Cognitive Heterogeneity and Complex Belief Elicitation

Ingrid Burfurd and Tom Wilkening[*]

May 27, 2021

Abstract

The Stochastic Becker-DeGroot-Marschak (SBDM) mechanism is a theoretically elegant way of eliciting incentive-compatible beliefs under a variety of risk preferences. However, the mechanism is complex and there is concern that some participants may misunderstand its incentive properties. We use a two-part design to evaluate the relationship between participants' probabilistic reasoning skills, task complexity, and belief elicitation. We first identify participants whose decision-making is consistent and inconsistent with probabilistic reasoning using a task in which non-Bayesian modes of decision-making lead to violations of stochastic dominance. We then elicit participants' beliefs in both easy and hard decision problems. Relative to Introspection, there is less variation in belief errors between easy and hard problems in the SBDM mechanism. However, there is a greater difference in belief errors between consistent and inconsistent participants. These results suggest that while the SBDM mechanism encourages individuals to think more carefully about beliefs, it is more sensitive to heterogeneity in probabilistic reasoning. In a follow-up experiment, we also identify participants with high and low fluid intelligence with a Raven task, and high and low proclivities for cognitive effort using an extended Cognitive Reflection Test. Although performance on these tasks strongly predict errors in both the SBDM mechanism and Introspection, there is no significant interaction effect between the elicitation mechanism and either ability or effort. Our results suggest that mechanism complexity is an important consideration when using elicitation mechanisms, and that participants' probabilistic reasoning is an important consideration when interpreting elicited beliefs.

Keywords: Belief Elicitation, Probabilistic Reasoning, Cognition, Complexity, Observer Effect

JEL Classifications: C81 C91

[*]Ingrid Burfurd: Centre for Market Design, Department of Economics, University of Melbourne. E-mail: ingrid.burfurd@gmail.com. Tom Wilkening: Department of Economics, The University of Melbourne. E-mail: tom.wilkening@unimelb.edu.au. We thank Amy Corman and Aaron Kamm for assistance in running the experiments. We also thank Ralph-Christopher Bayer, Glenn Harrison, Nikos Nikiforakis, and James Tremewan for helpful comments. We gratefully acknowledge the financial support of the Australian Research Council (DE140101014) as well as the Faculty of Business and Economics at the University of Melbourne. Raw data, the programs used to analyze the data, and the programs used to run the experiment are available at https://osf.io/ex26f/.

1 Introduction

Most economic theories describe the decision-making process as a confluence of preferences, beliefs, and cognitive processes. Disentangling these primitives is a challenge because they are all unobservable in most empirical data. An important advantage of experiments is that auxiliary revelation mechanisms can be used to elicit participants' beliefs. Accurate belief data can supplement choice data to facilitate stronger identification of the preferences and cognitive processes that guide choice.

It is well-known that heterogeneous preferences can make eliciting accurate beliefs difficult.[1] This is because heterogeneous preferences may also impact behavior in the revelation mechanism used to elicit beliefs. For example, participants may misreport in unincentized Introspection mechanisms (Introspection) if they find it arduous to think carefully about their beliefs or if revealing their true belief causes them discomfort. Explicit incentives can mitigate these issues, but incentive-compatible mechanisms must use lotteries and lotteries interact with risk preferences. This has led to the use of the sophisticated Stochastic Becker-Degroot-Marschak mechanism (SBDM), which is predicted to induce truthful revelation for a wide variety of preferences.[2]

Despite the impressive theoretical properties of the SBDM, there is little evidence that the SBDM mechanism outperforms Introspection in terms of belief accuracy (Hollard et al., 2016; Trautmann and van de Kuilen, 2015).[3] Further, participants using incentive-compatible belief elicitation mechanisms often misreport their beliefs even when the probability of an event occurring is objectively known (Hao and Houser 2012; Burfurd and Wilkening 2018). These results suggest that there may be a second potential difficulty for belief elicitation: an interaction between belief elicitation mechanisms and cognition.

Heterogeneous responses to belief elicitation based on cognition has potentially important implications for interpreting belief data and for choosing a belief elicitation method. If decision-making and reporting behavior vary systematically with knowledge, modes of reasoning, cognitive effort, or fluid intelligence—fundamental components of cognition—complex mechanisms might yield reliable reporting data from participants who make

[1] For a discussion about belief elicitation techniques and preferences, see Schlag et al. (2013), Schotter and Trevino (2014), and Trautmann and van de Kuilen (2015).

[2] The mechanism that we refer to as the SBDM has a variety of names in the literature. Ducharme and Donnell (1973) is the first empirical paper we are aware of that uses the procedure and refers to the mechanism as "bets mode" for eliciting beliefs. Schlag et al. (2013) refer to the mechanism as "reservation probabilities" while Trautmann and van de Kuilen (2015) use the term "probability matching". Many other papers refer to the mechanism as the "Karni mechanism" due to the theoretical contributions of Karni (2009). We prefer SBDM due to the strong similarities between the mechanism and the mechanism proposed by Becker et al. (1964) for eliciting valuations. The use of probabilities to control for risk aversion is discussed as early as Smith (1961) and Savage (1971). Varieties of the mechanism have been studied by Grether (1981), Allen (1987), and Holt (2006).

[3] Hollard et al. (2016) finds that both the SBDM and Introspection outperforms the quadratic scoring rule using a series of subjective tasks. Trautmann and van de Kuilen (2015) find that while accuracy between the SBDM mechanism and Introspection mechanism do not differ, there is some evidence that incentive-compatible belief mechanisms are better predictors of a participant's own actions.

better decisions, and less reliable data from participants who make sub-optimal ones. This could have serious implications for analysis, since belief errors would be correlated systematically with unobservable skills and abilities. It also suggests that researchers may face a tradeoff between catering for heterogenous preferences and heterogeneous cognition.

In this paper, we focus on a potential interaction between the SBDM mechanism and probabilistic reasoning. To study this interaction, we use a two-part design in which we first identify participants whose decision-making is *consistent* or *inconsistent* with probabilistic reasoning and then examine how participants of both types respond to different belief elicitation mechanisms.

To evaluate whether participants' decisions are consistent with probabilistic reasoning, we use a variant of an urn task introduced in Charness and Levin (2005) and Charness et al. (2007), which we refer to as the Bucket Game. In each period, participants are individually assigned one of two buckets (A or B) with equal probability. Each bucket is divided into two sides, and each side contains 20 balls; each ball may be black or white. Participants draw and replace a ball from the *left* side of their bucket at the start of each period and are paid $4 if they observe a black ball and $0 if they observe a white ball. The color of the ball is informative, and presents an opportunity for a participants to update their belief that they have been given Bucket A. Participants then choose whether they would like to draw an additional ball from the left or the right side of their bucket. They are paid $4 if their second ball is black and $0 if their second ball is white.

The task is structured so that it is optimal for an individual who updates her belief in the direction predicted by Bayes' rule to *switch* to the right side of her bucket if the first ball successfully earned $4, and to *stay* with the left side of her bucket if the first ball was unsuccessful. By contrast, if individuals use a simple reinforcement-learning heuristic or have an affective response to success, they will prefer to *stay* after a success and *switch* after a failure. This choice pattern directly violates stochastic dominance and reveals behavior that is not consistent with Bayes' rule. Thus, by observing decisions in the Bucket Game we can identify individuals whose choices are "consistent" and "inconsistent" with probabilistic reasoning and stochastic dominance.

After 20 iterations of the Bucket Game, we begin part two of our experiment. Participants continue to play the Bucket Game, but belief elicitation is introduced. In our main treatments, half of the participants are exposed to an Introspection mechanism; after observing their first ball, participants are asked to report the probability that they have been given Bucket A. The other half are exposed to the SBDM mechanism. This belief elicitation method is incentive-compatible under minimal assumptions about risk preferences but is fairly complex and likely unfamiliar to participants.

To allow for variation in the probabilistic difficulty of forming correct beliefs, we vary the composition of balls in the bucket and the number of balls drawn. A feature of our design is that all participants observe instances of two balls being drawn, and both a black and a white ball being observed. In these periods the combined signal is uninformative and thus it takes no effort to form a belief. Our design therefore allows us to observe belief errors in (i) problems in which signals are informative and beliefs are costly to compute and (ii) problems in which signals are uninformative and beliefs are easy to compute.

We interpret the Bucket Game as identifying individuals who have high and low crystallized intelligence related to probabilistic reasoning. Relative to fluid intelligence, which captures an individual's capacity for abstract reasoning and "inductive" capacity, crystallized intelligence relates to the knowledge an individual has acquired through experience (Horn and Cattell, 1966).[4] Ex-ante, we predicted that probabilistic reasoning would be important for the SBDM mechanism because it requires that an individual's choices are consistent with stochastic dominance and probabilistic sophistication in order for its incentive properties to hold. As our "inconsistent" group frequently violates stochastic dominance, we predicted that these individuals may not understand the incentive properties of the mechanism and may have additional belief errors as a result.

In the SBDM mechanism, errors may arise from two potential sources: (i) inaccurate underlying beliefs that are a result of incorrect Bayesian updating and (ii) misreported beliefs that are due to a misunderstanding of the incentive properties of the mechanism. As both types of errors are likely to differ with probabilistic reasoning, observing a difference in the belief errors between consistent and inconsistent participants in the SBDM mechanism does not necessarily imply that inconsistent participants are misunderstanding the incentive properties of the mechanism.

To isolate the mechanism-specific misreport channel, we employ a difference-in-difference approach in which we compare the difference in mean errors between consistent and inconsistent participants in the SBDM mechanism with the same difference in the Introspection mechanism. As the Introspection mechanism is not predicted to generate mechanism-specific misreports, we predict that the SBDM mechanism will have a larger difference in belief errors between consistent and inconsistent participants than the Introspection mechanism.[5]

[4] The terms crystallized intelligence and fluid intelligence come from Cattell's model of generalized intelligence (Cattell, 1963), which is widely used in the psychology, cognitive science and economic literatures. Our categorization of probabilistic reasoning as crystallized intelligence is based on the psychology literature, which suggests that individuals naturally think in frequencies rather than probabilities. See, for instance, Gigerenzer (1984) and Gigerenzer and Hoffrage (1995). Raven's Progressive Matrices is the most widely-used tool for measuring fluid intelligence across disciplines and within economics (Huepe et al., 2011; Li et al., 2013; Lilleholt, 2019). Tests of crystalized intelligence are more domain-specific, and typically involve metaphor comprehension tasks or tests of linguistic skills (Schipolowski et al., 2014). Tests of numeric and probabilistic abilities include the Berlin Numeracy Test, introduced by Cokely et al. (2012), and the Probabilistic Reasoning Scale (PRS) (Primi et al., 2016).

[5] Interpreting the difference-in-difference as a measure of SBDM-specific misreports relies on an assumption that any difference in errors between consistent and inconsistent participants that stem from inaccurate beliefs are similar for the two mechanisms. One of our two main hypotheses is that the SBDM mechanism improves the accuracy of underlying beliefs in decision problems that are cognitively costly. Thus, there is a concern that accuracy improvements may not be uniform across consistent and inconsistent participants. As such, we report the difference-in-difference estimate both for the full sample, in

Pooling the data from our initial experiments and pre-registered follow-up experiments, the results are in line with these predictions: the mean error of a consistent participant is 37.6 percent smaller than an inconsistent participant in the SBDM. However, it is only 22.4 percent smaller in the Introspection mechanism. The difference in sensitivities to probabilistic reasoning is significant in a permutation test that ensures independence between the main effects and the interaction effect. Further, the difference in sensitivities is strongest in easy decision problems in which Bayesian updating is not required. In these decision problems, identifying the correct belief is unlikely to be cognitively costly and differences in belief errors is most likely driven by confusion stemming from the SBDM mechanism itself.

A caveat to these results is that the magnitude of the estimated interaction effect between the SBDM mechanism and probabilistic reasoning is large and significant in our initial lab-based experiments but small and not significant in our online follow-up experiments.[6] Further, while the magnitude of the interaction effect in the follow-up experiments increases when the data is restricted to easy decision problems or when the most obvious outliers are removed, a difference in the magnitude of the interaction effect between the two samples persists. Thus, we see value in future independent replications of our experiments and in understanding whether there are differences in how belief elicitation mechanisms and incentives interact with lab and online environments.

The main rationale for using incentive-compatible belief elicitation is to induce participants to carefully report their beliefs and to provide high-quality information even when calculating beliefs is costly. Thus, we would predict that belief errors in the SBDM mechanism will be less sensitive to the difficulty of the decision problem than Introspection. Consistent with this second prediction, we find that in the SBDM mechanism, the mean error of participants in easy decision problems is 20.7 percent smaller than in hard decision problems. By contrast, in the Introspection mechanism, the mean error in easy decision problems is 53.8 percent smaller than the mean error in hard decision problems. The difference in sensitivities to task difficulty is significant in a permutation test that ensures independence between the main effects and the interaction effect.

In a follow-up experiment, we also identify individuals with high and low fluid intelligence ("ability") with a Raven task and high and low proclivity for cognitive effort using an extended Cognitive Reflection Test.[7] Although these tasks strongly predict errors in

which we rely on the additional assumption, and for a subset of easy decision problems in which the signal is uninformative and in which underlying beliefs are likely to be accurate.

[6] In our pre-analysis plan we committed to pooling the data from our original and follow-up experiments if there were no significant differences in errors in the full data set, the SBDM sample, or the Introspection sample. As seen in Appendix A, we find no differences in the sample along these dimensions and therefore used the pooled data to evaluate our main two hypotheses.

[7] Although performance in the Cognitive Reflection Test is correlated with cognitive ability (Frederick, 2005; Obrecht et al., 2009; Toplak et al., 2011; Brañas-Garza et al., 2012), the CRT captures a distinct dimension of cognition which is consistent with engagement and effort. Toplak et al. (2011) use regression analysis to demonstrate that the CRT is a unique predictor of performance in heuristics-and-biases tests,

both the SBDM mechanism and Introspection, there is no significant interaction effect between high and low-ability types and mechanisms, nor high and low-effort types and mechanisms.

Taken together, our results suggest that while the SBDM mechanism encourages participants to think carefully about their beliefs in difficult elicitation problems, some individuals struggle to understand the mechanism. This may lead to heterogeneity in belief errors across a population based on probabilistic reasoning skills. Our results help to clarify why earlier studies have found mixed evidence regarding the relative efficiency of the SBDM mechanism and the Introspection mechanism. It also highlights a potential confound in designs that rely on individual-level beliefs since belief errors may be correlated systematically with probabilistic reasoning.

The rest of the paper is arranged as follows. In Section 2 we describe the Stochastic Becker-DeGroot-Marschak mechanism and discuss the existing literature on belief elicitation and cognitive processes. In Section 3 we discuss the experiment, hypotheses, and analysis plan. Results are presented in Section 4.

2 The Stochastic Becker-DeGroot-Marschak Mechanism

Consider a participant in an experiment who has a subjective belief about the distribution of a discrete random variable X, with range $\mathcal{X}$. Her true beliefs P_X describes the probability that $X = x$ for each $x \in \mathcal{X}$, and the researcher wants to know belief p that event $P(X = x)$ will occur.

If participants have an aversion to lying, and if there are no cognitive costs from identifying or reporting p, unincentivized Introspection will be truth-telling. However, if the researcher is concerned that these conditions are not satisfied, she can use explicit incentives to induce truthful reporting. "Scoring rules" describe a payment schedule based on a participant's reported belief $r \in [0, 1]$ and the realisation of the random variable X. For a single realisation of X, a scoring rule S is a mapping $S : [0, 1] \times \mathcal{X} \to \mathbb{R}$. This means that $S(r, x)$ is paid when r is reported and outcome x is realized.

For a participant who has utility function u, in which u is a utility function in the class of von Neumann-Morgenstern Expected Utility functions, a rational participant faced with scoring rule S reports $r \in [0, 1]$ to maximize $\mathbb{E}u(S(r, X))$ where, by the expected utility

and argues that in addition to cognitive ability the CRT test captures important features of rational decision-making, which they term "thinking disposition". While self-reporting measures of cognitive effort exist, such as the Need for Cognition Scale (Cacioppo et al., 1984), the CRT is the most widely-used test in economics (see, for example, Thomson and Oppenheimer (2016); Oechssler et al. (2009); Brañas-Garza et al. (2012); Carpenter et al. (2013); Brañas-Garza et al. (2019)); it has the advantage of being positively correlated with self-reported heuristic tendencies (Juanchich et al., 2016) while offering an objective task-based approach. Thus, it is a natural test to use for an experiment in which heuristic tendencies may be important.

assumption,

$$\mathbb{E}u(S(r,X)) = \sum_{x\in\mathcal{X}} u(S(r,x))P(X=x).$$

Using the terminology introduced by Winkler and Murphy (1968), a "proper" scoring rule renders it optimal for risk-neutral agents to report their beliefs truthfully. That is, given a utility function $u(S(r,X)) = S(r,X)$, the scoring rule is "truth-telling" (or "incentive-compatible") in the sense that, for all $P_X \in \mathcal{P}_X$,

$$p \in \arg\max_{r\in[0,1]} \mathbb{E}u(S(r,X)).$$

As the definition suggests, truth-telling may not occur in cases in which $u(S(r,X)) \neq S(r,X)$. This may be problematic when participants have heterogeneous risk preferences that are unobservable to the researcher.[8]

As noted as far back as Smith (1961) and Savage (1971), moving from a deterministic scoring rule to a stochastic one makes it possible to induce truth-telling for all von Neumann-Morgenstern Expected Utility miximizers.[9] Here, we discuss a stochastic scoring rule that has garnered significant interest in the literature: the Stochastic Becker-DeGroot-Marschak mechanism (SBDM).

In the SBDM mechanism, the experimenter presents a participant with a choice under risk, described by lottery H_AL, which pays H if event A occurs and L if not. The participant forms a subjective belief that event A will occur. We denote this subjective probability p. The participant is then asked to issue report $r \in [0,1]$ about her belief p before making a decision based on her beliefs. A second lottery is created in parallel. A number z is realized from the distribution of random variable Z, which has distribution P_Z on support $[0,1]$. The participant does not know z, but does know that if z falls above her report r she will receive lottery H_zL, which makes a high payoff H with probability z. If z falls below r she receives lottery H_AL. The lotteries therefore offer identical payoffs with different probabilities. It is in the participant's best interest to report $r = p$, because a report of $r \neq p$ might mean the participant receives the less desirable lottery.

By construction, the SBDM uses the same two payoffs for the subjective and objective lotteries and thus the particular cardinal values assigned to the high and low payoffs are not predicted to influence reports. As a result, the SBDM induces truth-telling under minimal assumptions about preferences; namely, that $u(S(r,X))$ are consistent with stochastic dominance and probabilistic sophistication (Karni, 2009). As per Machina and Schmeidler (1992), probabilistic sophistication means that a participant will rank

[8] See Schlag et al. (2013) for a review of scoring rules and techniques that might be used to control for risk aversion. In addition to the stochastic elicitation techniques discussed below, researchers have also tried to separate risk preferences from beliefs econometrically. See, in particular, Offerman et al. (2009) and Andersen et al. (2014).

