, p = 0.022$). We find very similar (and somewhat stronger) results also when testing the interaction with our subjective complexity question (see Table A11 in the Appendix). This points to a potential advantage of the BDM method over the BSR one for some participants.

Result 5. *Reporting a higher perceived difficulty of the BDM method is associated with a reduction in the average size of the prior error compared to the BSR method, which is significant in the second elicitation. No significant interaction effect is observed for the two incentive compatible methods when compared to INTRO.*

We also check whether the frequency of reporting inaccurate priors differs depending on participants' perceived difficulty. The results are presented in Table 15 and further strengthen the previous findings regarding the positive effect of the BDM method for participants that perceive it as having higher difficulty. In particular, these participants are 11 percentage points less likely to report an inaccurate prior compared to those in INTRO in the first elicitation. This coefficient decreases in the second elicitation to 9 percentage points and becomes insignificant. The difference in the interaction coefficient between BSR and BDM is though significantly different both in the first ($\chi^2(1,2470) = 4.575, p = 0.032$) and the second elicitation ($\chi^2(1,2470) = 8.990, p = 0.003$).

The results regarding the sensitivity of the accuracy of elicited priors to perceived difficulty of the elicitation method could suggest that participants who report that the elicitation methods are more difficult, may, in fact, be the ones who understand the incentive structure and respond positively to it.

4.2.6 Probabilistic reasoning (numeracy) and belief accuracy

A second cognitive measure that we expected to influence the effect of incentive compatible methods is probabilistic reasoning. In particular, according to hypothesis 6, we expected that partici-

Table 15: Marginal effects from probit regression of prior belief being inaccurate

	First elicitation	Second elicitation
BSR	0.06	0.03
	(0.34)	(0.03)
BDM	0.04	0.05
	(0.34)	(0.03)
High Difficulty	0.05	0.01
	(0.32)	(0.03)
BSR x High Difficulty	-0.01	0.05
	(0.05)	(0.05)
BDM x High Difficulty	-0.11^{*}	-0.09
	(0.05)	(0.05)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

pants with high probabilistic reasoning will respond positively to incentive-compatible methods as they should have a higher ability to understand the underlying mechanism behind these methods. Recall that each participant's probabilistic reasoning was measured by the 9-item questionnaire from Primi et al. (2017). The median score was 7, with only 22% of participants answering all 9 questions correctly. Given our pre-registered categorization based on the median-split, participants with a numeracy score strictly greater than 6 were classified as having high probabilistic reasoning and those with a score of lower or equal to 6 as having low probabilistic reasoning. Overall, 40% of participants were classified as low probabilistic reasoning. Table 16 presents the results of the regression analysis for our sixth hypothesis.

First we find a large and significant effect of being categorized as having high numeracy (by scoring greater than the median score (7) on the numeracy questionnaire). In particular, these participants have an average error size in their reported beliefs of 6-7 points lower than those categorized as having low numeracy skills, both in the first and in the second elicitation.

Next, focusing on the interaction effects we find that high numeracy is systematically associated with a lower average error size in the BDM treatment compared to INTRO and this reduction is significant in the second elicitation. This is not the case for the BSR treatment. In fact, the interaction

Table 16: Prior belief error size and probabilistic reasoning (numeracy)

	First elicitation	Second elicitation
(Intercept)	10.69^{***}	9.71^{***}
	(1.27)	(1.15)
BSR	1.84^{*}	1.19
	(0.94)	(0.85)
BDM	1.83	3.49^{***}
	(0.94)	(0.85)
High Numeracy	-7.01^{**}	-5.91^{**}
	(0.87)	(0.78)
BSR x High Numeracy	-0.40	0.56
	(1.22)	(1.10)
BDM x High Numeracy	-1.97	-3.30^{**}
	(1.21)	(1.10)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

coefficient between BDM and high numeracy is significantly different in the second elicitation also when compared to that of BSR interacted with high numeracy ($\chi^2(1,2470) = 9.331, p = 0.003$).

Result 6. *Higher probabilistic reasoning led to a lower average error size in prior beliefs in BDM compared to INTRO and BSR which is significant in the second elicitation. Probabilistic reasoning does not matter when comparing the BSR method with INTRO.*

To better understand whether BDM's apparent advantage for high numeracy participants is due to these participants making smaller-sized errors and/or whether fewer of them are reporting inaccurate beliefs, we also run a probit analysis of the frequency of reporting a belief different than the objective prior. The results reported in Table 17 show that high numeracy individuals in the BDM treatment do not make significantly fewer reporting errors compared to INTRO, but they do so when compared to BSR both in the first elicitation ($\chi^2(1,2470) = 3.925, p = 0.048$) and in the second ($\chi^2(1,2470) = 10.869, p = 0.001$).

Table 17: Marginal effects from probit regression of prior belief being inaccurate

	First elicitation	Second elicitation
BSR	0.05	0.01
	(0.04)	(0.03)
BDM	0.02	0.04
	(0.03)	(0.03)
High Numeracy	-0.31^{***}	-0.32^{***}
	(0.03)	(0.78)
BSR x High Numeracy	0.02	0.07
	(0.05)	(0.04)
BDM x High Numeracy	−0.07	−0.08
	(0.05)	(0.05)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

4.3 Secondary analysis - posteriors

We rerun the main analysis with the posterior beliefs as dependent variables to investigate the quality of belief updating across the different incentive schemes. Since there is no previous literature on how elicitation mechanisms affect belief updating, our best guess is that similar patterns for prior beliefs will also be observed for posterior beliefs. We use Bayes' rule to derive the benchmark for accurate posterior beliefs (rounded to the nearest integer).

The first observation is that the posterior belief data are much more noisy: 94% of posterior beliefs are inaccurate and the average deviation from the Bayesian benchmark is 17 percentage points. However, we find patterns similar to those for prior beliefs also for the reports of posterior beliefs. The corresponding regression tables are in the Appendix.

In particular, the average posterior error size is larger in the first elicitation of posteriors in BDM and BSR than in INTRO, but the coefficient is significant only for the BSR treatment. The errors decrease across all methods in the second elicitation and no significant difference between treatments is observed. No significant treatment differences are observed in the likelihood of reporting central beliefs.

Regarding the interaction with perceived difficulty, we find that the error size (but not the error frequency) of posterior beliefs is significantly affected by whether the participant perceived the method as highly difficult. The direction of the effect is the same as for prior beliefs but larger such that now, perceiving the method as highly difficult significantly increases the average error size. However, this increase is eliminated in the BSR and BDM treatments. This is valid only for the first elicitation though as any significant differences disappear in the second elicitation of posteriors.