[9] Formally, a stochastic scoring rule is a mapping from $S : [0,1] \times \mathcal{X} \to \Delta(\mathbb{R})$, where $\Delta(\mathbb{R})$ is a lottery over one or more real outcomes.

lotteries according to the implied probability distribution over outcomes. "Stochastic dominance" is the condition that a participant has preference relation $\succeq$ over lotteries such that $H_q L \succeq H_{q'} L$ for all $H > L$ if and only if $q \geq q'$.

2.1 Cognitive Processes and Belief Elicitation

Although there is little research that empirically studies the interaction between cognitive processes and reporting behavior in belief elicitation mechanisms, a few papers suggest that cognition may influence behavior in the SBDM. Hao and Houser (2012) evaluate two implementations of the SBDM mechanism: the standard implementation in which a participant directly reports her beliefs, and an ascending clock mechanism.[10] While both mechanisms are incentive-compatible, the ascending clock mechanism is also obviously strategy-proof and more easily understood by cognitively limited agents (Li, 2017). Thus, differences in the quality of reports across these two implementations suggests that cognition may influence reporting. Hao and Houser identify "naive" subjects, who report $r \neq p$, and "sophisticated" subjects who report $r = p$. The clock mechanism reduces the sample of naive observations and improves the accuracy of reported beliefs.[11]

Freeman and Mayraz (2019) study how individuals choose between safe and risky lotteries in environments in which (i) participants are shown exactly one lottery, (ii) they are given a choice list and one decision is randomly selected for payment, and (iii) they are given a choice list but informed about the decision that will be paid before making their choice. The paper finds more risk-taking in the individual choice problem relative to the other two formats and conjectures that the choice list provides scaffolding that helps decision makers identify their true preferences. If cognition is an issue in the SBDM mechanism, then we should also find that belief errors in the SBDM are reduced with choice lists. Holt and Smith (2016) compares behavior between a direct elicitation method and a choice list using an "induced value" urn task in which participants receive one or more signals from an urn. The probability that the balls are drawn from a particular urn can be calculated explicitly via Bayes' rule. The paper does not find a significant difference in belief errors between the choice list approach and a direct elicitation implementation based on Holt and Smith (2009). However, it does find that the choice list reduces boundary reports. Burfurd and Wilkening (2018) also does not find differences in belief errors between a direct elicitation format based on Hao and Houser (2012) and a choice list format in urn problems with a single draw. However, Burfurd and Wilkening (2018)

[10] The clock implementation of Hao and Houser (2012) has each participant compete against a dummy bidder that exits the auction at (unknown) probability z. The clock starts at 0 and rises continuously as long as both the participant and the bidder is in the auction. The clock stops when one of the two bidders drop out. If it is the participant, the participant receives lottery $H_z L$. If the dummy bidder exits first, the bidder receives the original lottery $H_A L$.

[11] Although the ascending clock auction leads to better reports in Hao and Houser (2012), it censors data when the dummy bidder wins. We thus use direct reporting methods in this paper.

does find that there is significant heterogeneity in belief errors across individuals even when the probability of an event is objectively known.

In a concurrent project, Schlag and Tremewan (2019) studies a "frequency" based belief elicitation mechanism that can be used when multiple realisations of an outcome are available. The paper compares this mechanism to an SBDM mechanism based on the instructions of Dal Bó et al. (2017). The authors find that the frequency method performs well against the SBDM and that the difference in performance is driven by a large number of participants who choose a focal report of 50% in the SBDM mechanism. These focal reports are correlated with poor performance in a Cognitive Reflection Test. We do not find the same large spike of focal reports at 50% in our data, though we use a different analogy-based instruction format and include a control quiz.[12]

3 The Experiment

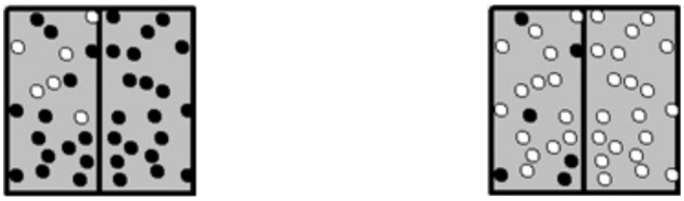
We use a two-part design in which we first identify whether participants are consistent or inconsistent probabilistic reasoners, making use of a computerized "Bucket Game". We then study how participants respond to different belief elicitation techniques. We describe the Bucket Game before introducing the treatments.

3.1 The Bucket Game

The Bucket Game is a variant of an urn task introduced in Charness and Levin (2005) and Charness et al. (2007). In each period, a participant is allocated one of two buckets (A or B) with equal probability. Each bucket is divided into a left and a right side and each side holds 20 balls. Subjects are not told which bucket they have been given, but are provided an illustration that shows the composition of balls in the two buckets. An example illustration is given in Figure 1. As can be seen, the left hand side of each bucket is composed of a mixture of black and white balls and there are more black balls in the left hand side of Bucket A than Bucket B. The right hand side of Bucket A is filled with only black balls and the right hand side of Bucket B is filled with only white balls. The buckets used in all treatments share these features.

In each period, the participant observes the color of a ball that is drawn (with replacement) from the left-hand side of her bucket. If the participant observes a black ball, she receive a stage-one payment of \$4. If the ball is white, she receive \$0. Next, the participant must decide whether to draw a second ball from the same (left) side of her bucket, or to switch to the other (right) side. The participant receives a payment of \$4 if she observes a black ball in this second stage and receives \$0 if she observes a white ball.

[12] Burfurd and Wilkening (2018) find that a control quiz significantly increases accuracy in the SBDM mechanism when using the analogy-based instruction format of Hao and Houser (2012).



(a) Bucket A (b) Bucket B

Figure 1: Illustrations of Bucket A and Bucket B, as presented to participants

There are more black balls on the left hand side of Bucket A than Bucket B. Thus, the first draw from the bucket is informative about the bucket that has been allocated to the participant. Participants whose updating is directionally *consistent* with Bayes' rule are predicted to use this information in their choice. If a consistent participant observes a black ball from the left-hand side of her bucket, her belief that she has been given Bucket A will exceed 0.5 and she should choose to switch to the right side of the bucket. If a consistent participant receives a white ball, her belief that she has been given Bucket A will be less than 0.5 and she should choose to continue to draw from the left side.

However, the game is designed so that the expected value maximizing choice is at odds with an intuitive reinforcement learning heuristic in which a decision maker repeats actions that are successful and changes actions when unsuccessful. When observing a black ball on the first draw, the participant is "successful" and receives \$4. Thus, reinforcement learning predicts that the participant will continue to choose left. After observing a white ball, the participant receives \$0 and reinforcement learning predicts that the participant switches to the right. We therefore predict that participants who use a reinforcement learning heuristic will always choose the side that is stochastically dominated.

3.2 Experimental Design and Treatments

We ran an initial experiment consisting of 239 participants and a follow-up experiment consisting of 244 participants. Our initial experiment was conducted in the University of Melbourne's Experimental Economics Lab and was conducted in a traditional lab setting. Participants were recruited using ORSEE (Greiner, 2015) from the university's experimental economics subject pool and sessions were conducted using z-Tree (Fischbacher, 2007). The follow-up experiment recruited participants from the same database but excluded those who participated in the initial experiment. The follow-up experiment was pre-registered with the Centre for Open Science (https://osf.io/t57vq) and was conducted online using oTree (Chen et al., 2016).

In the initial experiment, we randomized individuals to computers in the lab using a set of bingo balls. Each terminal was assigned one of six potential treatments. These treatments are summarized in Table 1. The treatments differed in the number of black balls in the left hand side of Bucket A, and in the belief elicitation method.

A session consisted of three blocks and each block consisted of 20 periods. In the first block, participants in all treatments received the same computerized instructions describing the Bucket Game and were required to successfully answer all questions in a computerized quiz before starting the experiment. Participants then played 20 periods of the Bucket Game. They were informed about whether they successfully drew a black ball from their chosen side of the bucket in each period.

Treatment	**Belief Elicitation Method (Blocks Two and Three)**	**Number of Black Balls in Left Side of Bucket A**
SBDM - 14	SBDM	14 of 20
SBDM - 12	SBDM	12 of 20
Introspection - 14	Introspection	14 of 20
Introspection - 12	Introspection	12 of 20
No Elicitation - 14	No Elicitation	14 of 20
No Elicitation - 12	No Elicitation	12 of 20

Table 1: Summary of Treatments

In the second block, we elicited beliefs with the SBDM mechanism in one-third of treatments and with an Introspection mechanism in one-third of treatments. The remaining treatments were not exposed to any belief elicitation mechanism and were used to test for an observer effect. We discuss the observer effect in Appendix B.

As with the first block, all participants received computerized instructions at the start of the second block and were required to take a quiz before continuing. The instructions in the Introspection and SBDM treatments explained the belief elicitation task and included additional control questions to ensure participant comprehension.

After reading the instructions for Block Two, participants played twenty more periods of the Bucket Game. We elicited beliefs after the participant had observed the draw from the left hand side of their bucket but before they chose left or right. All beliefs were expressed as the “chance-in-100” the participant had been given Bucket A.

In the Introspection treatments, there were no payments associated with belief reports. However, the instructions asked participants to think carefully about their beliefs.

In the SBDM treatments, we used an adaptation of the direct elicitation method developed in Hao and Houser (2012). This set of instructions was shown in Burfurd and Wilkening (2018) to yield high quality data and to be quick to implement relative to alternatives.

Block Three of the experiment was identical to Block Two, except that a participants initial draw consisted of two balls from the bucket instead of one. These draws were done with replacement and the participant was informed of the colour of both balls before reporting their belief and making their left/right choice. Subjects were paid for each black ball they received from the initial draws. As discussed in more detail in Section

3.2.1 below, this block was important because it created situations in which the signal was uninformative, which allows us to study how beliefs interact with task difficulty. Instructions for Block Three were short and discussed only the additional draw that the participant observed.

To avoid wealth effects and potential hedging strategies, participants were paid in cash for three randomly chosen periods announced at the end of the experiment—one chosen from each of the three blocks.[13] Participants were allowed to proceed at their own pace through the experiment and most participants completed the experiment in under 45 minutes. Including a show-up fee of \$10, the average payment of a participant was \$24.40 AUD. The experiments were run in November and December of 2015, when \$1 AUD $\approx$ \$0.72 USD.

The follow-up experiment was similar to the initial experiment except that we dropped the No-Elicitation treatment and included two additional questionnaires. The first was an expanded version of Frederick's Cognitive Reflection Test (Frederick, 2005), which used three additional questions from Primi et al. (2016) and an additional set of placebo questions taken from Thomson and Oppenheimer (2016). The questions on the CRT were given in a fixed order, with the original and well-known "bat-ball" CRT question asked last. The full list and ordering of questions is included in Appendix H.

The second survey was a short form version of Raven's Advanced Progressive Matrices test developed and validated in Bors and Stokes (1998). The short form consists of 12 questions extracted from the original 36, but does not include early questions in the test that most university students are able to answer correctly.

We randomly selected one question from each quiz and paid the participant \$4 if they answered the question correctly. Thus, the incentives offered in these quizzes were similar in magnitude to the main experiment.

Participants in the follow-up experiment worked at their own pace and no time limits were imposed when answering the main questions or surveys. The show-up fee was increased to \$15 to cover the time required to complete the two questionnaires. The average payment was \$35.55 AUD, with most participants completing the experiment in 75 minutes or less. The experiments were completed in December 2020, when \$1 AUD $\approx$ \$0.74 USD. Due to Covid-19 restrictions the follow-up experiments were conducted online using Zoom and oTree (Chen et al., 2016). All key protocols were preserved and participants were able to privately ask questions throughout. Participants' names and decisions were

[13] In Block One, the participant's profit for the selected period was the value of her first ball plus the value of her second ball. In Blocks Two and Three we used this same payment rule for participants in the Introspection and No Elicitation treatments. For participants in the SBDM treatments, we 'tossed a coin' to determine whether profit for the second ball was determined by her left/right choice—in which case a second ball was drawn from her nominated side of the bucket—or her beliefs. If her profit was determined by her beliefs, then we used the outcome of the SBDM mechanism to determine payment. Subjects could therefore earn \$0, \$4 or \$8 in Block One, \$0, \$4 or \$8 in Block Two, and \$0, \$4, \$8, or \$12 in Block Three.

not visible to other participants.

Following our pre-analysis plan, we compared the initial experiment to the follow-up experiment and did not find any statistically significant differences (see Appendix A). We therefore pooled the data from the two experiments when reporting averages and testing the main two hypotheses. We also show in Appendix F that our results are robust to outliers, which tended to be more frequent in the follow-up online experiment.

Treatment	**Belief Elicitation Method**	**Experiment Sample Size**		
		Initial	Follow-up	Total
SBDM - 14	SBDM	40	59	99
SBDM - 12	SBDM	41	63	104
Introspection - 14	Introspection	40	58	98
Introspection - 12	Introspection	38	64	102
Total	Both	159	244	403

Table 2: Sample sizes

3.2.1 Informative and Uninformative Signals

An important feature of our design is that all participants were exposed to periods in which they drew one black ball and one white ball before reporting their beliefs in Block Three. In these periods, the signals were jointly uninformative and the decision problem required no Bayesian updating to report the true belief. We conjecture that reporting the correct beliefs was not cognitively challenging in these periods, and we compare errors from these periods to periods with informative signals to test whether errors in the Introspection mechanism is influenced by task difficulty.

To generate additional variation in the difficulty of the belief updating task, we also used two different sets of buckets across the treatments and varied the number of balls drawn within a treatment. In our "high information" treatments, Bucket A contained 14 black balls and Bucket B contained 6 black balls. In Blocks One and Two of this treatment, receiving a single black signal results in a posterior of $\rho' = 0.7$ while receiving two black signals in Block Three results in a posterior of $\rho' = 0.84$. In the other half of the treatments, Bucket A contained 12 black balls and Bucket B contained 8 black balls. In these treatments, receiving a single black signal results in a posterior of $\rho' = 0.6$ and receiving two black signals results in a posterior of $\rho' = .69$.

All treatments were designed so that posteriors were an equal distance from the prior whether the participant observes a white or a black ball (i.e., the posteriors were 0.7 and 0.3 after receiving a black ball or a white ball in the high information treatments). This symmetry allows us to cleanly aggregate participants' reported beliefs: for example, in Block Two of the high information treatment, a participant who reported $r = 0.5$ has a belief error of 0.2 regardless of whether they observed a white or a black ball.

3.2.2 Measures of Cognitive Heterogeneity

We classify participants as consistent or inconsistent probabilistic reasoners based on their decisions in the last ten periods of Block One. We elected to use only the second half of the Block One sample to ensure that individuals were not being classified based on early experimentation.[14] A participant is classified as *consistent* if they made 7 or more correct left/right decisions in periods 11-20. Our type cutoff was set to achieve as close to a median split across consistent and inconsistent types as possible. Based on this classification there are 215 consistent participants and 188 inconsistent participants in our treatments with a belief elicitation mechanism. The proportion of consistent types is balanced across treatments, with 105 consistent participants in the Introspection treatments (53 percent of Introspection participants) and 110 inconsistent participants in the SBDM treatments (54 percent of SBDM participants).

Cognitive ability is often divided into crystallized intelligence, which relates to knowledge that an individual has acquired, and fluid intelligence, which relates to a individual's capacity for abstract reasoning, using the model proposed in Cattell (1963). As noted in the introduction, we interpret our Bucket Game as identifying individuals who have high and low crystallized intelligence related to probabilistic reasoning. Individuals who are inconsistent are observed to frequently violate stochastic dominance, which only requires updating in the *direction* predicted by Bayes' rule. We predict that such knowledge is important to the SBDM mechanism because stochastic dominance is one of the weak assumptions required for the mechanism to be incentive compatible.

In our follow-up experiment we use additional surveys to generate measures of fluid intelligence and cognitive effort. Following our analysis plan, we classify individuals as *high-ability* and *low-ability* using a median split of performance in the short-form Raven's Advanced Progressive Matrices test. Individuals are classified as high-ability if they got 9 or more of the 12 matrices questions correct and low-ability otherwise. 134 individuals were classified as high-ability and 110 participants were classified as low-ability. 74 of the high-ability participants were in the Introspection treatment (representing 60 percent of Introspection participants) and 60 of the high-intelligence participants were in the SBDM treatment (51 percent of SBDM participants).

We classify individuals into *high-effort* and *low-effort* groups using a median split of our extended CRT.[15] 137 participants who answered 4 or more CRT questions correctly

[14] In our initial experiment, all hypotheses hold under an alternative specification in which we use all 20 Block-One decisions to classify individuals. See Appendix E for this robustness check. We pre-registered the classification criterion before conducting our follow-up experiments.

[15] Cognitive effort is often analyzed using Stanovich and West's distinction between effortless engagement which draws on heuristics and intuition, referred to as System-1 thinking, and effortful mental operations referred to as System-2 engagement (Stanovich and West, 2000). Frederick's Cognitive Re-

are classified as high-effort, while 107 are classified as low-effort. 67 of the high-effort cohort belong to the Introspection treatment (representing 55 percent of Introspection participants) while 70 belong to the SBDM group (57 percent of SBDM participants).

3.3 Statistics and Hypotheses

Both of our main hypotheses come from a 2×2 factorial design. We are primarily interested in the interaction effect between factors. The standard approach to testing this type of model would be to use a parametric ANOVA specification. However, our dependent variable in this analysis, *Error*, is the absolute error of a participant's reports, relative to the objective Bayesian posterior. The distribution of errors is not normally distributed and thus the underlying assumption of parametric ANOVA is not satisfied. The permutation test represents an ideal alternative since it requires only minimal assumptions about the errors, is exact in some cases, and has high power relative to other approaches.

The main assumption of permutation tests is that the data is exchangeable under the null hypothesis. Data is exchangeable if the probability of the observed data is invariant with respect to random permutations of the indexes (Basso et al., 2009). In the 2×2 factor design, the observations are typically not exchangeable since units assigned to different treatments have different expectations. This implies that approaches that freely permute data may fail to separate main and interaction effects (Good, 2000). Instead, we use a variant of the synchronized permutation test of Perasin (2001) and Salmaso (2003), which restricts permutations to the same level of a factor to generate test statistics for main factors and interactions that are independent of each other (Basso et al., 2009).

A detailed explanation of the synchronized permutation test is included in Appendix D. We note that in some cases our data is not balanced, which can also confound main effects and interaction effects. To deal with this issue, we follow a suggestion in Montgomery (2017) of randomly dropping observations so that each cell has the same number of observations. Although we lose some power by reducing the size of the sample, the resulting data is a random sample of the original and the resulting test statistic is independent of the main effect. To ensure that our random subset of data is not driving our results, we use an outer loop in our testing procedure and perform our permutation test with 1000 sub samples. We report the average p-value over the 1000 sub samples in the main text.

A potential concern when using a permutation test is that it may be sensitive to heterogeneity in the dispersion of points across cells. This issue was raised in the context of the Mann-Whitney test by Fagerland and Sandvik (2009), who show that deviations in Type I error rates can be generated for a null of identical means or medians when the means and medians of two samples are the same but the skewness or kurtosis of the

flection Test (CRT) (Frederick, 2005) is the most widely-used tool for gauging a participant's tendency towards System 1-or-2 cognition.

samples differ. To at least partially address this concern, we also tested a Wald-type permutation statistic (WTPS) developed by Pauly et al. (2015). This procedure uses a free permutation of the dependent variable and is asymptotically valid in the case of heteroscedasticity in the errors across cells. As seen in Appendix F, results using this test are similar to those in the main text if we control for outliers.

Finally, in our tables, we also report the results from pairwise permutation tests. For these tests, we regress error on the mechanism treatment dummy and randomize assignment to treatments using the "ritest" command in Stata (Heß, 2017). These permutation tests are performed 10,000 times and the null hypothesis is that there are no differences between the test groups.

3.3.1 Hypotheses

Sensitivity to Probabilistic Reasoning: As shown by Karni (2009), the SBDM mechanism is incentive-compatible when individuals' preferences over risk satisfy probabilistic sophistication and stochastic dominance. Thus, for consistent participants, we would predict lower errors in the SBDM regardless of the difficulty of the belief updating problem.

By contrast, a participant who makes an incorrect decision in the Bucket Game is actively choosing a bucket with a lower expected value over one with a higher expected value. Such actions violate stochastic dominance. Thus, inconsistent participants may have difficulty understanding and interacting with the SBDM mechanism.

Using behavior in the Introspection treatments to control for inherent differences in accuracies between the two groups, we predict:

Hypothesis 1 *The SBDM mechanism is more sensitive to probabilistic reasoning than the Introspection mechanism.*

If Hypothesis 1 is true, we should see a larger difference in errors between consistent and inconsistent participants in the SBDM mechanism than in the Introspection mechanism. Let $i \in \{1, 2\}$ represent the assignment of an individual to the SBDM mechanism ($i = 1$) or the Introspection mechanism ($i = 2$). Likewise, let $j \in \{1, 2\}$ represent whether an individual is classified as consistent ($j = 1$) or inconsistent ($j = 2$). Then, using a standard additive ANOVA specification, we assume that the mean absolute error of individual k assigned to mechanism i and classified as type j, E_{ijk}, can be decomposed into a overall mean (μ), two main effects (α_i and β_j), an interaction effect $(\alpha\beta)_{ij}$, and an error term ϵ_{ijk}:

$$E_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}. \tag{1}$$

By including the additive constant μ, all main effects and interactions in the model can be defined to sum to zero. Thus, we assume that $\alpha_1+\alpha_2 = 0$, $\beta_1+\beta_2 = 0$, $(\alpha\beta)_{i1}+(\alpha\beta)_{i2} = 0$ for all i, and $(\alpha\beta)_{1j} + (\alpha\beta)_{2j} = 0$ for all j. In this construction, $\alpha_1 = -\alpha_2$ and thus, under the null of no effect of the mechanism on errors, each of the main effects $\alpha_1 = \alpha_2 = 0$. Under the alternative, α_1 represents the difference from a zero average, and the interaction term $(\alpha\beta)_{ij}$ represents the deviation from the sum $\alpha_i + \beta_j$.

Hypothesis 1 predicts that $(\alpha\beta)_{11} < 0$. This would imply that there is a greater difference in errors between consistent and inconsistent participants in the SBDM mechanism than in the Introspection mechanism. As seen in the Appendix, the estimate for $(\alpha\beta)_{11}$ is based on the difference between (i) the difference in mean errors between consistent and inconsistent types in the SBDM mechanism and (ii) the difference in mean errors between consistent and inconsistent types in the Introspection mechanism. Thus, when discussing our results, we will report the mean errors of each group and discuss the magnitude and one-sided significance of this difference-in-difference.

As noted in the introduction, a belief error in the Introspection treatment is based on inaccurate underlying beliefs that are a result of incorrect Bayesian updating while a belief error in the SBDM mechanism may be a combination of (i) inaccurate underlying beliefs and (ii) misreported beliefs that are due to a misunderstanding of the incentive properties of the mechanism. In order for the interaction effect to be interpreted as a measurement of SBDM-specific misreports, the difference in errors between consistent and inconsistent participants that stem from inaccurate beliefs must be similar for the two mechanisms.

As discussed below, we hypothesize that the SBDM mechanism is likely to improve accuracy in difficult questions in which decision making is cognitively costly. Thus, there is a concern that accuracy improvements may not be uniform across consistent and inconsistent participants. To address this concern, we report the difference-in-difference estimate for Hypothesis 1 using only the decision problems with an uninformative signal in addition to reporting the estimate from the full sample. In this subset of decision problems, underlying beliefs require no updating and we have no reason to believe that belief accuracy should differ across mechanisms.

Sensitivity to Task Difficulty: While the Introspection mechanism may be easier for inconsistent participants to understand, a concern is that participants may not have an incentive to think carefully about their belief when updating is cognitively costly. This would imply that the quality of data in the Introspection mechanism may be strongly dependent on the difficulty of forming accurate beliefs.

In our design, participants are exposed to decision problems in which signals are informative and in which Bayesian updating is challenging. Participants are also exposed to simple problems in which signals are uninformative and no Bayesian updating is needed. Using behavior in the SBDM treatments to control for inherent differences in belief errors between these two types of problems, we would predict:

Hypothesis 2 *The Introspection mechanism is more sensitive to task difficulty than the SBDM mechanism.*

To test for Hypothesis 2, we again let $i \in \{1, 2\}$ represent the assignment of an individual to the SBDM mechanism ($i = 1$) or the Introspection mechanism ($i = 2$), but divide our decision problems into hard problems in which the posterior is informative ($j = 1$) and easy problems in which the posterior is uninformative ($j = 2$). We predict that the difference is greater in the Introspection mechanism than in the SBDM mechanism. Thus, our test statistic is given by:

$$E_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \tag{2}$$

where E_{ijk} is the mean absolute error of participant k in mechanism i in decision problems of j difficulty. We predict that $(\alpha\beta)_{21} > 0$ as this would indicate that there is greater variation in belief errors under Introspection when participants encounter easy decision problems relative to difficult decision problems. We note that $(\alpha\beta)_{21}$ is based on the difference between (i) the difference in mean errors between informative and uninformative problems in the Introspection mechanism and (ii) the difference in mean errors between informative and uninformative problems in the SBDM mechanism. Thus, when discussing our results, we will again report the mean errors associated with each mechanism-difficulty combination, and discuss the magnitude and one-sided significance of this difference-in-difference.

Combining Hypotheses 1 and 2, we predict that the relative performance of the SBDM is likely to be best for consistent types in problems with informative signals and worst for inconsistent types in problems with uninformative signals. A priori, we cannot order the other two combinations of types and decision problems since the relative importance of mechanism complexity and task difficulty are unknown.

4 Results

4.1 Probabilistic Reasoning

Result 1 *Consistent with Hypothesis 1, the SBDM mechanism is more sensitive to probabilistic reasoning than the Introspection mechanism.*

Table 3 reports mean errors of reports under the SBDM mechanism and the Introspection mechanism for (i) consistent participants, (ii) inconsistent participants, and (iii) both consistent and inconsistent participants combined. We report mean errors for each informative posterior pair starting with the most informative posteriors and ending with the least informative signal. Thus, for instance, the $\rho' \in \{0.16, 0.84\}$ column corresponds to data from Block Three of the high-information treatments when a participant has drawn either two black balls or two white balls. We then show mean errors for all informative

Belief	**Cognitive**	**Informative Signals**				**All Informative**	**Uninformative**	**All**
Elicitation	**Type**					**Signals**	**Signals**	**Signals**
Method		$\rho' \in \{0.16, 0.84\}$	$\rho' \in \{0.30, 0.70\}$	$\rho' \in \{0.31, 0.69\}$	$\rho' \in \{0.40, 0.60\}$	$\rho' \neq 0.5$	$\rho' = 0.5$	
SBDM	Consistent	10.10	9.58	14.46	11.40	11.01	8.29	10.37
Introspection	Consistent	15.37	15.75	14.38	12.12	14.35	6.62	12.62
- Permutation Test:		(p-value: 0.001)	(p-value: 0.004)	(p-value: 0.973)	(p-value: 0.748)	(p-value: 0.008)	(p-value: 0.283)	(p-value: 0.051)
SBDM	Inconsistent	19.91	17.55	17.60	15.61	17.25	14.43	16.63
Introspection	Inconsistent	18.31	19.75	20.10	16.79	18.53	8.53	16.26
- Permutation Test:		(p-value: 0.556)	(p-value: 0.300)	(p-value: 0.192)	(p-value: 0.630)	(p-value: 0.384)	(p-value: 0.003)	(p-value: 0.787)
SBDM	Full sample	14.31	12.88	16.07	13.50	13.90	11.02	13.24
Introspection	Full sample	16.65	17.51	17.30	14.50	16.33	7.54	14.35
- Permutation Test:		(p-value: 0.132)	(p-value: 0.003)	(p-value: 0.385)	(p-value: 0.549)	(p-value: 0.013)	(p-value: 0.007)	(p-value: 0.226)

Table 3: Mean error of reports under the SBDM mechanism and the Introspection mechanism for (i) consistent participants, (ii) inconsistent participants, and (iii) all participants combined. The reported p-values are based on permutation tests using 10,000 iterations in which the subset of participants is held fixed and participants are randomly allocated to the SBDM or Introspection mechanism in each iteration of a regression on the treatment effect. The null hypothesis is that the treatment coefficient is equal to 0 (i.e. that there is no difference in belief error between the SBDM and Introspection). The two-sided test statistic is reported.

signals combined and for the case of an uninformative signal. Finally, mean errors over all decision problems are shown in the last column.

In Section 3.3.1 we showed that the interaction effect is based on the difference between (i) the difference in mean errors of consistent and inconsistent types in the SBDM mechanism and (ii) the difference in mean errors of consistent and inconsistent types in the Introspection mechanism. As seen in the last column, the mean error for consistent participants in the SBDM mechanism is 10.37 while the mean error for inconsistent participants is 16.63. Thus, there is a -6.25 percentage point difference in means in the SBDM mechanism. In percentage terms, the mean error of a consistent participant is 37.6 percent smaller than an inconsistent participant in the SBDM mechanism.

The mean error for consistent participants in the Introspection mechanism is 12.62 while the mean error for inconsistent participants is 16.26. Thus, there is a -3.64 percentage point difference in means in the Introspection mechanism and the mean error of a consistent participant is only 22.4 percent smaller than an inconsistent participant. The difference-in-difference estimate of -2.61 $(-6.25 + 3.64)$ is significant using the one-sided synchronized test described in the last section (p-value = .027). The effect is also large in magnitude given that the mean error in the sample is 13.79.

We note that the difference-in-difference estimate is particularly large in decision problem with uninformative signals. In these problems, the difference-in-difference estimate is -4.23 and the effect is significant using the same one-sided synchronized test as above (p-value = .017). In these questions, there is no Bayesian updating necessary. Thus, identifying the correct belief is unlikely to be cognitively costly and the difference in belief errors is likely driven by inconsistent participants being confused by the SBDM mechanism itself. The difference-in-difference estimate is not significant when the sample is restricted to informative signals (p-value = .075).[16]

In Appendix F, we also report robustness results when we exclude outliers. If we remove individuals whose reports are almost always above or below 50, the difference-in-difference estimates become larger and the p-values fall. Thus, our results in this section do not appear to be the result of an allocation of outliers to treatments.

Turning to our second hypothesis, regarding task difficulty, we find:

Result 2 *Consistent with Hypothesis 2, the Introspection mechanism is more sensitive to task difficulty than the SBDM mechanism.*

Recall from the last section that our parameter of interest for Hypothesis 2 is the difference in mean errors between (i) informative and uninformative questions in the

[16] Following our pre-analysis plan, we tested for the interaction effect in informative signals by first calculating the mean error in Block 2 and the mean error for informative questions in Block 3 separately, and then taking the average of these two means. This approach reduces variation in errors caused by a different number of informative questions being asked to participants in Block 3.

Introspection mechanism and (ii) informative and uninformative questions in the SBDM mechanism. Referring back to Table 3 and looking at the rows corresponding to the full sample, the mean errors under Introspection is 7.54 when the signal is uninformative and 16.33 when the signal is informative. Mean errors under the SBDM mechanism are 11.02 in problems in which the signal is uninformative and 13.90 in problems in which the signal is informative. Thus, under Introspection, the difference in mean errors is 8.79 while it is 2.88 under SBDM. The difference-in-difference estimate of 5.91 is significant in the one-sided synchronized permutation test described in Section 3 (p-value $< .001$).

4.2 Focal Reports in the SBDM and Introspection Mechanisms

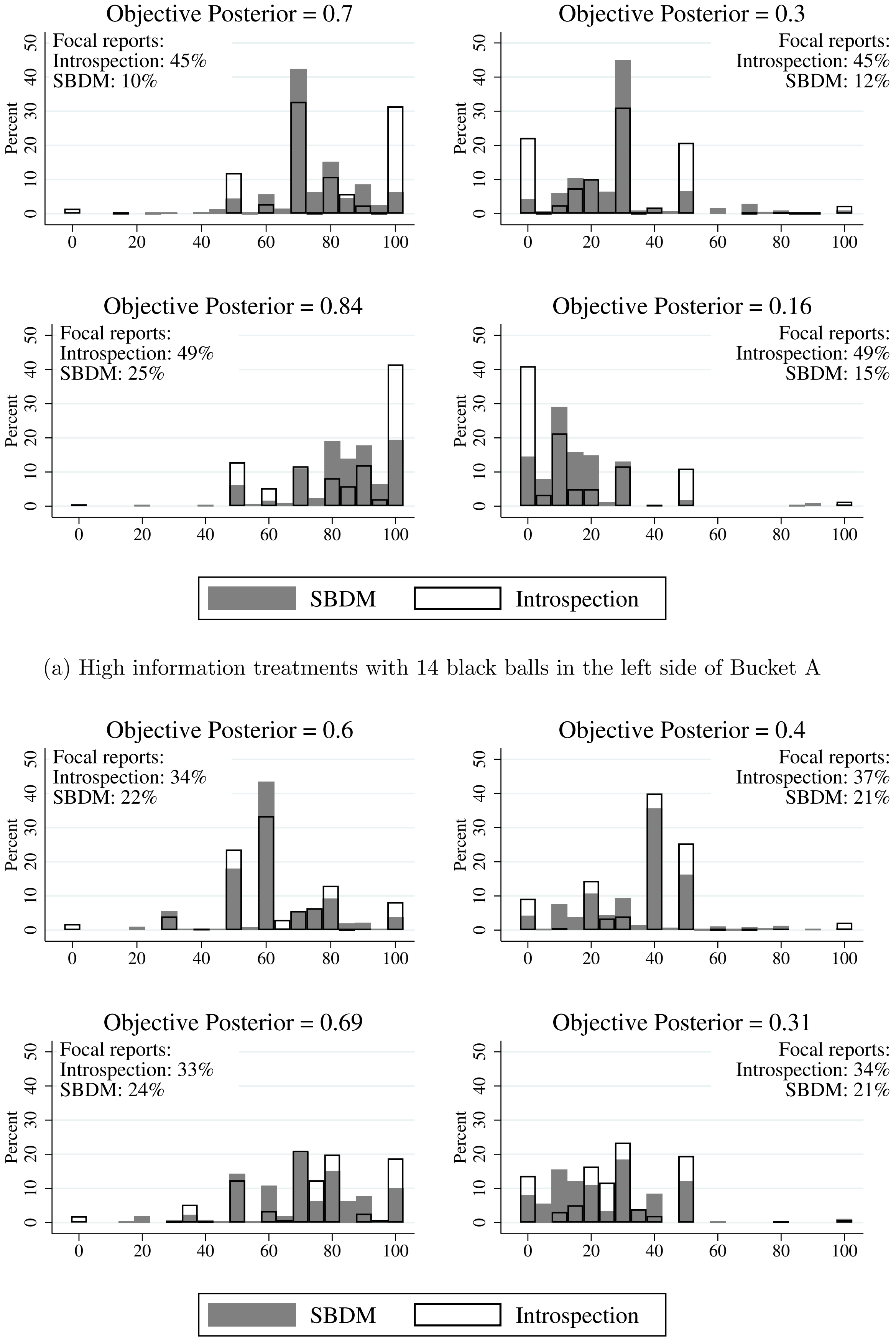
Having found evidence that the SBDM mechanism is more sensitive to heterogeneity in probabilistic reasoning, and that the Introspection mechanism is more sensitive to task difficulty, we now take a deeper look at the data to understand what is driving the differences in mechanism performance. We begin by comparing consistent participants' responses to both mechanisms when signals are informative.

Result 3 *In decision problems with an informative signal, consistent participants have significantly smaller belief errors in the SBDM mechanism than in the Introspection Mechanism. The difference is due in part to the larger number of focal reports observed in the Introspection mechanism.*

As seen by comparing the first two rows of Table 3, the SBDM is more accurate for consistent participants when we combine the data from all the informative priors (p-value = 0.008). Thus reports in the SBDM mechanism have lower mean errors than reports in the Introspection mechanism for consistent participants when signals are informative.

Figure 2 shows the distribution of reports for consistent participants for each of the eight informative signals under the SBDM mechanism and Introspection. Introspection has more focal reports of 0, 50, and 100 than the SBDM mechanism. Aggregating over the eight informative priors, focal reports by consistent participants occur in 41 percent of cases in the Introspection mechanism and in only 18 percent of cases in the SBDM mechanism. This difference is significant when we compare the average proportion of focal reports made in the two mechanisms in a permutation test using data from periods with informative signals (p-value < 0.001).

Excluding the focal reports, the mean error of consistent participants in the Introspection mechanism is 7.63 in periods with an informative signal while the mean error in the SBDM mechanism is 8.81 in the same periods. Thus, the larger number of focal reports in the Introspection mechanism appears to be the main driver of differences between the two mechanisms for consistent participants.



(a) High information treatments with 14 black balls in the left side of Bucket A

(b) Low information treatments with 12 black balls in the left side of Bucket A

Figure 2: Distribution of reported beliefs by consistent participants

Result 4 *In decision problems with an uninformative signal, there is no significant differences in mean errors between the SBDM mechanism and Introspection for consistent participants. However, consistent participants in the Introspection mechanism make significantly more correct and incorrect focal reports.*

In periods with uninformative signals, the mean error for consistent participants is 8.29 in the SBDM mechanism and 6.62 in the Introspection mechanism and there is no significant difference between the two mechanisms (p-value = 0.283). In the Introspection mechanism 72.97 percent of consistent participants report the correct belief of 50 while only 57.75 percent of consistent participants report the correct belief in the SBDM. This difference in correct focal reports is significant (p-value = 0.009). However, incorrect focal reports are also common under Introspection when signals are uninformative: 6.63 percent of reports are extreme reports of 0 or 100 in the Introspection mechanism, while 2.13 percent of reports are extreme reports of 0 of 100 in the SBDM mechanism (p-value = 0.041).

Result 5 *In decision problems with informative signals, there is no significant difference in mean errors between the SBDM mechanism and Introspection for inconsistent participants. However, in decision problems with an uninformative signal, inconsistent individuals have significantly smaller belief errors in the Introspection mechanism than in the SBDM mechanism.*

As seen by comparing rows 3 and 4 of Table 3, inconsistent participants have slightly lower errors in the SBDM mechanism than the Introspection mechanism in each of the four cases with informative signals. However, none of these differences are significant.