Finally, with respect to the effect of numeracy (probabilistic reasoning), we find, as for prior beliefs, a very strong decrease in the average posterior error for participants with high numeracy, but no significant interaction with the incentive-compatible elicitation methods.

5 Discussion

This study aimed to evaluate the performance of three belief elicitation methods—the Binarized Scoring Rule (BSR), the stochastic Becker-DeGroot-Marschak mechanism (BDM), and unincentivized introspection (INTRO) in an online experimental setting. Using a between-subjects design, we investigated how these methods influence the accuracy of elicited beliefs about objective probabilities, the perceived difficulty of each method, and the implementation time. Besides this being the first comparison of the BDM and the BSR mechanisms for induced probabilities, our study contributes to the literature by focusing on the online research market which has become one of (if not the) primary research ground in experimental economics. There are many ways in which the online settings differs from the controlled, lab environments with undergraduate student samples. This makes extending the results drawn from lab studies troublesome, a view expressed in Charness, Gneezy and Rasocha (2021) and Svorenčík and Maas (2016). Additionally, we incorporated measures of probabilistic reasoning to examine the interaction between participants' cognitive abilities and the efficacy of belief elicitation methods.

Our findings challenge the assumption that incentivized belief elicitation methods necessarily outperform simpler, unincentivized alternatives in online settings. While the BSR and BDM methods are theoretically robust and incentive-compatible, they were associated with higher perceived difficulty and longer implementation durations. Moreover, incentivized methods did not outperform introspection in terms of belief accuracy. In fact, the BSR method systematically induced larger errors, particularly during the initial rounds of belief elicitation. Errors under BDM were less consistent but still generally higher than introspection. Interestingly, participants who perceived BDM as more difficult appeared to engage more cognitive effort, which reduced the frequency of prior belief errors and the error size of posterior beliefs. This was not the case for BSR, where higher perceived difficulty did not translate into improved performance.

Furthermore, individual differences in probabilistic reasoning significantly influenced outcomes. Participants with high probabilistic reasoning skills performed better with BDM, demonstrating lower errors compared to those with lower skills. However, this advantage was not observed with BSR, which performed poorly across all participant groups. 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. They also 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. Our probabilistic reasoning measure, which significantly moderates the difference between BDM, BSR, and INTRO, may be better tailored to the type of cognitive sophistication necessary to understand these complex methods, rather than the cognitive measure Burfurd and Wilkening (2021) used (a type of Cognitive Reflection Test) which was intended to capture how willing people are to think through the task and arrive at an answer.

Our findings have several implications for the design and selection of belief elicitation methods in online experiments. First, despite its lack of incentive compatibility, introspection demonstrated strong performance in terms of belief accuracy and participant comprehension. Its simplicity and cost-effectiveness make it a viable option for studies where incentivized mechanisms may overcomplicate tasks without yielding better data quality. Another potential method, not covered by our horserace, is the frequency method. This method has a significantly simpler structure (Charness, Gneezy and Rasocha, 2021) and, in our exploratory survey, turned out to be perceived as having low likelihood of being the basis of a journal rejection. Nevertheless, our experts were somewhat less familiar with it. Moreover, this method is not applicable for all types of probabilistic beliefs.

A second implication of our findings relates to the interaction between cognitive ability and method efficacy. This underscores the importance of tailoring belief elicitation methods to the characteristics of the target population. For populations with lower probabilistic reasoning skills, simpler methods like introspection may be more appropriate, while BDM could be more effective for cognitively sophisticated participants. Therefore, given the higher implementation costs of BSR and BDM, researchers should carefully consider whether the added complexity is justified by the research objectives and target audience.

Several avenues for future research emerge from our study. First, the development of simplified incentivized methods that maintain theoretical robustness while reducing complexity could improve data quality in diverse participant pools. Second, targeted training interventions to improve participants' understanding of complex elicitation mechanisms may help bridge the gap between theoretical advantages and practical outcomes. This, however, needs to take into account the time investment necessary to complete such training. Finally, further exploration of belief elicitation in less controlled environments could provide deeper insights into the trade-offs between accuracy, complexity, and participant comprehension.

References

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

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.

Banovetz, James, and Ryan Oprea. 2023. "Complexity and procedural choice." *American Economic Journal: Microeconomics*, 15(2): 384–413.

Benjamin, Daniel J. 2019. "Errors in probabilistic reasoning and judgment biases." *Handbook of Behavioral Economics: Applications and Foundations 1*, 2: 69–186.

Brace, Paul, Kellie Sims-Butler, Kevin Arceneaux, and Martin Johnson. 2002. "Public Opinion in the American States: New Perspectives Using National Survey Data." *American Journal of Political Science*, 46(1): 173.

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

Burdea, Valeria, and Jonathan Woon. 2022. "Online belief elicitation methods." *Journal of Economic Psychology*, 90: 102496.

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*, 1–36.

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

Danz, David, Lise Vesterlund, and Alistair J Wilson. 2022. "Belief elicitation and behavioral incentive compatibility." *American Economic Review*, 112(9): 2851–83.

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.

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

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. 2014. "Eliciting subjective probabilities with binary lotteries." *Journal of Economic Behavior & Organization*, 101: 128–140.

Hollard, Guillaume, Sébastien Massoni, and Jean-Christophe Vergnaud. 2016. "In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments." *Theory and Decision*, 80(3): 363–387.

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): 110–139.

Hossain, T., and R. Okui. 2013. "The Binarized Scoring Rule." *The 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.

Matousek, Jindrich, Tomas Havranek, and Zuzana Irsova. 2022. "Individual discount rates: a meta-analysis of experimental evidence." *Experimental Economics*, 25(1): 318–358.

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

Oprea, Ryan. 2020. "What makes a rule complex?" *American economic review*, 110(12): 3913–3951.

Plott, Charles R. 1986. "Rational Choice in Experimental Markets." *The Journal of Business*, 59(S4): S301.

Primi, Caterina, Kinga Morsanyi, Maria Anna Donati, Silvia Galli, and Francesca Chiesi. 2017. "Measuring probabilistic reasoning: The construction of a new scale applying item response theory." *Journal of Behavioral Decision Making*, 30(4): 933–950.

Primi, Caterina, Maria Anna Donati, Sara Massino, Elisa Borace, Edoardo Franchi, and Kinga Morsanyi. 2019. "Measuring probabilistic reasoning: the development of a brief version of the Probabilistic Reasoning Scale (PRS-B)." Freudenthal Group; Freudenthal Institute; ERME.

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

Schlag, Karl, and James Tremewan. 2021. “Simple belief elicitation: An experimental evaluation.” *Journal of Risk and Uncertainty*.