When the signals are uninformative, the mean error for inconsistent participants in the Introspection mechanism is 8.53 while the mean error in the SBDM mechanism is 14.43. This difference is significant in a permutation test (p-value = 0.003). The distribution of reports indicates a correct report of 50 is made in 73.70 percent of cases in the Introspection mechanism and in only 39.76 percent of cases in the SBDM mechanism. This difference is significant when we compare the average proportion of correct reports made in the two mechanisms in a permutation test using data from periods with uninformative signals (p-value < 0.001). The strong reduction in correct focal reports of 50 suggests that some individuals do not understand the truth-telling properties of the SBDM mechanism and misreport as a result.

4.3 Differences in the Initial Experiment and Follow-up Experiment

In our pre-analysis plan we committed to pooling the data from our original and follow-up experiments if there were no significant differences in errors in the full data set, the SBDM sample, or the Introspection sample. As seen in Appendix A, we find no differences in the samples along these dimensions and have therefore used the pooled data as the basis for our evaluation of Hypotheses 1 and 2. In this section we deviate from our pre-analysis plan to discuss an important difference in the two samples as they relate to Hypothesis 1.

Result 6 *The magnitude of the estimated interaction effect between the SBDM mechanism and probabilistic reasoning is much larger in the initial experiment than in the follow-up experiment.*

Tables 5 and 6 in Appendix A show the mean error of reports for the initial experiment and follow-up experiment separately. As seen in Table 5, in the initial experiment's Introspection treatment, the mean error of consistent participants is 14.58 and the mean error of inconsistent participants is 15.01. In the initial experiment's SBDM treatment, the mean error for consistent participants is 10.45 and the mean error for inconsistent participants is 16.30. Thus, for the initial experiment, the difference-in-difference estimate related to Hypothesis 1 is -5.42 $(10.45 - 16.30 - (14.58 - 15.01))$, which is significant in the one-sided synchronized permutation test described in Section 3 (p-value = .009).

By contrast, as seen in Table 6 in Appendix A, in the follow-up experiment's Introspection treatment, the mean error for consistent participants is 11.32 and the mean error for inconsistent participants is 17.01. In the follow-up experiment's SBDM treatment, the mean error for consistent participants is 10.32 while the mean error for inconsistent participants is 16.86. Thus, for the follow-up experiment, the difference-in-difference estimate related to Hypothesis 1 is -0.85 $(10.32 - 16.86 - (11.32 - 17.01))$, which is not significant in the one-sided synchronized permutation test described in Section 3 (p-value = 0.314).

As noted in Section 3.3.1, interpreting the difference-in-difference as a measure of SBDM-specific misreport-errors when using all decision problems relies on the assumption that any difference in errors between consistent and inconsistent participants that stem from inaccurate beliefs are similar for the two mechanisms. Thus, one potential reason for the difference in point estimates is that this assumption is violated in one of the two experiments.

To explore this issue, we also calculated the difference-in-difference estimates using only the easy decision problems in which signals were uninformative. These decision problems provide the cleanest estimate of SBDM-specific misreports because most individuals are likely to have correct latent beliefs. In these problems, the difference in point estimates diminishes but does not go away: in our initial treatment, the difference-in-difference estimate is -6.41 (p-value = .029), while in the follow-up experiment, the point estimate is -2.68 (p-value = 0.150).

A second potential reason for the difference in point estimates are changes to the experimental environment. Covid-19 restrictions prevented us from using the lab and our follow-up experiments were conducted online. While we worked hard to maintain identical protocols in the two experiments, it is possible that the online environment generated new sources of errors. As discussed in Appendix F, we find some evidence that this may be the case. In the follow-up data, there are a number of participants who appear to be reporting their beliefs out of 40 (the total number of balls in the bucket) rather than 100. A conservative removal of the 16 most extreme outliers (individuals whose reports almost always fell below 50 or above 50) increases the magnitude of the difference-in-difference from -0.85 to -1.74 (p-value = 0.163). However, this estimate is still smaller in magnitude than the point estimate from our original experiment using the same criterion for removing outliers (-5.45; p-value = .010).[17]

Thus, restricting attention to easy decision problems or controlling for outliers in the follow-up sub-sample results in point estimates that are similar to the pooled difference-in-difference estimate that we use throughout the paper. We cannot, however, fully explain the difference between the original experiment and follow-up experiment. We hope that future replications will be conducted that can improve our understanding of this issue and to understand if there are systematic differences in how belief elicitation mechanisms and incentives interact with lab and online environments.

4.4 Fluid Intelligence and Cognitive Effort

In our follow-up experiment, we divided participants into high-ability and low-ability groups based on their performance on a short-form version of the Raven's Advanced Progressive Matrices task, and high-effort and low-effort groups based on their performance on an extended Cognitive Reflection Test. Our pre-analysis plan predicted the following hypotheses:

Hypothesis 3 *The SBDM mechanism is more sensitive to variation in fluid intelligence than the Introspection Mechanism.*

Hypothesis 4 *The SBDM mechanism is more sensitive to variation in cognitive effort than the Introspection Mechanism.*

To test for these hypotheses, we repeated the analysis we used to test for Hypothesis 1, but split groups based on their classification in the Raven task and the CRT. Our pre-analysis plan called for a one-sided test with a greater difference in errors between high and low types in the SBDM mechanism than in the Introspection mechanism.

Result 7 *Both fluid intelligence and cognitive effort strongly predict errors in both the SBDM and Introspection Mechanisms. However, there is no significant difference in sensitivity to fluid intelligence nor to cognitive effort.*

[17] Removing outliers and looking only at decision problems with an uninformative signal leads to a difference-in-difference estimate of -3.63 (p-value = 0.084) in the follow-up experiment and -7.13 (p-value = 0.018) in the original experiment.

Support for Result 7 is given in Table 4. Panel A of this table reports mean errors under the SBDM mechanism and the Introspection mechanism for (i) high-ability participants and low-ability participants. We first report errors for all decision problems that were informative and for all decision problems that were uninformative. In the last column, we report the mean error on all decision problems. Panel B is identical to Panel A except that it divides individuals into high-effort and low-effort groups based on the extended CRT.

Table 4: Fluid Intelligence and Cognitive Effort

Panel A: Fluid Intelligence

Belief Elicitation Method	**Cognitive Type**	**All Informative Signals** $\rho' \neq 0.5$	**Uninformative Signals** $\rho' = 0.5$	**All Signals**
SBDM	High	11.70	9.24	11.12
Introspection	High	12.33	2.53	10.04
- Permutation Test:		(p-value 0.684)	(p-value 0.000)	(p-value 0.435)
SBDM	Low	17.70	12.00	16.45
Introspection	Low	19.81	11.14	17.98
- Permutation Test:		(p-value 0.272)	(p-value 0.751)	(p-value 0.399)

Panel B: Cognitive Effort

Belief Elicitation Method	**Cognitive Type**	**All Informative Signals** $\rho' \neq 0.5$	**Uninformative Signals** $\rho' = 0.5$	**All Signals**
SBDM	High	10.74	8.37	10.18
Introspection	High	11.53	4.76	10.00
- Permutation Test:		(p-value 0.620)	(p-value 0.045)	(p-value 0.900)
SBDM	Low	18.54	12.97	17.30
Introspection	Low	21.82	9.09	19.03
- Permutation Test:		(p-value 0.044)	(p-value 0.116)	(p-value 0.272)

Reported p-values in both panels are based on permutation tests using 10,000 iterations in which the subset of participants is held fixed and participants are randomly allocated to the SBDM or Introspection mechanism in each iteration of a regression on the treatment effect. The null hypothesis is that the treatment coefficient is equal to 0 (i.e. that there is no difference in belief error between the SBDM and Introspection). The two-sided test statistic is reported.

As seen in Panel A, mean error in the SBDM mechanism is 11.12 for high-ability participants and 16.45 for low-ability participants. The mean error in the Introspection mechanism is 10.04 for high-ability participants and 17.98 for low-ability participants. The difference-in-difference estimate is therefore 2.61, which has the opposite sign from the one predicted in Hypothesis 3; it is not significant using the one-sided synchronized test (p-value = .914).

As seen in Panel B, mean error in the SBDM mechanism is 10.18 for high-effort participants and 17.30 for low-effort participants. The mean error in the Introspection mechanism is 10.00 for high-effort participants and 19.03 for low-effort participants. The difference-in-difference estimate is therefore 1.91, which has the opposite sign from the one predicted in Hypothesis 4 and is again not significant (p-value = .868).

Although we do not observe an interaction effect for either case, we note that low-ability and low-effort individuals have very large errors relative to high-ability and high-effort ones. Thus, while we do not find a significant difference in sensitivity to different belief elicitation mechanisms, both Fluid Intelligence and Cognitive Effort are strongly predictive of belief errors. This finding is highly consistent with earlier papers in which Raven test scores have been found to correlate positively with fewer Bayesian updating errors (Charness et al., 2018) and with more accurate beliefs (Burks et al., 2009). It is also consistent with the finding of Schlag and Tremewan (2019) that focal reporting in the SBDM is correlated with scores on the Cognitive Reflection Test.

5 Discussion and Conclusion

The Stochastic Becker-DeGroot-Marschak (SBDM) mechanism is a theoretically elegant way of eliciting incentive-compatible beliefs under a variety of risk preferences. However, the mechanism is complex and there is concern that some participants may misunderstand its incentive properties. We use a two-part design in which we identify participants whose decision-making is consistent and inconsistent with probabilistic reasoning, and elicit their beliefs in both easy and hard decision problems. Relative to Introspection, there is less variation in mean belief errors between easy and hard problems in the SBDM mechanism. However, there is a greater difference in belief errors between consistent and inconsistent participants. These results suggest that while the SBDM mechanism encourages individuals to think more carefully about beliefs, it is more sensitive to probabilistic reasoning skills. Our results show that mechanism complexity is an important consideration when using elicitation mechanisms, and identifies probabilistic reasoning as an important consideration when interpreting elicited beliefs.

By identifying different channels by which errors occur in the two mechanism, we can better understand the mixed results from earlier studies that compare the two mechanisms. In particular, our finding that errors in the Introspection mechanism varies with task difficulty implies that any horse race between the two mechanisms is likely to be strongly task dependent and that task difficulty may be an important consideration in whether to offer explicit incentives for beliefs.

Our finding that errors in both the SBDM and Introspection mechanism vary with participants' probabilistic reasoning ability, fluid intelligence, and cognitive effort suggests that researchers should be cautious when using individual beliefs to identify types. For example, in the literatures on overconfidence, it is common to use the difference between an agent's true ability and their reported belief about this ability as a proxy for overconfidence. If errors are correlated with cognitive ability, then individuals who are assigned to the overconfident group may also include a large set of low-ability types who struggle to optimize in other situations.

We see value in an independent replication of our experiments and in understanding whether there are differences in how belief elicitation mechanisms interact with lab and online environments. As seen in Section 4.3, the point estimate for the interaction effect related to probabilistic reasoning is large and significant in our initial lab-based experiments but small and not significant in our online follow-up experiments. Further, while the magnitude of the interaction effect in the follow-up experiments increases when the data is restricted to easy decision problems or when the most obvious outliers are removed, a difference in the magnitude of the interaction effect between the two samples persists.

It is an open question as to how to improve the SBDM to reduce the impact of probabilistic reasoning. Holt and Smith (2016) and Burfurd and Wilkening (2018) suggest that choice lists can reduce focal reports but neither paper finds accuracy improvements from using multiple choice lists. Nonetheless, choice lists may be important for a subset of individuals and it would be interesting to understand how they interact with cognition.

As an alternative, Hao and Houser (2012) suggests that combinatorial clocks might play an important role if researchers can overcome the censoring which results from using a single ascending clock. This result is consistent with the notion of obviously strategy-proof mechanisms (Li, 2017). One potential solution would be to conduct both an ascending and decreasing clock auctions against a dummy player with a common cutoff point $\hat{p}$ and pay the participant for the outcome of one of these clock auctions. In the ascending clock auction, the clock probability z goes from zero to one and the participant receives $H_A L$ if z reaches $\hat{p}$. If the participant drops out, she receives $H_{\hat{p}}L$. In the descending clock auction, the participant receives $H_{\hat{p}}L$ if z reaches $\hat{p}$ and she receives $H_A L$ if she drops out. In both mechanisms, it is a dominant strategy to drop out at one's true value. Further, at least one of the two mechanisms will have no censoring.

It is also an open question as to how probabilistic reasoning and cognition interacts with other elicitation methods, particularly those that are robust to heterogeneity in risk preferences. One alternative to the SBDM is to combine quadratic-scoring rules with a binary lottery procedure, which theoretical induces risk neutrality under subjective expected utility. This binary lottery procedure has been found to generate better data than the quadratic-scoring rule in objective settings (Harrison et al., 2013; Hossain and Okui, 2013), and subjective settings (Harrison and Phillips, 2014; Harrison et al., 2014, 2015, 2017), but not in settings in which subjective beliefs about others is elicited (Koh, 2017). A recent paper by Danz et al. (2020) finds that transparent information on the incentives of the binary lottery procedure actually increases belief errors, suggesting that cognition may also be important for this mechanism. A second alternative is the frequency method of Schlag and Tremewan (2019), which elicits beliefs in terms of natural frequencies and can be used when multiple realisations of an outcome are available.

Bibliography

Allen, F. (1987). Notes—discovering personal probabilities when utility functions are unknown. *Management Science 33*(4), 542–544.

Andersen, S., J. Fountain, G. W. Harrison, and E. E. Rutström (2014). Estimating subjective probabilities. *Journal of Risk and Uncertainty 48*(3), 207–229.

Basso, D., F. Pesarin, L. Salmaso, and A. Solari (2009). *Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applicaitons with R.* Springer Science+Business Media, LLC 2009.

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

Bors, D. A. and T. L. Stokes (1998). Raven's advanced progressive matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement 58*(3), 382–395.

Brañas-Garza, P., T. Garcia-Munoz, and R. H. González (2012). Cognitive effort in the beauty contest game. *Journal of Economic Behavior & Organization 83*(2), 254–260.

Brañas-Garza, P., P. Kujal, and B. Lenkei (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics 82*, 306–307.

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

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences 106*(19), 7745–7750.

Cacioppo, J. T., R. E. Petty, and C. Feng Kao (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment 48*(3), 306–307.

Carpenter, J., M. Graham, and J. Wolf (2013). Cognitive ability and strategic sophistication. *Games and Economic Behavior 80*, 115–130.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology 54*(1), 1–22.

Charness, G., E. Karni, and D. Levin (2007). Individual and group decision making under risk: An experimental study of Bayesian updating and violations of first-order stochastic dominance. *Journal of Risk and Uncertainty 35*(2), 129–148.

Charness, G. and D. Levin (2005). When optimal choices feel wrong: A laboratory study of Bayesian updating, complexity, and affect. *The American Economic Review 95*(4), 1300–1309.

Charness, G., A. Rustichini, and J. van de Ven (2018). Self-confidence and strategic behavior. *Experimental Economics 21*, 72–98.

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

Cokely, E. T., M. Galesic, E. Schulz, S. Ghazal, and R. Garcia-Retamero (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making 7*(1), 25–47.

Dal Bó, E., P. Dal Bó, and E. Eyster (2017). The demand for bad policy when voters underappreciate equilibrium effects. *The Review of Economic Studies 85*(2), 964–998.

Danz, D., L. Vesterlund, and A. Wilson (2020). Belief elicitation: Limiting truth telling with information on incentives. Technical report, CESifo Working Paper 8048.

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

Fagerland, M. W. and L. Sandvik (2009). The Wilcoxon–Mann–Whitney test under scrutiny. *Statistics in Medicine 28*(10), 1487–1497.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives 19*(4), 25–42.

Freeman, D. J. and G. Mayraz (2019). Why choice lists increase risk taking. *Experimental Economics 22*, 131–154.

Gigerenzer, G. (1984). External validity of laboratory experiments: The frequency-validity relationship. *The American Journal of Psychology 97*, 185–195.

Gigerenzer, G. and U. Hoffrage (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review 102*, 684–704.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer Science+Business Media, New York; 2nd Edition.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association 1*(1), 114–125.

Grether, D. M. (1981). Financial incentive effects and individual decision-making. Technical report, California Institute of Technology, Working Papers 401.

Guerra, G. and D. J. Zizzo (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior & Organization 55*(1), 25–30.

Hahn, S. and L. Salmaso (2017). A comparison of different sychronized permutation approaches to testing effects in two-level two-factor unablanced ANOVA designs. *Statistical Papers 58*(2), 123–146.

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

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

Harrison, G. W., J. Martínez-Correa, and J. T. Swarthout (2014). Eliciting subjective probabilities with binary lotteries. *Journal of Economic Behavior & Organization 101*, 430–448.

Harrison, G. W., J. Martínez-Correa, and J. T. Swarthout (2015). Eliciting subjective probabily distributions with binary lotteries. *Economics Letters 101*, 68–71.

Harrison, G. W., J. Martínez-Correa, and J. T. Swarthout (2017). Scoring rules for subjective probability distributions. *Journal of Economic Behavior & Organization 134*, 430–448.

Harrison, G. W. and R. D. Phillips (2014). Subjective beliefs and statistical forecasts of financial risks: The chief risk officer project. In T. Andersen (Ed.), *Contemporary Challenges in Risk Management*, pp. 163–202. Palgrave Macmillan, London.

Heß, S. (2017). Randomization inference with Stata: A guide and software. *Stata Journal 17*(3), 630–651.

Hollard, G., S. Massoni, and J.-C. 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, C. A. (2006). *Markets, Games, and Strategic Behavior: Recipes for Interactive Learning.* Pearson Addison Wesley.

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

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

Horn, J. L. and R. B. Cattell (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology 57*(5), 253.

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

Huepe, D., M. Roca, N. Salas, A. Canales-Johnson, Á. A. Rivera-Rei, L. Zamorano, A. Concepción, F. Manes, and A. Ibañez (2011). Fluid intelligence and psychosocial outcome: From logical problem solving to social adaptation. *PLoS One 6*(9), e24858.

Juanchich, M., C. Dewberry, M. Sirota, and S. Narendran (2016). Cognitive reflection predicts real-life decision outcomes, but not over and above personality and decision-making styles. *Journal of Behavioral Decision Making 29*(1), 52–59.

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

Koh, B. H. (2017). Belief elicitation: QSR vs. BSR. In *Essays in Leadership and Contests.* University of Melbourne PhD Thesis.

Li, S. (2017). Obviously strategy-proof mechanisms. *American Economic Review 107*(11), 3257–3287.

Li, Y., M. Baldassi, E. J. Johnson, and E. U. Weber (2013). Complementary cognitive capabilities, economic decision making, and aging. *Psychology and Aging 28*(3), 595.

Lilleholt, L. (2019). Cognitive ability and risk aversion: A systematic review and meta analysis. *Judgment & Decision Making 14*(3).

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

Manly, B. F. (2007). *Randomization Bootstrap and Monte Carlo Methods in Biology. Texts in Statistical Sciences, Third Edition.* CRC Press INC, New York.

Montgomery, D. C. (2017). *Design and Analysis of Experiments, 9th Edition.* John Wiley & Sons, Inc. Hoboken, NJ.

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

Obrecht, N. A., G. B. Chapman, and R. Gelman (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition 37*(5), 632–643.

Oechssler, J., A. Roider, and P. W. Schmitz (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization 72*(1), 147–152.

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

Pauly, M., E. Brunner, and F. Konietschke (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society. Series B (Statistical*

Methodology) 77(2), 461–473.