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.” *Annu. Rev. Econ.*, 6(1): 103–128.

Smith, Vernon L, and James M Walker. 1993. “Monetary rewards and decision cost in experimental economics.” *Economic Inquiry*, 31(2): 245–261.

Svorenčík, Andrej, and Harro Maas, ed. 2016. *The Making of Experimental Economics: Witness Seminar on the Emergence of a Field.* Cham:Springer International Publishing.

Trautmann, Stefan T, and Gijs van de Kuilen. 2014. “Belief elicitation: A horse race among truth serums.” *The Economic Journal*, 125(589): 2116–2135.

Vespa, Emanuel, and Alistair J Wilson. 2016. “Communication with multiple senders: An experiment.” *Quantitative Economics*, 7(1): 1–36.

Wang, Stephanie W. 2011. “Incentive effects: The case of belief elicitation from individuals in groups.” *Economics Letters*, 111(1): 30–33.

A Horserace of Methods for Eliciting Induced Beliefs Online

Daniel Banko-Ferran Valeria Burdea Jonathan Woon

Appendix

Appendix A Additional tables

A.1 Exploratory expert survey

The tabulation of results for the QSR and frequency methods are presented in the following tables.

		QSR	**FREQ**
Familiarity	Not at all	0	0.2
	Slightly – Moderately	0.12	0.44
	Extremely – Very	0.88	0.36
Use as primary outcome	Never	0.36	0.72
	Rarely – Sometimes	0.56	0.2
	Often	0.08	0.08
Use as secondary outcome	Never	0.52	0.72
	Rarely – Sometimes	0.56	0.2
	Often	0.08	0.08

Table A1: Tabulation of responses to question: "How familiar are you with the XX method for belief elicitation?" and "How often do you use the XX method for belief elicitation when beliefs are a primary (secondary) outcome variable in your work? N = 25. QSR = Quadratic Score Rule, FREQ = Frequency/Interval method.

		QSR	**FREQ**
Top 5 – likelihood to reject as primary outcome	Very/somewhat unlikely	0.6	0.68
	Neither unlikely nor likely	0.12	0.24
	Very/somewhat likely	0.28	0.08
Top 5 – likelihood to reject as secondary outcome	Very/somewhat unlikely	0.76	0.76
	Neither unlikely nor likely	0.16	0.2
	Very/somewhat likely	0.08	0.04
Non-top 5 (top field) – primary/secondary outcome	Very/somewhat unlikely	0.76	0.64
	Neither unlikely nor likely	0.08	0.28
	Very/somewhat likely	0.16	0.08

Table A2: Tabulation of responses to question: "Suppose you were reviewing a paper, for a top 5 economics journal [field journal (not a top 5)], that elicited beliefs from participants in an experiment. These beliefs are a primary [secondary] (primary or secondary) outcome variable. How likely would you be to reject the paper or require the experiment be re-run with a different method if the belief elicitation process used this method?". Online survey conducted in October 2022. N = 25. QSR = Quadratic Scoring Rule, FREQ = Frequency/Interval method.

A.2 Preliminaries

Table A3: Demographics

	N	%
Age		
18 - 24	21	4.26%
25 - 34	169	34.28%
35 - 44	169	34.28%
45 - 54	76	15.42%
55 - 64	40	8.11%
65 - 74	17	3.45%
75 - 84	1	0.20%
Gender		
Male	286	58.01%
Female	205	41.58%
Other	2	0.41%
Education level		
Less than High School	3	0.61%
High School / GED	73	14.81%
Some college	99	20.08%
2 year collge degree (Associate)	60	12.17%
4 year college degree (Bachelor)	208	42.19%
Post-graduate degree (Professional, Masters, Doctorate)	50	10.14%
Race/ethnicity		
White/Caucasian	392	79.51%
African American	50	10.14%
Hispanic	33	6.69%
Asian or Pacific Islander	46	9.33%
Native American	9	1.83%

Table A4: Sample summary statistics

	Elicitation treatment condition					
	INTRO		BDM		BSR	
	N	%	N	%	N	%
Age						p-val: 0.387
18 - 24	6	3.66%	8	4.79%	7	4.32%
25 - 34	60	36.59%	46	27.54%	63	38.89%
35 - 44	58	35.37%	61	36.53%	50	30.86%
45 - 54	17	10.37%	34	20.36%	25	15.43%
55 - 64	15	9.15%	13	7.78%	12	7.41%
65 - 74	7	4.27%	5	2.99%	5	3.09%
75 - 84	1	0.61%	0	0.00%	0	0.00%
Gender						p-val: 0.872
Male	97	59.15%	97	58.08%	92	56.79%
Female	66	40.24%	69	41.32%	70	43.21%
Other	1	0.61%	1	0.60%	0	0.00%
Education level						p-val: 0.731
Less than High School	2	1.22%	1	0.60%	0	0.00%
High School / GED	27	16.46%	25	14.97%	21	12.96%
Some college	34	20.73%	28	16.77%	37	22.84%
2 year collge degree	20	12.20%	20	11.98%	20	12.35%
4 year college degree	69	42.07%	75	44.91%	64	39.51%
Post-graduate degree	12	7.32%	18	10.78%	20	12.35%
White/Caucasian						p-val: 0.062
0	30	18.29%	28	16.77%	43	26.54%
1	134	81.71%	139	83.23%	119	73.46%
African American						p-val: 0.656
0	147	89.63%	157	94.01%	139	85.80%
1	17	10.37%	10	5.99%	23	14.20%
Hispanic						p-val: 0.656
0	155	94.51%	156	93.41%	149	91.98%
1	9	5.49%	11	6.59%	13	8.02%
Asian or Pacific Islander						p-val: 0.321
0	146	89.02%	156	93.41%	145	89.51%
1	18	10.98%	11	6.59%	17	10.49%
Native American						p-val: 0.701
0	162	98.78%	164	98.20%	158	97.53%
1	2	1.22%	3	1.80%	4	2.47%

Table A5: Difference in mean between first and second elicitation

		Prior	Prior	Prior	Prior	Prior
		.2	.3	.5	.7	.8
INTRO	(p-val)	0.19	0.23	0.51	0.70	0.05
BDM	(p-val)	0.74	0.28	0.88	0.64	0.31
BSR	(p-val)	0.06	0.20	0.13	0.11	0.00

Table A6: Difference in variance between first and second elicitation

		Prior	Prior	Prior	Prior	Prior
		.2	.3	.5	.7	.8
INTRO	(p-val)	0.10	0.96	0.00	0.00	0.00
BDM	(p-val)	0.74	0.89	0.00	0.00	1.00
BSR	(p-val)	0.01	0.03	0.00	0.02	0.00