Perasin, F. (2001). *Multivariate Permutation Tests with Applications in Biostatistics.* Wiley & Sons, Chichester.

Primi, C., K. Morsanyi, F. Chiesi, M. A. Donati, and J. Hamilton (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making 29*(5), 453–469.

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

Salmaso, L. (2003). Synchronized permutation tests in 2^k factorial designs. *Communications in Statistics - Theory and Methods 32*(7), 1419–1437.

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

Schipolowski, S., O. Wilhelm, and U. Schroeders (2014). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence 46*, 156–168.

Schlag, K. H. and J. Tremewan (2019). Simple belief elicitation: An experimental evaluation. Technical report, mimeo.

Schlag, K. H., J. Tremewan, and J. J. van der Weele (2013). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics 18*(3), 1–34.

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

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

Stanovich, K. E. and R. F. West (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences 23*(5), 645–665.

Thomson, K. S. and D. M. Oppenheimer (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making 11*(1), 99–113.

Toplak, M. E., R. F. West, and K. E. Stanovich (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition 39*(7), 1275.

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

Winkler, R. L. and A. H. Murphy (1968). "Good" probability assessors. *Journal of Applied Meteorology 7*(5), 751–758.

Appendix A: Data Aggregation

Data from the initial and follow-up experiments was tested to identify any statistically significant differences before pooling. Using participants' mean belief errors as the basis for comparison, and clustering at the participant level, there were no statistically significant differences across the full data set (p-value = 0.704), within the SBDM sample (p-value = 0.968), within the Introspection sample (p-value = 0.614), within the high information treatments with posteriors of $\rho = 0.7$ and 0.3 (p-value = 0.773) or within the low information treatments with posteriors of $\rho = 0.6$ and 0.4 (p-value = 0.805). As the data sets are not statistically significantly different in any dimension of our analysis, we use the pooled data set whenever possible.

Histograms comparing data in the initial and follow-up experiments are presented in Figures 3, 4 and 5. For completeness, we have also replicated Table 3 from the main text using data only from the initial experiment (Table 5) and only from the follow-up experiment (Table 6).

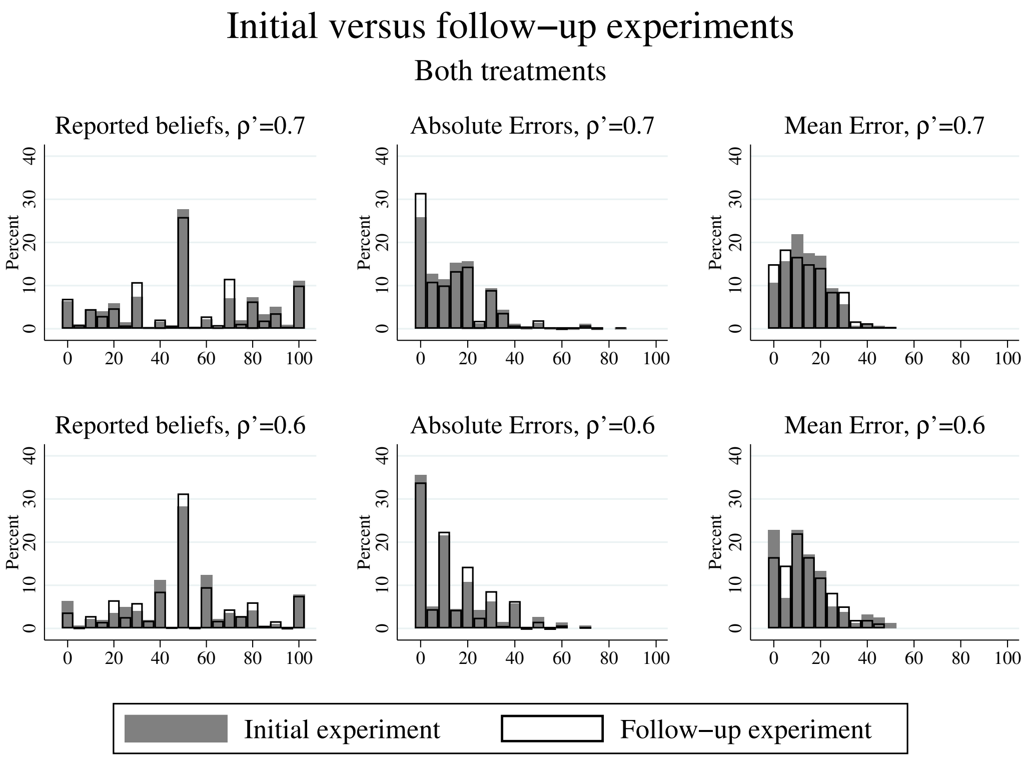


Figure 3: Comparison of reporting behavior in initial experiment and follow-up experiment: Pooled data from both the SBDM Treatment and the Introspection Treatments

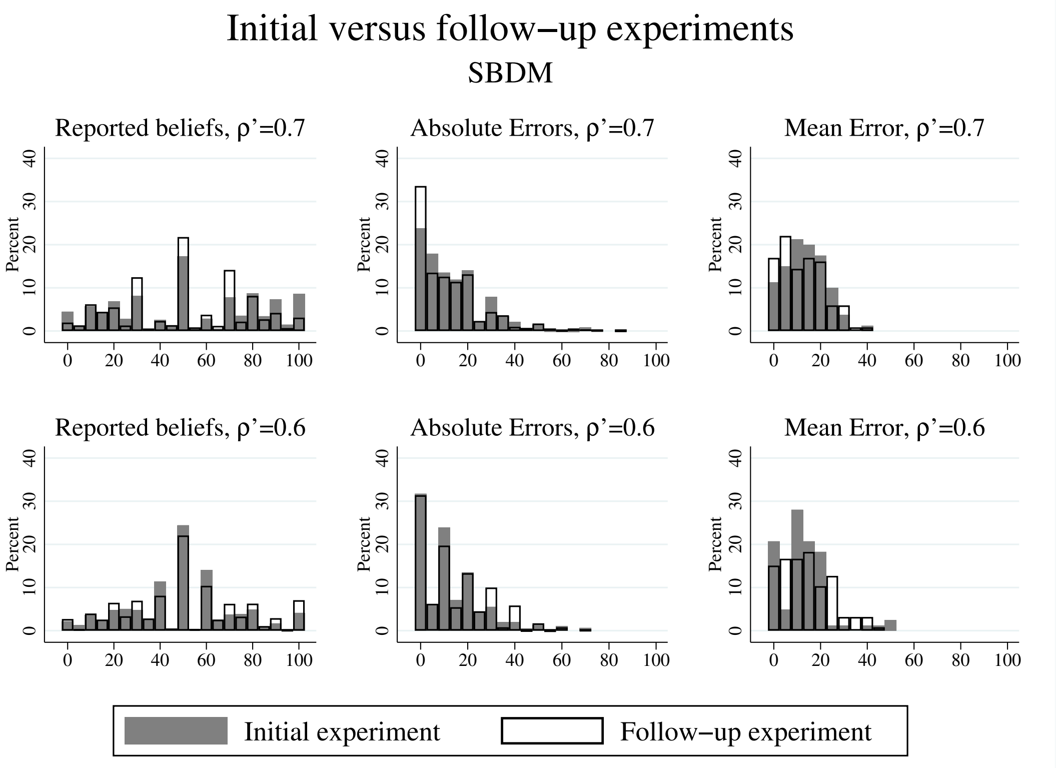


Figure 4: Comparison of reporting behavior in initial experiment and follow-up experiment: Data from SBDM Treatment

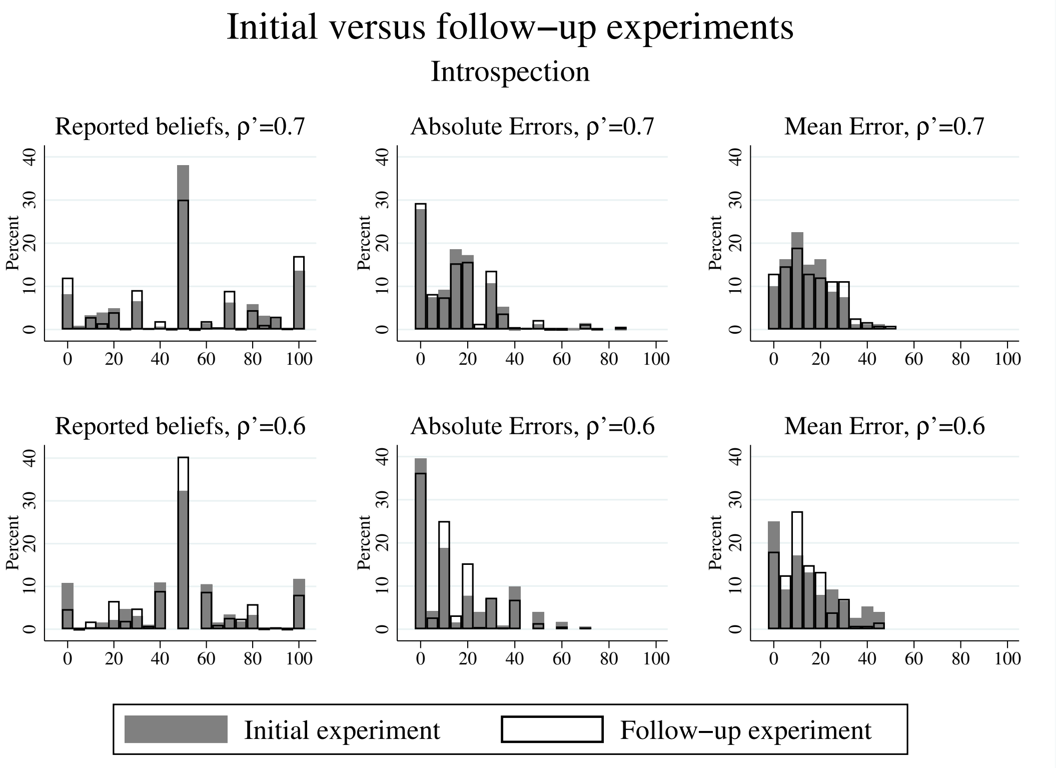


Figure 5: Comparison of reporting behavior in initial experiment and follow-up experiment: Data from Introspection Treatment

Belief Elicitation Method	**Cognitive Type**	**Informative Signals**				**All Informative Signals**	**Uninformative Signals**	**All Signals**
		$\rho' \in \{0.16, 0.84\}$	$\rho' \in \{0.30, 0.70\}$	$\rho' \in \{0.31, 0.69\}$	$\rho' \in \{0.40, 0.60\}$	$\rho' \neq 0.5$	$\rho' = 0.5$	
SBDM	Consistent	8.66	11.46	14.49	9.10	10.57	10.06	10.45
Introspection	Consistent	15.34	17.02	15.26	15.37	15.93	9.82	14.58
- Permutation Test:		(p-value: 0.003)	(p-value: 0.120)	(p-value: 0.837)	(p-value: 0.106)	(p-value: 0.013)	(p-value: 0.934)	(p-value 0.049)
SBDM	Inconsistent	16.30	18.22	19.30	14.44	16.85	14.42	16.30
Introspection	Inconsistent	16.91	18.68	19.39	15.06	17.26	7.77	15.01
- Permutation Test:		(p-value: 0.833)	(p-value: 0.881)	(p-value: 0.978)	(p-value 0.894)	(p-value: 0.855)	(p-value: 0.042)	(p-value: 0.535)
SBDM	Full Sample	12.16	14.84	16.80	11.58	13.60	12.15	13.27
Introspection	Full Sample	15.97	17.73	17.34	15.22	16.54	8.84	14.78
- Permutation Test:		(p-value: 0.053)	(p-value: 0.236)	(p-value: 0.833)	(p-value 0.203)	(p-value: 0.066)	(p-value: 0.136)	(p-value: 0.303)

Table 5: Mean belief errors in the initial experiment under the SBDM mechanism and the Introspection mechanism for (i) consistent participants, (ii) inconsistent participants, and (iii) both consistent and inconsistent participants combined. The reported p-values are based on permutation tests using 10,000 iterations in which the subset of participants is held fixed and participants are randomly allocated to the SBDM or Introspection mechanism in each iteration of a regression on the treatment effect. The null hypothesis is that the treatment coefficient is equal to 0 (i.e. that there is no difference in accuracy between the SBDM and Introspection). The two-sided test statistic is reported.

Belief	**Cognitive**	**Informative Signals**				**All Informative**	**Uninformative**	**All**
Elicitation	**Type**					**Signals**	**Signals**	**Signals**
Method		$\rho' \in \{0.16, 0.84\}$	$\rho' \in \{0.30, 0.70\}$	$\rho' \in \{0.31, 0.69\}$	$\rho' \in \{0.40, 0.60\}$	$\rho' \neq 0.5$	$\rho' = 0.5$	
SBDM	Consistent	11.21	8.59	14.43	13.09	11.28	7.25	10.32
Introspection	Consistent	15.38	14.84	13.86	10.13	13.29	4.51	11.32
- Permutation Test:		(p-value: 0.045)	(p-value: 0.016)	(p-value: 0.812)	(p-value: 0.259)	(p-value: 0.195)	(p-value: 0.104)	(p-value: 0.447)
SBDM	Inconsistent	23.13	16.91	16.64	16.27	17.54	14.44	16.86
Introspection	Inconsistent	19.13	20.46	20.51	17.79	19.28	9.02	17.01
- Permutation Test:		(p-value: 0.338)	(p-value: 0.245)	(p-value: 0.100)	(p-value: 0.606)	(p-value: 0.371)	(p-value: 0.034)	(p-value: 0.939)
SBDM	Full sample	15.86	11.55	15.61	14.76	14.09	10.27	13.22
Introspection	Full sample	17.10	17.36	17.28	14.08	16.19	6.68	14.07
- Permutation Test:		(p-value: 0.587)	(p-value: 0.005)	(p-value: 0.329)	(p-value: 0.741)	(p-value: 0.103)	(p-value: 0.016)	(p-value: 0.461)

Table 6: Mean belief errors in the follow-up experiment under the SBDM mechanism and the Introspection mechanism for (i) consistent participants, (ii) inconsistent participants, and (iii) all participants combined. The reported p-values are based on permutation tests using 10,000 iterations in which the subset of participants is held fixed and participants are randomly allocated to the SBDM or Introspection mechanism in each iteration of a regression on the treatment effect. The null hypothesis is that the treatment coefficient is equal to 0 (i.e. that there is no difference in belief error between the SBDM and Introspection). The two-sided test statistic is reported.

Appendix B: Observer Effects

An observer effect is a situation in which the introduction of belief elicitation helps participants perform optimally in the underlying task. There is mixed evidence that belief elicitation affects game play in other settings. Rutström and Wilcox (2009) compare behavior when participants do not have beliefs elicited, when participants participate in unpaid Introspection, and when participants' beliefs are elicited using the Quadratic Scoring Rule (QSR). They test the idea that the cognitively "intrusive" QSR might drive a sharper wedge between the "affective process" of belief formation and the "deliberative judgement" reporting process. In the case of the QSR they ultimately reject the hypothesis that belief elicitation does not affect game play, although they note than these effects are concentrated in earlier periods. Nyarko and Schotter (2002) also find that belief elicitation has unintended consequences, and that participants exposed to belief elicitation are more likely to use mixed strategies than pure strategies. However, Guerra and Zizzo (2004) and others have found no evidence that elicitation affects decision-making.

Since the instructions for the SBDM require an explicit discussion of probability and incentives, we predicted that this mechanism would generate an observer effect and that this effect could reduce belief errors in the belief elicitation task. To test for observer effects, we use data from the initial experiment to compare the proportion of correct left/right choices in our no-elicitation treatment to the proportion of correct left/right choices in the SBDM mechanism and Introspection mechanism.

Result 8 *There is no statistically significant observer effect in the data.*

Support for Result 8 is provided in Table 7, which shows the proportion of correct left/right choices in blocks one and two of the experiment with the data split into subsets of 10 periods. We focus the analysis on Periods 11-30 since these are the ten periods directly before and after the introduction of beliefs.

An observer effect would create larger improvements in the proportion of correct left/right choices at the start of Block Two in the treatments with belief elicitation relative to the No-Elicitation treatment. There is no such pattern in the data: in the treatments with no belief elicitation, participants make mistakes in 35.1 percent of cases in Periods 11-20 and in 30.5 percent of cases in Periods 21-30. This difference of 4.63 percentage points is not significantly different to the difference of 4.7 percentage points observed in the SBDM mechanism in a difference-in-difference permutation test in which we restrict data to periods 11-30 (p-value = 0.898). It is also not different to the difference of 7.3 percentage points observed in the Introspection mechanism (p-value = 0.533).

In Block Two, the proportion of incorrect left/right decisions in the SBDM mechanism are not significantly different than the Introspection mechanism (p-value = 0.761).

	Block One		Block Two	
	Periods 1-10	**Periods 11-20**	**Periods 21-30**	**Periods 31-40**
No Belief Elicitation	0.405	0.351	0.305	0.346
SBDM	0.335	0.340	0.293	0.275
Introspection	0.400	0.347	0.274	0.274

Table 7: Proportion of incorrect left/right decisions in initial experiment

Appendix C: Additional Figures

Result 3 presented histograms of reported beliefs for consistent participants across all of the informative priors. Here we provide the histograms of reported beliefs for the other cases. Figure 6 shows the reported beliefs of consistent and inconsistent participants in the case of an uninformative signal for both the SBDM mechanism and the Introspection mechanism using data from both the high-information treatments with 14 black balls in the left side of Bucket A and the low-information treatments with 12 black balls. Figure 7 shows the reported beliefs for the inconsistent participants across the eight potential informative posteriors.

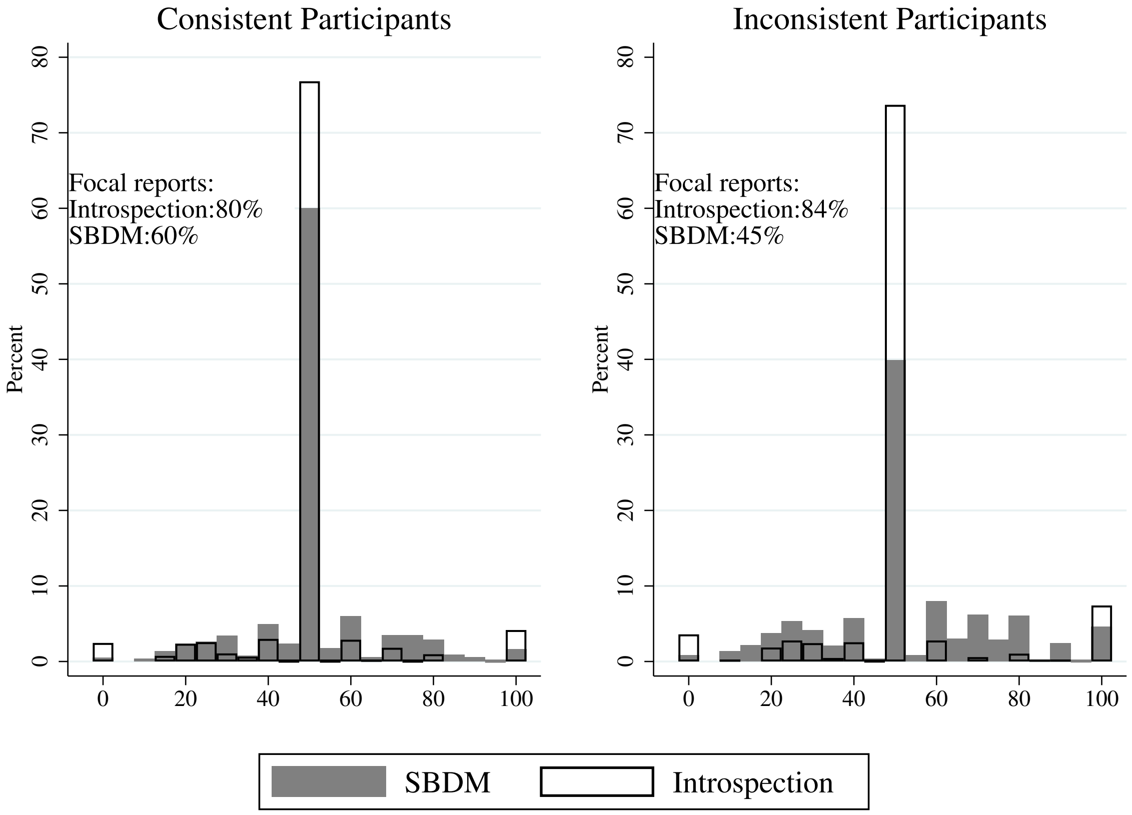
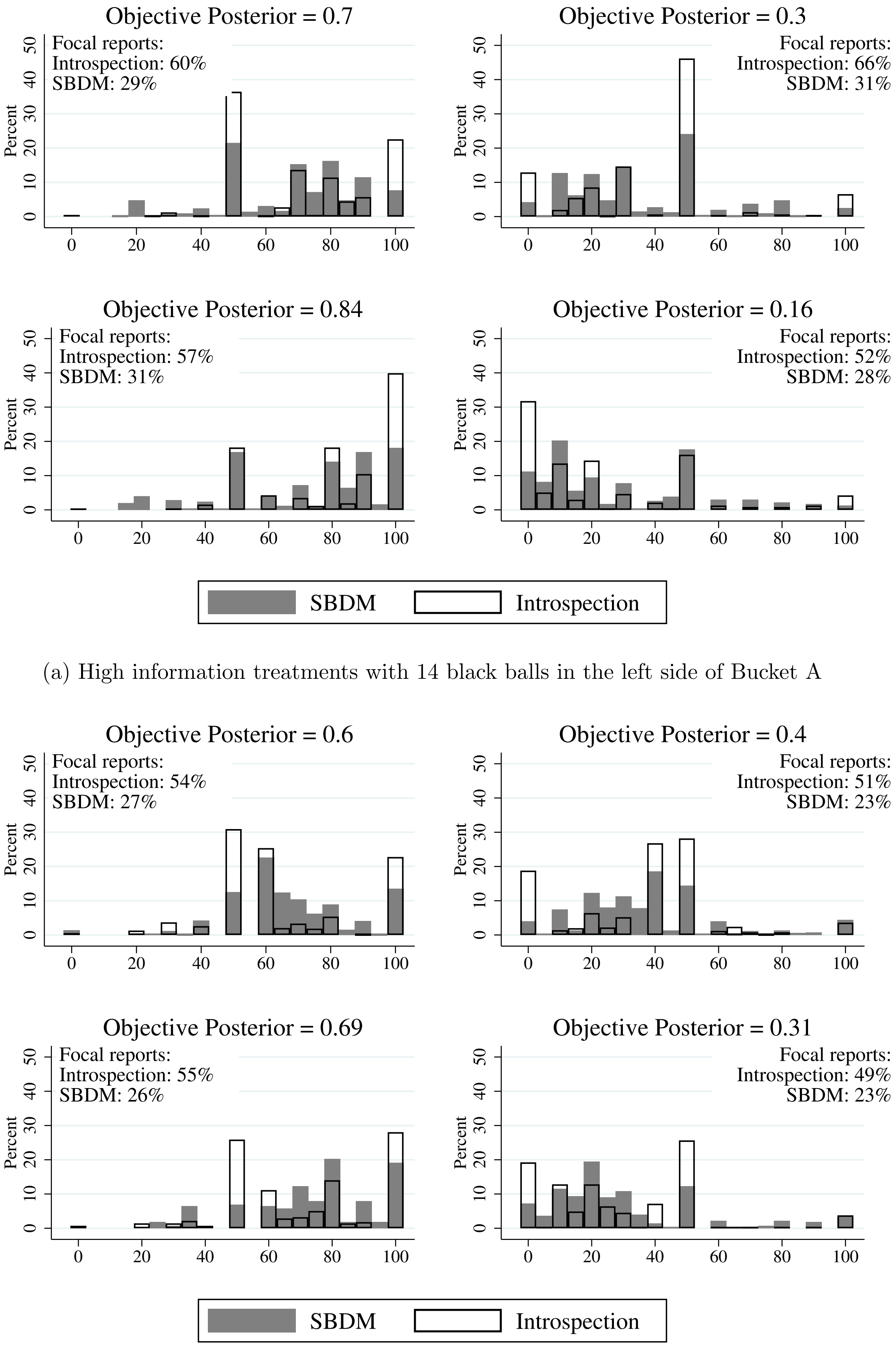


Figure 6: Distribution of reported beliefs when the posterior is 0.50



(a) High information treatments with 14 black balls in the left side of Bucket A

(b) Low information treatments with 12 black balls in the left side of Bucket A

Figure 7: Distribution of reported beliefs by inconsistent participants

Appendix D: Permutation Tests for Interactions

In this Appendix we briefly outline the Synchronized Permutation test of Perasin (2001) and Salmaso (2003) that we used to test for the interaction effect in Hypotheses 1 and 2. A more general introduction to permutation tests can be found in Good (2000) and Manly (2007). More details on Synchronized Permutation tests can be found in Basso et al. (2009) and Hahn and Salmaso (2017).

Hypotheses 1 and 2 both use a 2×2 factorial design. We are primarily interested in the interaction effect between factors. The standard approach would be to use a parametric ANOVA specification. However, as seen in the main text, the error distribution in the data is not normally distributed and thus the underlying assumption of parametric ANOVA is not satisfied. The permutation test represents an ideal alternative since it requires only a minimal assumptions about the errors, is exact in some cases, and has high power relative to other approaches.

The main assumption of permutation tests is that the data is exchangeable under the null hypothesis. Data is exchangeable if the probability of the observed data is invariant with respect to random permutations of the indexes (Basso et al., 2009). In the 2×2 factor design, the observations are typically not exchangeable since units assigned to different treatments have different expectations. This implies that approaches that freely permute data across cells may fail to separate main and interaction effects (Good, 2000). The synchronized permutation test of Perasin (2001) and Salmaso (2003) restricts permutations to the same level of a factor to generate test statistics for both main factors and, separately, interactions that depend only on the effect being tested and a combination of errors (Basso et al., 2009).

For clarity, we will concentrate the discussion on Hypothesis 1, in which each observation E_{ijk}, represents the mean error of an individual who has been assigned to mechanism $i = \{1, 2\}$ and who is of cognitive ability $j = \{1, 2\}$. We note that permutation tests will assign all observations of an individual to the same factor combination, which we refer to as a cell. Thus, there is no loss in power in using average effort as our dependent variable rather than treating each decision made by an individual as an observation.

Following the main text, we assume that each observation can be decomposed into a mean, two main effects, an interaction, and an error term:

$$E_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \tag{3}$$

in which $i = \{1, 2\}$ is the belief elicitation mechanism assigned to an individual, $j = \{1, 2\}$ is the cognitive ability of the individual, and $k = \{1, \dots, n_{ij}\}$ is the index of an observation within a treatment cell E_{ij}.

By including the additive constant μ, all main effects and interactions in the model can be defined to sum to zero. Thus, we assume that $\alpha_1 + \alpha_2 = 0$, $\beta_1 + \beta_2 = 0$, $(\alpha\beta)_{i1} + (\alpha\beta)_{i2} =$

0 for all i, and $(\alpha\beta)_{1j}+(\alpha\beta)_{2j}=0$ for all j. In this construction, $\alpha_1=-\alpha_2$ and thus, under the null of no effect of the mechanism on errors, each of the main effects $\alpha_1=\alpha_2=0$. Under the alternative, α_1 represents the difference from a zero average, and the interaction term $(\alpha\beta)_{ij}$ represents the deviation from the sum $\alpha_i+\beta_j$. The model assumes that errors ϵ_{ijk} are *exchangeable* and $\mathbb{E}(\epsilon_{ijk})=0$. Errors are exchangeable if the probability of the observed error is invariant with respect to random permutation of the data (Basso et. al, 2009).

We begin by considering a balanced design in which all cells have n observations and first construct a statistic for comparing the first factor (i.e., the mechanism) at each of the two levels of the second. Let $T_{A|1}=\sum_k E_{11k}-\sum_k E_{12k}$ and $T_{A|2}=\sum_k E_{21k}-\sum_k E_{22k}$. Further let $T_{AB}=T_{A|1}-T_{A|2}$ be a test for the interaction term. In a synchronized permutation, we select ν observations at random from the n observations in cell E_{11} and exchange them at random with observations from E_{12}. At the same time we select ν observations at random from E_{21} and exchange them at random with elements of E_{22}.

Noting that $\beta_1=-\beta_2$ and $(\alpha\beta)_{11}=-(\alpha\beta)_{12}$, a perturbation of $T_{A|1}$, will be equal to,

$$T^*_{A|1}=2(n-2\nu)\beta_1+2(n-2\nu)(\alpha\beta)_{11}+\sum_k\epsilon^*_{11k}-\sum_k\epsilon^*_{12k},$$

with the $*$ denoting a permutation of the data and ϵ^*_{ijk} denoting the permuted error. Likewise, a perturbation of $T_{A|2}$ is equal to

$$T^*_{A|2}=2(n-2\nu)\beta_1+2(n-2\nu)(\alpha\beta)_{21}+\sum_k\epsilon^*_{21k}-\sum_k\epsilon^*_{22k}.$$

Noting that $(\alpha\beta)_{11}=-(\alpha\beta)_{21}$, the expected value of the test statistic is

$$\mathbb{E}[T^*_{AB}]=4(n-2\nu)(\alpha\beta)_{11}.$$

This test statistic is independent of both main effects and relies only on the exchangeability of the errors.

We also calculate the test statistic T_{BA} where the second factor (i.e., cognitive type) is compared at each of the two levels of the first factor (i.e. the belief elicitation mechanism). Let $T_{B|1}=\sum_k E_{11k}-\sum_k E_{21k}$ and $T_{B|2}=\sum_k E_{12k}-\sum_k E_{22k}$. Then the test statistic $T^*_{BA}=T^*_{B|1}-T^*_{B|2}$ is also independent of both main effects. Since T_{AB} is obtained from synchronized permutations involving the row factor A and T_{BA} is obtained from permutations involving the column factor B, both are jointly and equally informative. It follows that their linear combination $T=T_{AB}+T_{BA}$ is a separate exact test for interaction. Following Basso et al. (2009), we use this linear combination as our main test statistic throughout the paper.

Note that in a balanced design, we can divide our test statistic by the number of observations in each cell without changing the relative value of the original test statistic and the value of the permutations. By doing so, both T_{AB} and T_{BA} are equal to the difference between (i) the difference in mean error in cells E_{11} and E_{12} and (ii) the difference in mean error in cells E_{21} and E_{22}. Thus, as described in the main text, the interaction term is based on the difference between (i) the difference in mean errors between consistent and inconsistent participants in the SBDM mechanism and (ii) the difference in mean errors between consistent and inconsistent participants in the Introspection mechanism.

We follow Basso et al. (2009) and use constrained synchronized permutations in which we exchange the observations in the same location within each cell on each iteration. This is done by permuting the observations in cells E_{11} and E_{12} and then using the same permutation of columns when shuffling observations in cells E_{21} and E_{22}, cells E_{11} and E_{21}, and cells E_{12} and E_{22}. The constrained synchronized permutation ensures that the same number of exchanges is made between each pair of cells. We perform an initial permutation of each cell to ensure that the original position of observations is irrelevant. This ensures that each permutation of the data is equally likely.

Finally, while we have aimed for a balance design, the median split of types was not always exactly 50:50 and our data is not balanced. As discussed in Good (2000), this has the potential of confounding the interaction and main effects. Basso et al. (2009) provides an approach of weighting observations that can be used to conduct synchronized permutations in an unbalanced 2×2 factor design. However, Hahn and Salmaso (2017) shows that these weights also influence the error terms and can lead to a test statistic that is too permissive. The alternative weights proposed by Hahn and Salmaso (2017), which can be used if there is balance in one direction, can be applied only in a subset of our analysis and restrict us to tests using only T_{AB} when it can be applied.

Rather than taking a weighting approach, we instead follow a suggestion in Montgomery (2017) of randomly dropping observations so that each cell has the same number of observations. Although we lose some power by reducing the size of the sample, the resulting data is a random sample of the original and the resulting test statistic is independent of the main effect. To ensure that our random subset of data is not driving our results, we use an outer loop in our testing procedure and perform our permutation test with 1000 sub samples. We report the average p-value over the 1000 samples in the main text. In Table 8 below we also report the percentage of iterations in which the individual p-value corresponds to the acceptance/rejection decision of the average p-value. For example, if the p-value of a test is 0.03 and we reject the null of no interaction, column 2 reports the percentage of sub samples in which the null was rejected.

The test for Hypothesis 2 is similar to the that of Hypothesis 1 with one major exception. In Hypothesis 2, we are comparing behavior of the same individual in informative and uninformative questions and thus the errors of observation E_{i1k} will be correlated with E_{i2k}. This correlation implies that we cannot randomly permute across informative and uninformative questions without changing the expected error distribution. In this case, we restrict attention only to the permutation test T_{BA} where we shuffle the same observations between cells E_{11} and E_{21} and then permute the same columns in cells E_{12} and E_{22}. This permutation keeps pairs of observations together and does not change the underlying error distribution.

As a robustness test, we also analyzed the data using the Wald-Type Permutation Statistic (WTPS) developed by Pauly et al. (2015). This procedure uses a free permutation of the dependent variable and is asymptotically valid in the case of heteroscedasticity in the errors across cells. In our experiment, this may be an issue if inconsistent participants have larger variation in errors. As the test is based on a Wald test, it is more sensitive to outliers. As such, we apply the test to the cleaned version of our dataset that drops outliers according to the criterion in Appendix F. As seen below, the acceptance/rejection decisions of the two tests coincide in all four of the main reported tests.

	Synchronized Permutation Test		**WTPS**
	***p*-value**	**Percentage of Iterations Matching Decision**	***p*-value**
Hypothesis 1: All Signals	0.027	89%	0.032
Hypothesis 1: Uninformative Signals	0.017	98%	0.023
Hypothesis 1: Informative Signals	0.075	74%	0.112
Hypothesis 2: All Signals	0.001	100%	0.001

Table 8: Comparison of p-values from Synchronized Permutation test and alternative Wald-Type Permutation Statistic

Appendix E: Robustness Check: Classification Criterion for Consistent and Inconsistent Participants in the Initial Experiment

In the main text we classified individuals into “consistent” and “inconsistent” types based on their decisions in the last ten periods of Block One of the experiment (Periods 11-20). This selection criterion was used to ensure that individuals were not being classified into type based on early experimentation. However, as this selection criterion could be interpreted as arbitrary, we also explored how variation of this criterion influences our results in the initial experiment where the criterion was not pre-specified.

We reported mean belief errors in the initial experiment in Appendix A in Table 5. Table 9 presents mean belief errors in the initial experiments using an alternative classification in which we do a median split using all 20 periods. As seen by comparing the two tables, mean belief errors are similar across the two classifications.

Concentrating on Table 9, which reports the mean errors for the alternative classification, the mean error for consistent participants in the SBDM mechanism is 10.76 while the mean error for inconsistent participants is 15.98. Thus, there is a -5.22 percentage point difference in means in the SBDM mechanism. The mean error for consistent participants in the Introspection mechanism is 14.57 while the mean error for inconsistent participants is 14.96. Thus there is a -0.39 percentage point difference in means in the Introspection mechanism. The difference-in-difference estimate of -4.83 is significant using the one-sided synchronized test used throughout the paper (p-value = .019). This is comparable to the difference-in-difference estimate of -5.40 that exists when we classify participants based on Periods 11-20, which is the measure we use throughout the paper (p-value = .009).

Belief	**Cognitive**	**Informative Signals**				**All Informative**	**Uninformative**	**All**
Elicitation	**Type**					**Signals**	**Signals**	**Signals**
Method		$\rho' \in \{0.16, 0.84\}$	$\rho' \in \{0.30, 0.70\}$	$\rho' \in \{0.31, 0.69\}$	$\rho' \in \{0.40, 0.60\}$	$\rho' \neq 0.5$	$\rho' = 0.5$	
SBDM	Consistent	8.41	11.25	13.74	9.72	10.57	11.39	10.76
Introspection	Consistent	16.13	18.03	15.43	13.97	16.05	9.31	14.57
- Permutation Test:		(p-value 0.001)	(p-value 0.061)	(p-value 0.672)	(p-value 0.331)	(p-value 0.015)	(p-value 0.516)	(p-value 0.078)
SBDM	Inconsistent	17.39	19.70	19.35	13.18	16.87	12.96	15.98
Introspection	Inconsistent	15.79	17.43	18.85	16.22	16.99	8.45	14.96
- Permutation Test:		(p-value 0.594)	(p-value 0.464)	(p-value 0.880)	(p-value 0.443)	(p-value 0.953)	(p-value 0.147)	(p-value 0.616)
SBDM	Full sample	12.16	14.84	16.80	11.58	13.60	12.15	13.27
Introspection	Full sample	15.97	17.73	17.34	15.22	16.54	8.84	14.78
- Permutation Test:		(p-value 0.053)	(p-value 0.236)	(p-value 0.833)	(p-value 0.203)	(p-value 0.066)	(p-value 0.134)	(p-value 0.306)

Table 9: Data from Initial Experiment Using Alternative Classification with all Observations From Block One: Mean belief errors in the SBDM mechanism and the Introspection mechanism for (i) consistent participants, (ii) inconsistent participants, and (iii) both consistent and inconsistent participants combined. The reported p-values are based on permutation tests using 10,000 iterations in which the subset of participants is held fixed and participants are randomly allocated to the SBDM or Introspection mechanism in each iteration of a regression on the treatment effect. The null hypothesis is that the treatment coefficient is equal to 0 (i.e. that there is no difference in accuracy between the SBDM and Introspection). The two-sided test statistic is reported.

Appendix F: Robustness Check: Results with Outliers Excluded

Due to Covid-19 restrictions our follow-up experiment was conducted online. The resulting data set was noisier than the data generated by the lab-based initial experiment. As part of our robustness check we removed outliers to ensure that these were not affecting results. We found that statistical tests on the reduced dataset led to statistically stronger conclusions that the results reported in the body of this paper.

When classifying participants as outliers, we began by counting the number of times that an individual reported a belief that was (i) less than or equal to 50 and (ii) greater than or equal to 50 in the final 10 periods of Blocks 2 and 3. Next, we classified an individual as an outlier if (i) either of the two counts were 19 or 20 and (ii) less than 50% of belief reports were exactly 50. This leads to the exclusion (for example) of participants who report a single number like 20 or 100 throughout the experiment, or whose probabilities are reported out of 40—the number of balls in the bucket—rather than 100.