Table A7: Difference in proportion of belief = 50 between first and second elicitation

		Prior	Prior	Prior	Prior	Prior
		.2	.3	.5	.7	.8
INTRO	(p-val)	0.00	0.00	0.00	0.00	0.00
BDM	(p-val)	0.00	0.00	0.00	0.00	0.00
BSR	(p-val)	0.00	0.00	0.00	0.00	0.00

A.3 Further analysis for prior beliefs

Table A8: Implementation duration - breakdown by experimental subpart

	1	2	3	pval(bdm==bsr)	pval(bdm==intro)	p(bsr==intro)
	bdm	bsr	intro			
Instructions	64.44	43.24	15.00	0.00	0.00	0.00
	(41.53)	(35.74)	(17.61)			
Comprehension (Common)	155.72	153.09	137.29	0.23	0.00	0.00
	(96.14)	(82.73)	(64.90)			
Comprehension (Unique)	163.03	121.65	94.88	0.00	0.00	0.00
	(84.66)	(62.97)	(52.03)			
Mean(Prior)—First Elicitation	12.48	12.69	12.78	0.67	0.55	0.86
	(9.71)	(10.13)	(10.96)			
Mean(Prior)—Second Elicitation	7.58	7.77	7.95	0.48	0.21	0.54
	(5.20)	(5.33)	(6.65)			
Mean(Posterior)—First Elicitation	12.35	12.22	10.86	0.82	0.00	0.01
	(10.69)	(10.92)	(10.03)			
Mean(Posterior)—Second Elicitation	8.11	7.60	7.98	0.12	0.71	0.24
	(7.16)	(5.97)	(7.26)			
Sum(Prior)—First Elicitation	61.56	62.26	62.89	0.6	0.34	0.66
	(26.56)	(27.14)	(29.97)			
Sum(Prior)—Second Elicitation	37.92	38.56	39.19	0.45	0.18	0.51
	(16.82)	(16.96)	(21.67)			
Sum(Posterior)—First Elicitation	60.62	59.98	53.76	0.7	0.00	0.00
	(33.56)	(34.31)	(33.39)			
Sum(Posterior)—Second Elicitation	40.38	37.48	39.48	0.01	0.5	0.10
	(25.98)	(21.93)	(27.49)			

All values are means (SD); p-values from pairwise t-tests.

Table A9: Marginal effects from probit regression of prior belief being inaccurate

	First elicitation	Second elicitation
BSR	0.07^{**}	0.06^{**}
	(0.02)	(0.02)
BDM	-0.01	0.00
	(0.02)	(0.02)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: BSR vs BDM: Column 1 - $\chi^2(1,2470) = 12.192, p = 0.000$; Column 2 - $\chi^2(1,2470) = 6.127, p = 0.013$. The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A10: Average prior belief error size across elicitation rounds

	INTRO	BSR	BDM
(Intercept)	4.90***	7.95***	14.76***
	(1.18)	(1.55)	(1.51)
Second elicitation	−1.45*	−1.61*	−0.63
	(0.52)	(0.63)	(0.62)
Num. obs.	1650	1620	1670
Urn FE	Yes	Yes	
Includes covariates?	Yes	Yes	

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A11: Average prior belief error size across treatments and perceived complexity

	First elicitation	Second elicitation
(Intercept)	7.47***	6.91***
	(1.30)	(1.17)
BSR	1.41	0.58
	(0.98)	(0.89)
BDM	2.25*	4.01***
	(1.06)	(0.96)
High Complexity	0.16	0.52
	(0.88)	(0.80)
BSR x High Complexity	0.70	1.63
	(1.29)	(1.16)
BDM x High Complexity	−2.03	−3.39**
	(1.35)	(1.21)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

A.4 Regression tables for the analysis of posterior beliefs

Table A12: Average posterior belief error size across treatments

	First elicitation	Second elicitation
(Intercept)	14.28^{***}	12.93^{***}
	(1.58)	(1.52)
BSR	2.61^{**}	0.71
	(0.80)	(0.78)
BDM	0.93	−0.23
	(0.80)	(0.77)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A13: Marginal effects from probit regression of posterior belief being equal to 50%

	First elicitation	Second elicitation
BSR	−0.00	−0.01
	(0.01)	(0.01)
BDM	0.00	0.01
	(0.01)	(0.01)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The data includes cases where the objective prior was equal to 0.50. The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A14: Posterior belief error size and perceived difficulty (median-split)

	First elicitation	Second elicitation
(Intercept)	12.99^{***}	12.74^{***}
	(1.54)	(1.58)
BSR	3.92^{**}	0.41
	(1.14)	(1.10)
BDM	2.46^{*}	0.31
	(1.13)	(1.09)
High Difficulty	3.67^{**}	0.79
	(1.18)	(1.14)
BSR x High Difficulty	-3.64^{*}	0.20
	(1.66)	(1.60)
BDM x High Difficulty	-4.00^{*}	-1.22
	(1.65)	(1.59)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A15: Marginal effects from probit regression of posterior belief being inaccurate

	First elicitation	Second elicitation
BSR	0.02	0.04^{**}
	(0.02)	(0.02)
BDM	-0.02	0.03
	(0.01)	(0.02)
High Difficulty	0.02	0.02
	(0.02)	(0.02)
BSR x High Difficulty	-0.02	-0.06^{***}
	(0.03)	(0.02)
BDM x High Difficulty	0.02	-0.04
	(0.02)	(0.02)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A16: Posterior belief error size and probabilistic reasoning (numeracy)

	First elicitation	Second elicitation
(Intercept)	17.45^{***}	15.62^{***}
	(1.66)	(1.61)
BSR	2.57^{*}	-0.07
	(1.23)	(1.19)
BDM	1.75	0.92
	(1.24)	(1.20)
High Numeracy	-7.20^{**}	-6.25^{**}
	(1.14)	(1.10)
BSR x High Numeracy	-0.47	0.90
	(1.59)	(1.54)
BDM x High Numeracy	-1.59	-2.12
	(1.59)	(1.54)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table A17: Marginal effects from probit regression of posterior belief being inaccurate

	First elicitation	Second elicitation
BSR	-0.01	0.01
	(0.02)	(0.02)
BDM	-0.01	0.05
	(0.02)	(0.02)
High Numeracy	-0.05^{*}	0.01
	(0.02)	(0.02)
BSR x High Numeracy	0.03	0.00
	(0.03)	(0.02)
BDM x High Numeracy	-0.01	-0.07^{*}
	(0.03)	(0.03)
Num. obs.	2470	2470
Urn FE	Yes	Yes
Includes covariates?	Yes	Yes

Note: The covariates include gender, age and education level. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Appendix B Expert survey instructions

Consent

Welcome to this study!

This study is part of a research project about belief elicitation methods. We are interested in your opinions because you have published research in experimental economics, decision research, or have scholarly expertise in a related area.

It is expected to take approximately **10 minutes** to complete. All participants must be 18 years of age or older.

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 opinions. Your responses will not be associated with 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. Any incomplete responses will be deleted from our database.

The study is being conducted by researchers at the University of Pittsburgh and their colleagues. If you have any questions about the study, you may send an email to grpwoon@pitt.edu.
[I have read the above and consent to take part in this study; I do not wish to participate]

experiments_freq How often does your research involve conducting experiments with **human subjects**?
[1 – Never; 2; 3; 4; 5 – Always]

beliefs_freq How often does your research involve **belief elicitation**?
[1 – Never; 2; 3; 4; 5 – Always]

Next, we will present you with a brief description of some of the most frequently used methods for belief elicitation in experiments. We will then ask you a few questions regarding your experience with each of these methods.

METHODS QUESTIONS

BDM method

The stochastic Becker-DeGroot-Marshak mechanism is also referred to as probability matching, the crossover method, or the Karni mechanism (Allen, 1987; DuCharme and Donnell, 1973; Grether, 1992; Holt and Smith, 2009; Karni, 2009).

With this method, subjects report the belief for which they prefer getting a prize based on the accuracy of the expressed belief over getting this prize based on an objective lottery.

bdm_familiar How familiar are you with the **BDM** method for belief elicitation?
[1 – Not at all familiar; 2; 3; 4; 5 – Extremely familiar]

bdm_use_primary How often do you use the **BDM** method for belief elicitation when beliefs are a **primary outcome variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

bdm_use_secondary How often do you use the **BDM** method for belief elicitation when beliefs are a **secondary outcome or control variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

QSR method

Here we refer to the Quadratic Scoring Rule (Brier, 1950).

With this method, subjects receive a score for the reported belief about an event based on the Brier score (Brier, 1950), which is the sum of the squared errors of the reported probabilities. Higher scores correspond to a smaller “distance” of the belief from the realization of the event.

qsr_familiar How familiar are you with the **QSR** method for belief elicitation?
[1 – Not at all familiar; 2; 3; 4; 5 – Extremely familiar]

qsr_use_primary How often do you use the **QSR** method for belief elicitation when beliefs are a **primary outcome variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

qsr_use_secondary How often do you use the **QSR** method for belief elicitation when beliefs are a **secondary outcome or control variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

BSR method

Here we refer to the Binarized Scoring Rule (Hossain and Okui, 2013; Harrison et al, 2014). This method is like QSR, but the scoring rule determines the probability of receiving a fixed prize instead of the amount of a variable payment.

bsr_familiar How familiar are you with the **BSR** method for belief elicitation?
[1 – Not at all familiar; 2; 3; 4; 5 – Extremely familiar]

bsr_use_primary How often do you use the **BSR** method for belief elicitation when beliefs are a **primary outcome variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

bsr_use_secondary How often do you use the **BSR** method for belief elicitation when beliefs are a **secondary outcome or control variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

Frequency/interval method

This method elicits beliefs by offering a prize if the reported belief matches an underlying, objective probability (Schlag and Tremewan, 2021). In case the frequency is used, the prize is awarded if the reported belief matches exactly the objective probability. In case the interval is used, the prize is awarded if the reported belief is within some number of percentage points from the objective probability.

freq_familiar How familiar are you with the **frequency/interval** method for belief elicitation?
[1 – Not at all familiar; 2; 3; 4; 5 – Extremely familiar]

freq_use_primary How often do you use the **frequency/interval** method for belief elicitation when beliefs are a **primary outcome variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

freq_use_secondary How often do you use the **frequency/interval** method for belief elicitation when beliefs are a **secondary outcome or control variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

Introspection

When using this method, subjects are simply asked to report their belief about the likelihood of an event. This report is not tied to any potential monetary payment.

intro_familiar How familiar are you with the **introspection** method for belief elicitation?
[1 – Not at all familiar; 2; 3; 4; 5 – Extremely familiar]

intro_use_primary How often do you use the **introspection** method for belief elicitation when beliefs are a **primary outcome variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

intro_use_secondary How often do you use the **introspection** method for belief elicitation when beliefs are a **secondary outcome or control variable** in your work?
[1 – Never; 2; 3; 4; 5 – Always]

PREFERRED METHOD QUESTIONS

method_pref_multi If you were conducting a new study now, which method(s) would you use to elicit beliefs? Check all that you would be willing to implement in your own research:
[BDM; QSR; BSR; Frequency/interval; Introspection; Other; None of the above]

method_pref_uni If you had to choose only one method to use in a new study, which one would it be? [BDM; QSR; BSR; Frequency/interval; Introspection; Other; None of the above]

method_pref_explain Why do you prefer the method you specified in the previous question?

top5_primary Suppose you were reviewing a paper, for a **top 5 economics journal**, that elicited beliefs from participants in an experiment. These beliefs are **a primary outcome variable**.
For each of the methods below, how likely would you be to **reject the paper or require the experiment be re-run with a different method** if the belief elicitation process used this method?
BDM [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
QSR [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
BSR [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
Frequency/interval [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
Introspection [1 – Very unlikely; 2; 3; 4; 5 – Very likely]

top5_secondary
What if the elicited beliefs were instead for a **secondary outcome**?

That is, suppose again you were reviewing a paper, for a **top 5 economics journal**, that elicited beliefs from participants in an experiment. These beliefs are **a secondary outcome variable and not a primary one**.

For each of the methods below, how likely would you be to **reject the paper or require the experiment be re-run with a different method** if beliefs were elicited using this method?'
BDM [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
QSR [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
BSR [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
Frequency/interval [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
Introspection [1 – Very unlikely; 2; 3; 4; 5 – Very likely]

non_top5
What if you were reviewing a paper for a **field journal?**
That is, suppose you were reviewing a paper, for a **field journal (not a top 5)**, that elicited beliefs from participants in an experiment. These beliefs may be **primary or secondary outcome variables**.
For each of the methods below, how likely would you be to **reject the paper or require the experiment be re-run with a different method** if beliefs were elicited using this method?
BDM [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
QSR [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
BSR [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
Frequency/interval [1 – Very unlikely; 2; 3; 4; 5 – Very likely]
Introspection [1 – Very unlikely; 2; 3; 4; 5 – Very likely]

importance_criteria How important is each of the following criteria when choosing how to elicit beliefs in an experiment?
Incentive compatibility [1 – Not at all important; 2; 3; 4; 5 – Extremely likely]
Subjects' comprehension of the incentives compatibility [1 – Not at all important; 2; 3; 4; 5 – Extremely likely]
Length of time compatibility [1 – Not at all important; 2; 3; 4; 5 – Extremely likely]
Scope of belief (primary measure, secondary measure, control) compatibility [1 – Not at all important; 2; 3; 4; 5 – Extremely likely]

comments Do you have any other comments about the use of belief elicitation methods in experimental economics or decision research?