This rule leads to the exclusion of 9 participants from the initial experiment (4 from the Introspection treatment; 5 from the SBDM treatment), and 16 from the follow-up experiment (5 from the Introspection treatment; 11 from the SBDM treatment).

Table 10 reports the mean errors from the pooled data when outliers are excluded. The difference-in-difference estimate for Hypothesis 1 is -3.23, which is statistically significant in a one-sided test using the estimator described in Appendix D (p-value = .009). Tables 11 and 12 report the mean errors for the initial experiments and follow-up experiments separately when the outliers are excluded.

Belief	**Cognitive**	**Informative Signals**				**All Informative**	**Uninformative**	**All**
Elicitation	**Type**					**Signals**	**Signals**	**Signals**
Method		$\rho' \in \{0.16, 0.84\}$	$\rho' \in \{0.30, 0.70\}$	$\rho' \in \{0.31, 0.69\}$	$\rho' \in \{0.40, 0.60\}$	$\rho' \neq 0.5$	$\rho' = 0.5$	
SBDM	Consistent	10.14	9.26	14.11	10.31	10.48	8.03	9.91
Introspection	Consistent	15.37	15.75	13.73	11.47	14.06	6.20	12.31
- Permutation Test:		(p-value: 0.001)	(p-value: 0.003)	(p-value: 0.843)	(p-value: 0.598)	(p-value: 0.004)	(p-value: 0.229)	(p-value: 0.033)
SBDM	Inconsistent	17.38	17.02	17.05	15.32	16.44	13.75	15.85
Introspection	Inconsistent	16.87	18.46	19.78	15.58	17.42	6.85	15.02
- Permutation Test:		(p-value: 0.837)	(p-value: 0.514)	(p-value: 0.156)	(p-value: 0.914)	(p-value: 0.496)	(p-value: 0.000)	(p-value: 0.538)
SBDM	Full sample	13.01	12.24	15.66	12.87	13.18	10.52	12.58
Introspection	Full sample	15.99	16.89	16.69	13.48	15.59	6.50	13.54
- Permutation Test:		(p-value: 0.035)	(p-value: 0.003)	(p-value: 0.459)	(p-value: 0.718)	(p-value: 0.013)	(p-value: 0.001)	(p-value: 0.280)

Table 10: Data with both experiments with outliers removed: mean belief errors under the SBDM mechanism and the Introspection mechanism for (i) consistent participants, (ii) inconsistent participants, and (iii) both consistent and inconsistent participants combined. The reported p-values are based on permutation tests using 10,000 iterations in which the subset of participants is held fixed and participants are randomly allocated to the SBDM or Introspection mechanism in each iteration of a regression on the treatment effect. The null hypothesis is that the treatment coefficient is equal to 0 (i.e. that there is no difference in accuracy between the SBDM and Introspection). The two-sided test statistic is reported.

Belief Elicitation Method	**Cognitive Type**	**Informative Signals**				**All Informative Signals**	**Uninformative Signals**	**All Signals**
		$\rho' \in \{0.16, 0.84\}$	$\rho' \in \{0.30, 0.70\}$	$\rho' \in \{0.31, 0.69\}$	$\rho' \in \{0.40, 0.60\}$	$\rho' \neq 0.5$	$\rho' = 0.5$	
SBDM	High	8.23	10.67	14.49	9.10	10.34	9.09	10.06
Introspection	High	15.34	17.02	13.49	13.78	15.24	8.82	13.82
- Permutation Test:		(p-value: 0.004)	(p-value: 0.094)	(p-value: 0.775)	(p-value: 0.205)	(p-value: 0.020)	(p-value: 0.926)	(p-value 0.061)
SBDM	Low	16.69	18.46	18.74	14.38	16.93	14.57	16.41
Introspection	Low	16.97	17.84	20.84	14.56	17.09	7.17	14.72
- Permutation Test:		(p-value: 0.932)	(p-value: 0.846)	(p-value: 0.564)	(p-value 0.971)	(p-value: 0.949)	(p-value: 0.028)	(p-value: 0.449)
SBDM	Full sample	12.42	14.67	16.45	11.40	13.47	11.66	13.07
Introspection	Full sample	15.97	17.36	17.05	14.16	16.06	8.05	14.22
- Permutation Test:		(p-value: 0.079)	(p-value: 0.287)	(p-value: 0.817)	(p-value 0.358)	(p-value: 0.113)	(p-value: 0.096)	(p-value: 0.456)

Table 11: Data from initial experiment with outliers removed: Mean belief errors under the SBDM mechanism and the Introspection mechanism for (i) consistent participants, (ii) inconsistent participants, and (iii) all participants combined. The reported p-values are based on permutation tests using 10,000 iterations in which the subset of participants is held fixed and participants are randomly allocated to the SBDM or Introspection mechanism in each iteration of a regression on the treatment effect. The null hypothesis is that the treatment coefficient is equal to 0 (i.e. that there is no difference in belief error between the SBDM and Introspection). The two-sided test statistic is reported.

Belief	**Cognitive**	**Informative Signals**				**All Informative**	**Uninformative**	**All**
Elicitation	**Type**					**Signals**	**Signals**	**Signals**
Method		$\rho' \in \{0.16, 0.84\}$	$\rho' \in \{0.30, 0.70\}$	$\rho' \in \{0.31, 0.69\}$	$\rho' \in \{0.40, 0.60\}$	$\rho' \neq 0.5$	$\rho' = 0.5$	
SBDM	High	11.21	8.59	13.81	11.38	10.57	7.40	9.81
Introspection	High	15.38	14.84	13.86	10.13	13.29	4.51	11.32
- Permutation Test:		(p-value: 0.045)	(p-value: 0.016)	(p-value: 0.983)	(p-value: 0.645)	(p-value: 0.081)	(p-value: 0.091)	(p-value: 0.255)
SBDM	Low	18.18	15.31	16.15	15.81	16.08	13.14	15.43
Introspection	Low	16.81	18.87	19.21	16.16	17.62	6.63	15.20
- Permutation Test:		(p-value: 0.721)	(p-value: 0.257)	(p-value: 0.176)	(p-value: 0.903)	(p-value: 0.405)	(p-value: 0.006)	(p-value: 0.893)
SBDM	Full sample	13.46	10.58	15.14	13.87	12.99	9.77	12.24
Introspection	Full sample	16.01	16.57	16.50	13.10	15.29	5.48	13.11
- Permutation Test:		(p-value: 0.198)	(p-value: 0.004)	(p-value: 0.419)	(p-value: 0.704)	(p-value: 0.064)	(p-value: 0.002)	(p-value: 0.432)

Table 12: Data from follow-up experiment with outliers removed: Mean belief errors under the SBDM mechanism and the Introspection mechanism for (i) consistent participants, (ii) inconsistent participants, and (iii) all participants combined. The reported p-values are based on permutation tests using 10,000 iterations in which the subset of participants is held fixed and participants are randomly allocated to the SBDM or Introspection mechanism in each iteration of a regression on the treatment effect. The null hypothesis is that the treatment coefficient is equal to 0 (i.e. that there is no difference in belief error between the SBDM and Introspection). The two-sided test statistic is reported.

Appendix G: Instructions

Instructions and Quizzes

The experiment included 3 blocks of 20 periods, which were referred to in the instructions as Experiments 1, 2 and 3. Statements in parentheses and italics provide additional details or discuss differences between the treatments and do not form part of the experiment instructions.

Experiment One

Thank you for choosing to participate today. We appreciate your time. This experiment is an opportunity to earn money. You will be paid in cash at the end of the experiment. You will be paid a $10 attendance fee. You will also receive payments based on the outcome of three experiments. You will not learn your total payoff until the end of the experiment.

There is a very short, anonymous questionnaire at the end of the experiment. You will be paid when the questionnaire is completed.

If you have any questions during the experiment, please sit quietly and raise your hand. An experiment assistant will be with you as soon as possible.

Payment for the first experiment: You will play the first experiment 20 times. Each repetition is called a "period." In each period you will get a payoff of $0, $4, or $8. At the end of the experiment, 1 of the 20 periods will be chosen randomly by the computer. Each period is equally likely to be chosen. Your cash payment for the first experiment will be your payoff in the randomly chosen period.

*(In bold text:)*Although you will play 20 periods in the first experiment, you are only paid in cash for the payoff you earn in a single period.

You are going to participate in a decision-making task, which is referred to as the "Choose-A-Side Game." There are two buckets: Bucket A and Bucket B. Each bucket contains 40 balls. Each bucket is divided in half, with 20 balls in each side.

There is a 50-in-100 chance (50% chance) that you have been given Bucket A. The left side of Bucket A contains 12 black balls and 8 white balls. The right side of Bucket A contains 20 black balls and 0 white balls.

(Stylized illustration of Bucket A: a rectangle divided vertically in two, with black or white dots to illustrate the ratio of black and white balls in each half of the bucket.)

There is a 50-in-100 chance (50% chance) that you have been given Bucket B. The left side of Bucket A contains 8 black balls and 12 white balls. The right side of Bucket A contains 0 black balls and 20 white balls. (The buckets and balls are all computerized.)

(*Stylized illustration of Bucket B: a rectangle divided vertically in two, with black or white dots to illustrate the ratio of black and white balls in each half of the bucket.*)

One of the buckets will be randomly chosen by the computer. Both buckets have an equal chance of being chosen. This means that both buckets have a 50-in-100 chance of being chosen (50%). (You might imagine that the computer tosses a coin to decide which bucket will be used.) You will not be told which bucket has been chosen by the computer. The computer will randomly select a ball from the left hand side of your bucket. Each ball has an equal chance of being chosen. You will be told the colour of the ball. After you see the ball, it is put back in the left hand side of your bucket. If the ball is black, you receive $4. If it is white, you receive $0 (nothing). This is your Stage-1 payoff.

You then have a second chance to draw a ball from your bucket. As before, black balls are worth $4. White balls are worth $0 (nothing). You must decide whether you would like the computer to draw the ball from the left hand side of your bucket, or the right hand side. The computer randomly selects a ball from the side you choose. If it is black, you receive $4. If it is white, you receive $0 (nothing). This is your Stage-2 payoff.

Your payoff for the period is your Stage-1 payoff plus your Stage-2 payoff. In total you might have a payoff of $0, $4, or $8 across both stages of the Choose-A-Side Game.

In each period there is a 50-in-100 (50%) chance of being given Bucket A or Bucket B. Your bucket is randomly determined by the computer and is not affected by the bucket you have been given in previous periods.

Summary: Choose-A-Side Game

There are 2 buckets, Bucket A and Bucket B. Each bucket has a 50-in-100 chance (50%) of being chosen. Each bucket is divided in half, with 20 balls in each half. The computer randomly selects a bucket for you. You do not know which bucket you have been given. You will see a randomly chosen ball from the left-hand side of your bucket. If it is black, your Stage-1 payoff is $4. If it is white, your payoff is $0 (nothing). You then choose whether you want a second ball drawn from the left or right side of your bucket. The computer draws a ball from your chosen side. If it is black, your Stage-2 payoff is $4. If it is white, your payoff is $0 (nothing). Your period payoff is your Stage 1 plus your Stage 2 payoff. In each period you might get a payoff of $0 (nothing), $4, or $8 in the Choose-A-Side Game. 1 of the 20 periods will be randomly chosen. Each period has an equal (1-in-20) chance of being chosen. Your will be paid your earnings from that period in cash at the end of the experiment.

When you have finished Experiment 1 you will be given instructions for a second experiment.

Quiz

At the start of a period the computer randomly selects a bucket for you.

1. What is the chance-in-100 that you get Bucket A? (50)

2. What is the chance-in-100 that you get Bucket B? (50)

The bucket has 20 balls in each side; 40 in total. The computer shows you a ball from the left-hand side of your bucket, tells you its colour, and tells you whether it is worth \$4 or \$0. This is your payoff for Stage 1. The computer puts the ball back in your bucket. The computer asks whether you want the next ball drawn from the left-hand side or the right-hand side of the bucket.

3. How many balls are there in the left-hand side? (20)

4. How many balls are there in the right-hand side? (20)

The computer draws a ball from the side you choose, tells you its colour, and tells you if you have won \$4 or \$0. This is your payoff for Stage 2.

5. What is the minimum payoff possible in a period (Stage 1 + Stage 2)? (0)

6. What is the maximum payoff possible in a period (Stage 1 + Stage 2)? (8)

You then finish the period.

7. How many periods are there in this experiment? (20)

8. Do you receive a cash payment for your payoff in every period? (No)

9. Every period has a 1-in-? chance of being paid? (20)

10. When the next period starts, what is the chance-in-100 that you get Bucket A? (50)

11. What is the chance-in-100 that you get Bucket B? (50)

(Experiment begins when all questions are answered correctly. At the end of Experiment One:)

Thank you! You have now played 20 periods and finished the first experiment. At the end of the third experiment you will find out which period was randomly chosen. You will be paid your payoff from the randomly chosen period. You will be pain in cash. You will now read instructions for the second experiment.

Experiment Two

You will play the second experiment 20 times. Each repetition is called a ‘period.’ In each period you get a payoff of \$0, \$4, or \$8. At the end of the second experiment, 1 of the 20 periods will be chosen randomly by the computer. Each period is equally likely to be chosen. Your cash payment for the second experiment will be your payoff in the randomly chosen period. Your total payment today will include:

- Your show-up fee of \$10
- A cash payment for a randomly chosen period from the first experiment

- A cash payment for a randomly chosen period from the second experiment
- A cash payment for a randomly chosen period from the third experiment

Although you will play 20 periods in this second experiment, you only receive cash for your payoff from a single period.

The set-up for Experiment 2 is the same as Experiment 1. There are two buckets: Bucket A and Bucket B. Each bucket contains 40 balls. Each bucket is divided in half, with 20 balls in each side.

There is a 50-in-100 chance (50% chance) that you have been given Bucket A. The left side of Bucket A contains 12 black balls and 8 white balls. The right side of Bucket A contains 20 black balls and 0 white balls.

(*Stylized illustration of Bucket A: a rectangle divided vertically in two, with black or white dots to illustrate the ratio of black and white balls in each half of the bucket.*)

There is a 50-in-100 chance (50% chance) that you have been given Bucket B. The left side of Bucket A contains 8 black balls and 12 white balls. The right side of Bucket A contains 0 black balls and 20 white balls. (The buckets and balls are all computerized.)

(*Stylized illustration of Bucket B: a rectangle divided vertically in two, with black or white dots to illustrate the ratio of black and white balls in each half of the bucket.*)

One of the buckets will be randomly chosen by the computer. Both buckets have an equal (50-in-100) chance of being chosen. You will not be told which bucket has been chosen by the computer. The computer will randomly select a ball from the left hand side of your bucket. Each ball has an equal chance of being chosen. You will be told the colour of the ball. After you see the ball, it is put back in the left hand side of your bucket. If the ball is black, you receive $4. If it is white, you receive $0 (nothing). This is your Stage-1 payoff.

(The three treatments involve different instructions from this point.)

SBDM mechanism treatment

After seeing the colour of the ball, you need to think about the chance that the ball was drawn from Bucket A. This is your "belief" that the ball was drawn from Bucket A. Your "belief" is a number between 0 and 100, to indicate the chance-in-100 that the ball has been drawn from Bucket A.

For example: If you are sure that Bucket A is being used, your belief is that there is a 100-in-100 chance that Bucket A is being used. If you are sure that Bucket A is not being used, your belief is that there is a 0-in-100 chance that Bucket A is being used. If you believe that it is equally likely that Bucket A is being used as Bucket B, then your belief is that there is a 50-in-100 chance that Bucket A is being used. (These are just examples. You can enter any chance-in-100 belief between 0 and 100.)

You then answer 2 questions. Question 1: What is your belief that the ball was drawn from Bucket A? Question 2: Do you want the computer to draw a second ball from the left or right-hand side of your bucket?

The computer then tosses a coin to determine which question is used to determine your Stage-2 payoff. Tails: Question 1. Heads: Question 2. If the computer throws a Heads, your Stage-2 payoff will be determined the same way as Experiment 1 (the Choose-A-Side Game). The computer will draw a ball from the side of the bucket you choose. As before, black balls are worth $4. White balls are worth $0 (nothing).

We will now explain how Stage-2 payoffs are determined if the computer throws “Tails.”

In Question 1 you tell the computer your belief (the chance-in-100) that the first ball was drawn from Bucket A. If the computer throws “Tails”, this is how we determine your Stage-2 payoff:

The computer creates a Lottery Bag. The computer randomly chooses a number between 0 and 100. Each number is equally likely to be chosen. Although the computer knows this number, you do not. We call this randomly chosen number “?”. The computer fills a bag with 100 chips. ”?” chips are black, and the rest are white. ?-in-100 chips are black. There are now two ways to get a payoff of $4: The ‘Belief about Bucket A Game,” and the Lottery Bag Game.

(Table/illustration here.)

THE BELIEF ABOUT BUCKET A GAME:
Prize of $4 if the ball was from Bucket A.
Prize of $0 if the ball was from Bucket B.

Chance-in-100 of winning $4:

Belief (chance-in-100) that ball is from Bucket A

THE LOTTERY BAG GAME:
Prize of $4 if you draw a black chip.
Prize of $0 if you draw a white chip.

Chance-in-100 of winning $4:

“?”-in-100

The computer knows the chance of winning $4 in the Lottery Bag Game. Based on your reported belief that the ball was drawn from Bucket A, the computer will select the game that gives you the highest chance of winning $4. (If the games give you an equal chance of winning, you will play the Lottery Bag Game.)

You should think carefully about your belief that the ball has been drawn from Bucket A, as the computer will use your reported belief to decide whether you are paid according to your “Belief about Bucket A” or the ”Lottery Bag” Game. This experiment might feel very detailed and complicated, but it is set up this way so that it is in your best interest to report your beliefs honestly and carefully. If you make a report that is not your true belief, your payoff might be determined by the Lottery Bag Game when you would prefer to be paid based on your belief that the ball was drawn from Bucket A (or vice-versa).

The best thing you can do is report your belief honestly, so that you are given the game with the highest chance of a payoff of $4.

Summary: Experiment 2

You have a 50-in-100 chance of being given Bucket A or Bucket B in each period. You will be shown a ball from the left-hand side of your bucket. You will answer 2 questions. Question 1: What is your belief that the ball was drawn from Bucket A? Question 2: Do you want the computer to draw a second ball from the left or right side of your bucket? The computer then tosses a coin to determine which question is used to determine your Stage-2 payoff. Tails: Question 1. Heads: Question 2.

If the computer throws "Heads" your payoff is determined in the same way as Experiment 1. A second ball will be drawn from your bucket, from the side you choose. If the computer throws "Tails" your payoff will be determined by the "Belief about Bucket A" Game or a Lottery Bag Game. The best thing you can do is report your belief honestly, so that you are given the game with the highest chance of a payoff of $4.

Your period payoff is your Stage-1 payoff plus your Stage-2 payoff. In each period you might get a payoff of $0, $4, or $8. 1 of the 20 periods will be randomly chosen. Each period has an equal (1-in-20) chance of being chosen. Your will be paid your earnings from that period in cash at the end of the experiment.