<u>DEMOGRAPHICS</u>

phd_year When did you receive your PhD?
[1 – 2016-2022; 2011-2015; 2001-2010; 1991-2000; 1990 or before; 6 – Not applicable]

gender What is your gender?
[1 – Female; 2 – Male; 3 – Non-binary]

university_country In which country is your current university or institution located?
[Afghanistan (1) ... Zimbabwe (1357)]

academic_position Which position do you have at your university or institution?
[1 – PhD student; Postdoc or research scientist; Faculty (without tenure); Faculty (with tenure);
5 – Other]

email If you are interested in receiving a summary of the findings of this survey or have other comments and suggestions for the study investigators, please email us at grpwoon@pitt.edu.

Appendix C Experiment instructions

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 $4 for successfully completing the study. It is also possible for some participants to earn an additional bonus of up to $8.00. 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]

Questions about your background

1. What is your gender?
    - [Male; Female; Other]
2. What is your age?
    - [Under 18; 18-24; 25-34; 35-44; 45-54; 55-64; 65-74; 75-84; 85 or older]
3. What is your race/ethnicity? (check all that apply)
    - [White/Caucasian; African American; Hispanic; Asian or Pacific Islander; Native American; Other]
4. 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)]

[ATTENTION CHECK]

Some people think the government should provide fewer services, even in areas such as health and education, 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)]

General Instructions

This study is composed of two parts. You will receive a detailed description of each part before it begins.

[NUMERACY]

Part 1

In this part, we will ask you a series of questions about chance events. Please try to be as accurate as possible.

1. A fair coin is tossed nine times. Which of the following sequences of outcomes is a more likely result of nine flips of the fair coin?
   - [THHTHTTHH; HTHTHTHTH; Both sequences are equally likely **[CORRECT]**]
2. A marble bag contains 15 blue and 15 green marbles. After you drew 5 marbles (the marble drawn was always put back into the bag), a sequence of 5 green marbles was obtained. What is the most likely outcome if a marble is drawn a sixth time?
   - [a green marble; a blue marble; blue and green are equally likely **[CORRECT]**]
3. A bingo game is played with 25 numbers (from 1 to 25). At the first draw, which of the following results is the most likely?
   - [It is more likely to be an even number; It is more likely to be an odd number **[CORRECT]**; It is just as likely to be an even or an odd number]
4. Two decks, labeled A and B, are composed of cards with a star (star cards) and cards without any figure (white cards) on the reverse side. Deck A contains 100 cards, 80 white cards and 20 with a star. Deck B contains 10 cards, 8 white cards and 2 with a star. After choosing one of the decks, you must draw a card (without peeking, of course). Which deck gives you a better chance of drawing a star card?
   - [Deck A (with 80 white and 20 star cards); Deck B (with 8 white and 2 star cards); Equal chances from each deck **[CORRECT]**]
5. A marble bag contains 10 blue and 20 green marbles. After you drew 5 marbles (the marble drawn was always put back into the bag), a sequence of 5 green marbles was obtained. What is the most likely outcome if a marble is drawn a sixth time?
   - [a green marble **[CORRECT]**; a blue marble; blue and green are equally likely]
6. 60% of the population in a city are men and 40% are women. 50% of the men and 30% of the women smoke. We select a person from the city at random. What is the probability that this person is a smoker?
   - [42% **[CORRECT]**; 50%; 85%]
7. According to a recent survey, 90% of the population in a city usually lie and 30% of those usually lie about important matters. If we pick a person at random from this city, what is the probability that the person usually lies about important matters?
   - [60%; 30%; 27% **[CORRECT]**]
8. In a choir there are 100 children: 30 boys and 70 girls. Half of the boys and 1 in 10 girls learn to play the piano. We select a child from the choir at random. What is the probability that he/she plays the piano?
   - [22 out of 100 **[CORRECT]**; 30 out of 100; 50 out of 100]

9. In a medical center a group of people were asked whether they had a heart attack. The number of people who answered yes or no as well as the total number of respondents is presented in the following table:

	55 years-old or younger	**Over 55**	**Total**
Previous heart attack	29	75	104
No previous heart attack	401	275	676
Total	430	350	780

Suppose we select a person from this group at random. Based on the table, what is the probability that the person had a heart attack?

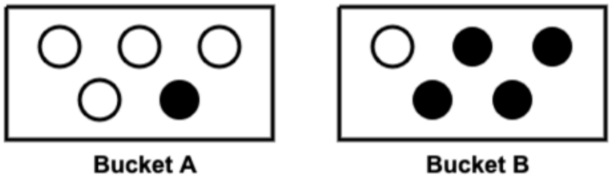
- [104 out of 780 **[CORRECT]**; 104 out of 676; 390 out of 780]

[INDUCED BELIEF ELICITATION INSTRUCTIONS]

Part 2

The Bucket Game

In this part you will play a computerized game that involves two buckets: Bucket A and Bucket B. Bucket A contains 4 light balls and 1 dark ball. Bucket B contains 1 light ball and 4 dark balls. These buckets and balls are all computerized and they are depicted below:



You will play the game 10 times. Each time you play the game, one of the buckets will be randomly and secretly chosen by the computer as the **Mystery Bucket**, with an X-in-10 chance that Bucket A is selected. The number X may be different each time you play the game, but you will always know in advance what the number is. For example, suppose X=6, then there is a six-in-ten chance (60 percent) that the computer selects Bucket A, and a four-in-ten chance (40%) that the computer selects Bucket B.

Given this X-in-10 chance, the selected bucket is determined as follows:

- The computer rolls a fair 10-sided die and compares the outcome to X. Each number on the die, from 1 to 10, is equally likely.
- If the die roll is less than or equal to X then Bucket A is selected.
- If the die roll is greater than X then Bucket B is selected.

Note that the Mystery Bucket may be the same or different each time you play the game. The computer will roll the die separately each time and the Mystery Bucket will depend on the outcome of that roll.

The computer will then randomly select one ball from the Mystery Bucket and show you the color of the ball. Each ball within the Mystery Bucket is equally likely to be selected.

Your task is to **report your Belief** about the percentage chance the ball came from Bucket A. You will report your belief twice, once before seeing the ball's color and then again after seeing the color of the ball. The game ends after you see the color of the ball and report your second Belief.

Summary

1. **Computer tells you the value of X.**
   - The value of X may be different each time you play the game.
2. **Computer fills the buckets.**
   - The two buckets are filled with five balls each.
   - Bucket A has 4 light balls and 1 dark ball.
   - Bucket B has 1 light ball and 4 dark balls.
   - You will always see the exact number of light and dark balls in the two urns.
3. **Computer selects Mystery Bucket.**
   - There is an X-in-10 chance that Bucket A is selected as the Mystery Bucket.
4. **You report your first Belief that the Mystery Bucket is Bucket A.**
   - Provide your best guess about the chance that the computer selected Bucket A knowing X and *before* seeing the ball.
5. **Computer shows you a randomly selected ball from the Mystery Bucket.**
   - Each ball in the Mystery Bucket is equally likely to be selected.
6. **You report your second Belief that the Mystery Bucket is Bucket A.**
   - Provide your best guess about the chance that the computer selected Bucket A knowing X *after* seeing the ball.

1. What does X mean for the Bucket Game?
   - [There is an X-in-10 chance that Bucket A is the Mystery Bucket. **[CORRECT]**; There is an X-in-10 chance that Bucket B is the Mystery Bucket.; There is an X-in-10 chance that a dark ball is drawn.; There is an X-in-10 chance that a light ball is drawn.]
2. Suppose X=1. Which statement is correct?
   - [There is a 10 percent chance that Bucket A is the Mystery Bucket. **[CORRECT]**; There is a 1 percent chance that Bucket A is the Mystery Bucket.; There is a 10 percent chance that a dark ball is drawn.; There is a 1 percent chance that a light ball is drawn.]
3. The first time you will be asked to report your belief, you will know:
   - [The value of X and the number of light and dark balls in Buckets A and B. **[CORRECT]**; Only the value of X.; The value of X, the number of light and dark balls in Buckets A and B, and the color of a randomly selected ball from the Mystery Bucket.; Only the color of a randomly selected ball from the Mystery Bucket.]
4. The second time you will be asked to report your belief, you will know:
   - [The value of X, the number of light and dark balls in Buckets A and B, and the color of a randomly selected ball from the Mystery Bucket. **[CORRECT]**; Only the value of X.; The value of

X and the number of light and dark balls in Buckets A and B.; Only the color of a randomly selected ball from the Mystery Bucket.]

Your Belief

You will report your Belief as a number between 0 and 100 to indicate your best guess about the percent chance (chance-in-100) that the Mystery Bucket is Bucket A.

To report your belief you will move a slider bar like the one below. For the slider to appear, you need to click anywhere on the 0-100 axis. The selected number will then appear underneath the axis.

The number appearing after moving the slider indicates your Belief about the percent chance that Bucket A was selected and it is interpreted as follows:

Your Belief	**This means:**
100	You think there is a 100 percent chance that the Mystery Bucket is Bucket A. That is, you are certain, beyond any doubt, that the Mystery Bucket is Bucket A.
51-99	You think there is a higher percent chance that the Mystery Bucket is Bucket A than Bucket B (higher numbers indicate greater certainty it is Bucket A).
50	You think there is an equal percent chance that the Mystery Bucket is Bucket A or Bucket B.
1-49	You think there is a lower percent chance that the Mystery Bucket is Bucket A than Bucket B (lower numbers indicate greater certainty it is Bucket B).
0	You think there is a 0 percent chance that the Mystery Bucket is Bucket A. That is, you are certain, beyond any doubt, that the Mystery Bucket is Bucket B.

Next, we will ask you a series of questions to check that you understand how different values for your Belief are interpreted.

1. Suppose you report that your Belief is equal to 35. Which of the following interpretations is correct?
   - [You believe there is a 35 percent chance that Bucket A is the Mystery Bucket. **[CORRECT]**; You believe there is a 35 percent chance that Bucket B is the Mystery Bucket.]
2. Suppose you report that your Belief is equal to 90. Which of the following interpretations is correct?
   - [You believe there is a 90 percent chance that Bucket A is the Mystery Bucket. **[CORRECT]**; You believe there is a 90 percent chance that Bucket B is the Mystery Bucket.]
3. Suppose you report that your Belief is equal to 23. Which of the following interpretations is correct?

  - [You believe that Bucket B is more likely to be the Mystery Bucket than Bucket A. **[CORRECT]**; You believe that Bucket A is more likely to be the Mystery Bucket than Bucket B.]

4. Consider two potential beliefs: 58 and 74. Which of the following statements is correct?
   - [If you report 74, you believe that Bucket A is more likely to be the Mystery Bucket than if you report 58. **[CORRECT]**; If you report 74, you believe that Bucket A is less likely to be the Mystery Bucket than if you report 58.]

[BDM CONDITION ONLY]

Payment Rule

We now explain how your Belief is used to determine whether you might win the $8 bonus.

There are now two possible ways to win the bonus of $8: the Bucket Game or the Lottery Bag Game.

- In the Bucket Game, you win the bonus if the Mystery Bucket is Bucket A.
- In the Lottery Bag Game, you win if a winning lottery ticket is drawn from the Lottery Bag (described next).

In the Lottery Bag Game, the computer creates a Lottery Bag by randomly choosing a number W between 0 and 100. Each number is equally likely to be chosen. Although the computer knows this number, you do not. You can think of the computer as filling a bag with 100 lottery tickets. W out of 100 tickets in the Lottery Bag are winning tickets, and the rest are not. The computer will then randomly draw one ticket from the Lottery Bag. You win the bonus in the Lottery Bag Game if one of the W winning tickets is drawn.

Based on the report of your Belief that Bucket A was selected, the computer will select the game that gives you the highest chance of winning $8.

- Your chance of winning in the Bucket Game is effectively your Belief out of 100.
- Your chance of winning in the Lottery Bag Game is W out of 100.
- If your Belief is greater than W then you will play the Bucket Game. If your Belief is lower than W, you will play the Lottery Bag Game. (If the Bucket Game and Lottery Bag Game both give you an equal chance of winning, you will play the Lottery Bag Game.)

You should think carefully about your Belief, as this procedure is designed so that you have the best chance of winning the bonus when you state your Belief as accurately as possible about the likelihood you think the computer selected Bucket A.

[BSR CONDITION ONLY]

Payment Rule

We now explain how your Belief is used to determine whether you might win the $8 bonus.

After you state your Belief, the computer randomly draws two whole numbers, Y and Z, each with values between and including 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 Y in no way affects the value selected for Z and vice versa.