Quiz

Imagine that you are shown a ball. Based on its colour, you report your belief that there is a 20-in-100 chance that the ball is from Bucket A. The computer flips a coin and it lands on "Tails." The computer creates a Lottery Bag game, and randomly includes 25 black chips. It has a 25-in-100 chance of winning $4. Based on your report, the computer chooses the game that gives you a higher chance of winning $4.

1. Which game will be used to determine your payoff for the period? (Lottery Bag Game)

2. What is your chance-in-100 of winning $4? (25)

3. What is your chance-in-100 of winning $0? (75)

Imagine you start a new period. You are shown a new ball. This time, you believe there is an 81-in-100 chance that the ball was taken from Bucket A... but you make an error! You type 18 by mistake. This is your reported belief.

The computer doesn't know your belief, only your reported belief. The computer thinks you believe there is an 18-in-100 chance of winning $4 in the Belief-About-Bucket-a Game.

The computer creates a Lottery Bag Game and randomly includes 36 black chips. It has a 36-in-100 chance of winning $4.

4. What do you believe is your chance-in-100 of winning $4 if you play the Belief-About-Bucket-A Game? (81)
5. What does the computer think you believe is the chance-in-100 of winning $4 if you play the Belief-About-Bucket-A Game? (18)
6. What is your chance-in-100 of winning $4 if you play the Lottery Bag Game? (36)

Based on your report, the computer chooses the game that it thinks will give you a higher chance of winning $4.

7. Which game will be used to determine your prize for the period? (Lottery Bag Game)

(Experiment begins when all questions are answered correctly.)

Unpaid Introspection Treatment

After seeing the colour of the ball, you need to think about the chance that the ball was drawn from Bucket A. This is your "belief" that the ball was drawn from Bucket A. Your "belief" is a number between 0 and 100 to indicate the chance in 100 that the ball has been drawn from Bucket A.You should think carefully about your belief that the ball has been drawn from Bucket A.

For example: If you are sure that Bucket A is being used, your belief is that there is a 100-in-100 chance that Bucket A is being used. If you are sure that Bucket A is not being used, your belief is that there is a 0-in-100 chance that Bucket A is being used. If you believe that it is equally likely that Bucket A is being used as Bucket B, then your belief is that there is a 50-in-100 chance that Bucket A is being used. (These are just examples. You can enter any chance-in-100 belief between 0 and 100.)

You then answer 2 questions.

Question 1: What is your belief that the ball was drawn from Bucket A

Question 2: Do you want the computer to draw a second ball from the left or right side of your bucket?

The computer randomly selects a ball from the side you choose. If it is black, you receive $4. If it is white, you receive $0 (nothing). This is your Stage-2 payoff. Your payoff for the period is your Stage-1 payoff plus your Stage-2 payoff. In total you might have a payoff of $0 (nothing), $4 or $8 across both stages of the Choose-A-Side Game. In each period there is a 50-in-100 (50%) chance of being given Bucket A or Bucket B.Your bucket is randomly determined by the computer and is not affected by the bucket you have been given in previous periods.

Summary: Experiment 2 You have a 50-in-100 (50%) chance of being given Bucket A or Bucket B. Each bucket is divided in half, with 20 balls in each half. You will be shown a ball from the left hand side of your bucket. If it is black, your Stage-1 payoff is $4. If it is white, your payoff is $0 (nothing). You will answer 2 questions:

Question 1: What is your belief that the ball was drawn from Bucket A?

Question 2: Do you want the computer to draw a second ball from the left or right side of your bucket?

You should think carefully about your belief that the ball has been drawn from Bucket A. If it is black, your Stage-2 payoff is $4. If it is white, your payoff is $0 (nothing). Your period payoff is your Stage 1 payoff plus your Stage 2 payoff. In each period you might get a payoff of $0 (nothing), $4 or $8. 1 of the 20 periods will be randomly chosen. Each period has an equal (1-in-20) chance of being chosen. You will be paid your payoff from that period in cash at the end of the experiment. When you have finished Experiment 2 you will be given instructions for a third experiment.

Quiz

At the start of a period the computer randomly selects a bucket for you. The bucket has 20 balls in each side: 40 in total. The computer shows you a ball from the left-hand side of the bucket, tells you its colour, and tells you whether it is worth $0 or $4. This is your payoff for Stage 1. The computer puts the ball back in your bucket. The computer asks whether you want the next ball drawn from the left hand side or the right hand side of the bucket. The computer also asks your belief about the chance-in-100 that the ball was drawn from Bucket A.

Imagine that you're sure the ball is drawn from Bucket A. How would you report this as a chance-in-100 that the ball is drawn from Bucket A?

1. The chance-in-100 of the Ball being from Bucket A is: (100)

Imagine that you're sure the ball is not drawn from Bucket A. How would you report this as a chance-in-100 that the ball is drawn from Bucket A?

2. The chance-in-100 of the Ball being from Bucket A is: (0)

Imagine that you think there's an equal chance the ball is drawn from Bucket A. How would you report this as a chance-in-100 that the ball is drawn from Bucket A?

3. The chance-in-100 of the Ball being from Bucket A is: (50)

The computer draws a ball from the side you choose, tells you its colour, and tells you if you have won $0 or $4. This is your payoff for Stage 2.

4. What is the minimum payoff possible in a period (Stage 1 + 2)? (0)

5. What is the maximum payoff possible in a period (Stage 1 + 2)? (8)

You then finish the period.

6. How many periods are there in this experiment? (20)
7. Do you receive a cash payments for your payoff in every period? (No)
8. Every period has a 1-in-? chance of being paid? 1-in: (20)

(Experiment begins when all questions are answered correctly.)

No-Elicitation Treatment

You must then decide whether you would like the computer to draw a second ball from the left-hand side of your bucket, or the right-hand side.

The computer randomly selects a ball from the side you choose. If it is black, you receive \$4. If it is white, you receive \$0 (nothing). This is your Stage-2 payoff. Your payoff for the period is your Stage-1 payoff plus your Stage-2 payoff. In total you might have a payoff of \$0 (nothing), \$4, or \$8 across both stages of the Choose-A-Side Game. In each period there is a 50-in-100 (50%) chance of being given Bucket A or Bucket B. Your bucket is randomly chosen by the computer and is not affected by the bucket you have been given in previous rounds.

Summary: Experiment 2 You have a 50-in-100 (50%) chance of being given Bucket A or Bucket B. Each bucket is divided in half, with 20 balls in each half. You will be shown a ball from the left-hand side of your bucket. If it is black, your Stage-1 payoff is \$4. It if is white, your payoff is \$0 (nothing). You must then decide whether you would like the computer to draw a second ball from the left-hand side of your bucket, or the right-hand side. A second ball will be drawn from your bucket, from the side you choose. If it is black, your Stage-2 payoff is \$4. If it is white, your payoff is \$0 (nothing). Your period payoff is your Stage 1 playoff plus your Stage 2 payoff. In each period you might get a payoff of \$0 (nothing), \$4, or \$8 in the Choose-A-Side Game. 1 of the 20 periods will be randomly chosen. Each period has an equal (1-in-20) chance of being chosen. You will be paid your payoff from that period in cash at the end of the experiment. When you have finished Experiment 2 you will be given instructions for a third experiment.

Quiz

(The quiz for the No-Elicitation Treatment is the same as the quiz for Experiment 1—that is, the "Choose-A-Side" Experiment.)

(Experiment begins when all questions are answered correctly.)

Experiment Three

You are about to start Experiment 3. This is the final experiment. You will repeat the experiment 20 times. Each repetition is called a "period." In each period you get a payoff of \$0, \$4, \$8 or \$12. At the end of the third experiment, 1 of the 20 periods will be chosen randomly by the computer. Each period is equally likely to be chosen. Your cash payment for the third experiment will be your payoff in the randomly chosen period. Your total payment today will include:

- Your show-up-fee of \$10.
- A cash payment for a randomly chosen period from the first computerized experiment
- A cash payment for a randomly chosen period from the second computerized experiment
- A cash payment for a randomly chosen period from the third computerized experiment

Although you will play 20 periods in this third experiment, you are paid cash for your payoff in a single period. Experiment 3 is the same as Experiment 2, except you will see **two** balls drawn from your bucket

The computer will randomly select a ball from the left-hand side of your bucket. The computer will tell you the colour of the ball, and whether your payoff is \$0 or \$4. The computer will put the ball back in the left-hand side of your bucket. The computer will randomly select a **second** ball from the left-hand side of your bucket. Because the ball is randomly chosen, this might be the same ball (chosen a second time) or it might be a new ball. The computer will tell you the colour of the second ball, and whether your payoff is \$0 or \$4.

You have two chances to get a payoff of \$4 in Stage 1. This means you can secure a payoff of \$0, \$4 or \$8 in Stage 1.

(The three treatments involve different instructions from this point.)

SBDM Treatment

You then answer 2 questions:

- Question 1: What is your belief that the two balls were drawn from Bucket A?
- Question 2: Do you want the computer to draw a third ball from the left or right side of your bucket?

The computer then tosses a coin to determine which Question is used to determine your Stage-2 payment.

- Tails: Question 1

- Heads: Question 2.

Just like Experiment 2: If you throw a Heads, your Stage-2 payoff will be determined by the “Choose-A-Side Game.” The computer will draw a ball from the side of the bucket you choose. You will get a payoff of $4 if the ball is black, and $0 (nothing) if it is white. If you throw a Tails, your Stage-2 payoff will be determined by the Lottery Bag Game, or the Belief-About-Bucket-A Game (whether the two balls were drawn from Bucket A). These games are played in exactly the same way as in Experiment 2. Based on your reported belief that the ball was drawn from Bucket A, the computer will select the game that gives you the highest chance of winning $4.

You should think carefully about your belief that the balls have been drawn from Bucket A, as the computer will use your reported belief to decide whether you are paid according to your “Belief-About-Bucket A” or the Lottery game. The experiment is set up so that it is in your best interest to report your belief honestly and carefully. If you make a report that is not your true belief your payoff might be determined by the Lottery Game when you would prefer to be paid based on your belief that the ball was drawn from Bucket A.

Summary: Experiment 3

You have a 50-in-100 (50%) chance of being given Bucket A or Bucket B. You will be shown **2 balls** from the left hand side of your bucket. You will answer 2 questions:

- Question 1: What is your belief that the 2 balls were drawn from Bucket A?
- Question 2: Do you want the computer to draw a third ball from the left or right side of your bucket?

The computer then tosses a coin to determine which question is used to determine your Stage-2 payment:

- Tails: Question 1
- Heads: Question 2

If the computer throws “Heads” your payoff is determined in the same way as Experiment 1. A second ball will be drawn from your bucket, based on the side you choose. If the computer throws “Tails” your payoff will be determined by the “Belief about Bucket A” game or a Lottery Game. The best thing you can do is report your belief honestly, so that you are given the game with the highest chance of a payoff of $4. Your period payoff is your Stage 1 payoff plus your Stage 2 payoff. In each period you might get a payoff of $0 (nothing), $4, $8 or $12 in the third experiment. 1 of the 20 periods will be randomly chosen. Each period has an equal (1-in-20) chance of being chosen. You will be paid your payoff from that period in cash at the end of the experiment.

Quiz

At the start of a period the computer randomly selects a bucket for you. Both buckets are equally likely to be chosen. The bucket has 20 balls in each side: 40 in total. The computer shows you a ball from the left side of your bucket, tells you its colour, and tells you whether your payoff is \$4 or \$0. The computer puts the ball back in the left-hand side of your bucket.

1. How many balls are there in the left hand side? (20)

The computer draws a second ball from the left-hand side of your bucket and tells you whether your second payoff is \$4 or \$0.

2. Is it possible that the computer drew the same ball twice? (Yes)

The computer puts the second ball back in your bucket.

3. How many balls are there in the left hand side?(20)

The computer asks whether you want the third ball drawn from the left hand side or the right hand side of the bucket. The computer also asks you to report your belief that the ball is from Bucket A.

4. What is the minimum payoff possible in a period (both balls from Stage 1 + ball from Stage 2)? (0)
5. What is the maximum payoff possible in a period (both balls from Stage 1 + ball from Stage 2)? (12)

(Experiment begins when all questions are answered correctly.)

Unpaid Introspection Treatment

After seeing the colour of the **two balls**, you need to think about the chance that the balls were drawn from Bucket A. This is your “belief” that the balls were drawn from Bucket A. Your “belief” is a number between 0 and 100 to indicate the chance-in-100 that the ball has been drawn from Bucket A. You should think carefully about your belief that the ball has been drawn from Bucket A. You then answer 2 questions:

- Question 1: What is your belief that the ball was drawn from Bucket A?
- Question 2: Do you want the computer to draw a second ball from the left or right side of your bucket?

The computer randomly selects a ball from the side you choose. If it is black, you receive \$4. If it is white, you receive \$0 (nothing). This is your Stage-2 payoff. Your payoff for the period is your Stage-1 payoff plus your Stage-2 payoff. In total you might have a payoff of $0, $4, $8 or $12 across both stages of the third experiment. In each period there is a 50-in-100 (50%) chance of being given Bucket A or Bucket B. Your bucket is randomly determined by the computer and is not affected by the bucket you have been given in previous periods

Summary: Experiment 3

You have a 50-in-100 (50%) chance of being given Bucket A or Bucket B. You will be shown **2 balls** from the left hand side of your bucket. You will answer 2 questions:

- Question 1: What is your belief that the 2 balls were drawn from Bucket A?
- Question 2: Do you want the computer to draw a third ball from the left or right side of your bucket?

You should think carefully about your belief that the ball has been drawn from Bucket A. A second ball will be drawn from your bucket, from the side you choose. If it is black, your Stage-2 payoff is $4. If it is white, your payoff is $0 (nothing). Your period payoff is your Stage 1 payoff plus your Stage 2 payoff. In each period you might get a payoff of $0 (nothing), $4, $8 or $8 in the third experiment. 1 of the 20 periods will be randomly chosen. Each period has an equal (1-in-20) chance of being chosen. You will be paid your payoff from that period in cash at the end of the experiment.

Quiz

At the start of a period the computer randomly selects a bucket for you. Both buckets are equally likely to be chosen. The bucket has 20 balls in each side: 40 in total. The computer shows you a ball from the left side of your bucket, tells you its colour, and tells you whether your payoff is $4 or $0. The computer puts the ball back in the left-hand side of your bucket.

1. How many balls are there in the left hand side? (2)

The computer draws a second ball from the left-hand side of your bucket and tells you whether your second payoff is $4 or $0.

2. Is it possible that the computer drew the same ball twice? (Yes)

The computer puts the second ball back in your bucket

3. How many balls are there in the left hand side? (20)

The computer asks whether you want the third ball drawn from the left hand side or the right hand side of the bucket. The computer also asks you to report your belief that the ball is from Bucket A.

4. What is the minimum payoff possible in a period (both balls from Stage 1 + ball from Stage 2)? (0)

5. What is the maximum payoff possible in a period (both balls from Stage 1 + ball from Stage 2)? (12)

(Experiment begins when all questions are answered correctly.)

No-elicitation Treatment

You then have a third chance to draw a ball from your bucket. As before, black balls are worth $4. White balls are worth $0 (nothing). You must decide whether you would like the computer to draw the ball from the left hand side of your bucket, or the right hand side. The computer randomly selects a ball from the side you choose.

If it is black, you receive $4. If it is white, you receive $0 (nothing). This is your Stage-2 payoff.

Your payoff for the period is your Stage-1 payoff plus your Stage-2 payoff. In total you might have a payoff of $0, $4, $8, or $12 across both stages of the third experiment.

In each period there is a 50-in-100 (50%) chance of being given Bucket A or Bucket B. Your bucket is randomly determined by the computer and is not affected by the bucket you have been given in previous periods.

Summary: Experiment 3

You have a 50-in-100 (50%) chance of being given Bucket A or Bucket B. You will be shown **2 balls** from the left hand side of your bucket You will answer a question: Question: Do you want the computer to draw a third ball from the left or right side of your bucket?

A third ball will be drawn from your bucket, from the side you choose. If it is black, your Stage-2 payoff is $4. If it is white, your payoff is $0 (nothing).Your period payoff is your Stage 1 payoff plus your Stage 2 payoff. In each period you might get a payoff of $0 (nothing), $4, $8, or $12 in the third experiment. 1 of the 20 periods will be randomly chosen. Each period has an equal (1-in-20) chance of being chosen. You will be paid your payoff from that period in cash at the end of the experiment

Quiz

At the start of a period the computer randomly selects a bucket for you. Both buckets are equally likely to be chosen. The bucket has 20 balls in each side: 40 in total. The computer shows you a ball from the left side of your bucket, tells you its colour, and tells you whether your payoff is $4 or $0. The computer puts the ball back in the left-hand side of your bucket.

1. How many balls are there in the left hand side? (20)

The computer draws a second ball from the left-hand side of your bucket and tells you whether your payoff is $4 or $0.

2. Is it possible that the computer drew the same ball twice? (Yes)

The computer puts the second ball back in your bucket.

3. How many balls are there in the left hand side? (20)

The computer asks whether you want the third ball drawn from the left hand side or the right hand side of the bucket.

4. What is the minimum payoff possible in a period (both balls from Stage 1 + ball from Stage 2)? (0)
5. What is the maximum payoff possible in a period (both balls from Stage 1 + ball from Stage 2)? (12)

(Experiment begins when all questions are answered correctly.)

Appendix H: Cognitive Response Questionnaire

Below is the Cognitive Reflection Test we used in the order in which questions appeared. Questions 2, 7, and 10 are from Frederick (2005). Questions 5, 6, and 9 are from Primi et al. (2016), which also uses Questions 2, 7, and 10. Placebo questions are from Thomson and Oppenheimer (2016). A participants' score was based on the number of questions correctly answered from Questions 2, 5, 6, 7, 9, and 10.

1. Sara, Emma, and Sophia embark on a river trip. Each of them brings one supply item for the trip: a kayak, a box of sandwiches, and a bag of apples. Sarah brought the apples and Emma didn't bring anything edible. What did Sophia bring? [Placebo A; correct answer = a box of sandwiches]
2. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? [correct answer = 47 days; heuristic answer = 24 days]
3. A mechanic shop had five silver cars, and one blue car in the garage. During the day, three silver cars and one blue car were picked up, and one black car was dropped off. How many silver cars are left at the end of the day? [Placebo B; correct answer = 2]
4. An expedition on a mountain climbing trip will be traveling with eleven horse packs. A horse can carry one, two, or three packs. What is the minimum number of horses that the expedition needs? [Placebo C; correct answer = 4]
5. If three elves can wrap three toys in hour, how many elves are needed to wrap six toys in 2 hours? [correct answer = 3 elves; heuristic answer = 6 elves]

6. Tall members of an athletics team are three times more likely to win a medal than short members. This year the team has won 60 medals so far. How many of these medals have been won by short athletes? [correct answer = 15 medals; heuristic answer = 20 medals]

7. If it takes 5 minutes for five machines to make five widgets, how long would it take for 100 machines to make 100 widgets? [correct answer = 5 minutes; heuristic answer = 100 minutes]

8. A ship has 500 crates of oranges. At the ship's first stop, 100 crates of oranges were unloaded. At the ship's second stop, 200 more crates were unloaded. How many crates of oranges were left on the ship after the second stop? [Placebo D: correct answer = 200]

9. Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are there in the class? [correct answer= 29 students; heuristic answer= 30 students]

10. A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost? [correct answer = 5 cents; heuristic answer = 10 cents]