The computer determines whether you win the $8 bonus according to which bucket was selected as the Mystery Bucket:

- If Bucket A is the Mystery Bucket, then you receive the bonus if your Belief is greater than or equal to either of the two numbers Y or Z.
- If Bucket B is the Mystery Bucket, then you receive the bonus if your Belief is less than either of the two numbers Y or Z.

You should think carefully about your Belief, as this procedure is designed so that you have the best chance of winning the bonus when you state your Belief as accurately as possible about the likelihood you think the computer selected Bucket A as the Mystery Bucket.

[INTROSPECTION CONDITION ONLY]

Payment Rule

We now explain how you might win the $8 bonus.

For each reported Belief, the computer will flip a fair coin. You will win the bonus if the outcome of the coin flip is Heads. You do not win the bonus if the outcome is Tails.

You will play this game 10 times. At the end of the experiment, only 1 of these 10 games will be randomly chosen by the computer to count for payment.

After the computer selects which game counts, we will then randomly choose whether the Belief you reported before or after seeing the ball will count to determine your actual bonus.

- Note that this means that every Belief you report has the same chance of being selected to count to determine your payment.
- We emphasize that this is a NO DECEPTION study. The computer will make the random draws and calculate your bonus behind the scenes following the procedures we described to you.
- You will not see any of the draws for any Belief you report until you finish playing the game 10 times.

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.

[INTROSPECTION CONDITION ONLY]

1. Imagine you report that your Belief is equal to 90. Which of the following is correct?
    - [You win the bonus if the outcome of the coin flip is Heads. **[CORRECT]**; You win the bonus only if Bucket B is the Mystery Bucket.; You win the bonus only if Bucket A is the Mystery Bucket.; You win the bonus if the outcome of the coin flip is Tails.]
2. Imagine you report that your Belief is equal to 40. Which of the following is correct?
    - [You win the bonus if the outcome of the coin flip is Heads. **[CORRECT]**; You win the bonus only if Bucket B is the Mystery Bucket.; You win the bonus only if Bucket A is the Mystery Bucket.; You win the bonus if the outcome of the coin flip is Tails.]
3. Imagine you report that your Belief is equal to 60. Moreover, the Mystery Bucket is Bucket A and the outcome of the coin flip is Heads. What is your bonus for this round?
    - [$8 **[CORRECT]**; $0; $4; $15]
4. Imagine you report that your Belief is equal to 12. Moreover, the Mystery Bucket is Bucket A and the outcome of the coin flip is Tails. What is your bonus for this round?
    - [$0; $4; $8; $15]

[BSR CONDITION ONLY]

1. Imagine you report that your Belief is equal to 90. Which of the following is correct?
    - [You win the bonus if Bucket A is the Mystery Bucket, and your Belief is greater than or equal to either Y or Z. **[CORRECT]**; You win the bonus only if Bucket B is the Mystery Bucket.; You win the bonus if Bucket B is the Mystery Bucket, and your Belief is greater than or equal to either Y or Z.; You win the bonus if Bucket A is the Mystery Bucket, and your Belief is smaller than both Y and Z.]
2. Imagine you report that your Belief is equal to 40. Which of the following is correct?
    - [You win the bonus if Bucket B is the Mystery Bucket, and your Belief is smaller than either Y or Z. **[CORRECT]**; You win the bonus only if Bucket B is the Mystery Bucket.; You win the bonus only if Bucket A is the Mystery Bucket.; You win the bonus if Bucket A is the Mystery Bucket, and your Belief is smaller than both Y and Z.]
3. Imagine you report that your Belief is equal to 60. Then, you find out that the randomly drawn numbers are Y = 58 and Z = 2. Moreover, the Mystery Bucket is Bucket A. What is your bonus for this round?
    - [$8 **[CORRECT]**; $4; $0; $15]
4. Imagine you report that your Belief is equal to 12. Then, you find out that the randomly drawn numbers are Y = 88 and Z = 19. Moreover, the Mystery Bucket is Bucket A. What is your bonus for this round?
    - [$0 **[CORRECT]**; $4; $8; $15]

[BDM CONDITION ONLY]

1. Imagine you report that your Belief is equal to 90. Which of the following is correct?
    - [You win the bonus if Bucket A is the Mystery Bucket, and your Belief is larger than or equal to W. **[CORRECT]**; You win the bonus only if Bucket B is the Mystery Bucket.; You win the bonus if a winning ticket is drawn from the Lottery Bag and your Belief is greater than or equal to W.; You win the bonus if a non-winning ticket is drawn from the Lottery Bag and your Belief is smaller than W.]
2. Imagine you report that your Belief is equal to 40. Which of the following is correct?

- [You win the bonus if a winning ticket is drawn from the Lottery Bag and your Belief is smaller than W. **[CORRECT]**; You win the bonus only if Bucket B is the Mystery Bucket.; You win the bonus only if Bucket A is the Mystery Bucket.; You win the bonus if a non-winning ticket is drawn from the Lottery Bag and your Belief is smaller than W.]

3. Imagine you report that your Belief is equal to 60. Then, you find out that the randomly drawn number of winning lottery tickets W is equal to 2. Moreover, the Mystery Bucket is Bucket A and a non-winning ticket was drawn from the Lottery Bag. What is your bonus for this round?
   - [$8 **[CORRECT]**; $0; $4; $15]
4. Imagine you report that your Belief is equal to 12. Then, you find out that the randomly drawn number of winning lottery tickets W is equal to 19. Moreover, the Mystery Bucket is Bucket A and a non-winning ticket was drawn from the Lottery Bag. What is your bonus for this round?
   - [$0 **[CORRECT]**; $4; $8; $15]

[ALL CONDITIONS]

5. Suppose you are completely uncertain so that you believe Bucket A and Bucket B are equally likely. What would give you the greatest chance of winning the bonus?
   - [Reporting a Belief equal to 50. **[CORRECT]**; Reporting a Belief less than 50.; Reporting a Belief greater than 50.; The Belief I report does not influence my chance of winning the bonus.]
6. Suppose you think there is a 67-in-100 chance that Bucket A is the Mystery Bucket. What would give you the greatest chance of winning the bonus?
   - [Reporting a Belief equal to 67. **[CORRECT]**; Reporting a Belief equal to 54.; Reporting a Belief equal to 79.; The Belief I report does not influence my chance of winning the bonus.]

[SUBJECTIVE COMPREHENSION]

1. 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]
2. How simple or complex did you find the instructions for determining the bonus?
   - [Extremely simple; Somewhat simple; Neither simple nor complex; Somewhat complex; Extremely complex]