Belief Elicitation:
Limiting Truth Telling with Information on Incentives

David Danz, Lise Vesterlund, Alistair J. Wilson[*]

June 2020

Abstract

Belief elicitation is central to inference on economic decision making. The recently introduced Binarized Scoring Rule (BSR) is heralded for its robustness to individuals holding risk averse preferences and for its superior performance when eliciting beliefs. Consequently, the BSR has become the state-of-the-art mechanism. We study truth telling under the BSR and examine whether information on the offered incentives improves reports about a known objective prior. We find that transparent information on incentives gives rise to error rates in excess of 40 percent, and that only 15 percent of participants consistently report the truth. False reports are conservative and appear to result from a biased perception of the BSR incentives. While attempts to debias are somewhat successful, the highest degree of truth telling occurs when information on quantitative incentives is withheld. Consistent with incentives driving false reports, we find that slow release of information decreases truth telling. Perversely, our results suggest that information on the BSR incentives substantially distorts reported beliefs.

Keywords: Incentive compatibility, belief elicitation, binarized scoring rule, experiments.

[*] All authors: Department of Economics, 230 Bouquet Street, Pittsburgh, PA. Danz: University of Pittsburgh and WZB, danz@pitt.edu. Vesterlund: University of Pittsburgh and NBER, vesterlund@pitt.edu. Wilson: alistair@pitt.edu. We are grateful to Tim Cason, Yoram Halevy, Paul J. Healy, Steffen Huck, Alex Imas, Dorothea Kübler, Ryan Oprea, Isabel Trevino, and seminar audiences at UCSB and WZB Berlin. We thank Felipe Araujo, Prottoy Akbar, Mallory Avery, Conor Brown, Ying Kai Huang, Matthew Raffensberger, Yuriy Podvysotskiy, and Tianyi Wang for help with the design and implementation of the experiment.

1. Introduction

Accurate information on individual beliefs is central to our ability to draw inference on economic decisions. Absent data on what people think and expect we are often unable to discriminate between alternative models of choice, gauge the limits of rationality, or test new equilibrium concepts. While belief measurement is of clear importance, beliefs, like preferences, are primitives of the decision maker, observable only through revealed choice. Incentive-compatible elicitation mechanisms therefore have a complex task: for each distinct belief there must be a corresponding choice in the mechanism that uniquely maximizes the agent's outcome, allowing the analyst to interpret the revealed choice as the agent's true belief.

Initial efforts to design incentive-compatible mechanisms centered on doing so for a narrow class of decision makers: risk-neutral expected-utility (EU) maximizers. However, conservative reporting and systematic deviations from truth telling initiated a search for mechanisms that were incentive compatible for a broader class of decision makers. Of concern was pull-to-center effects which suggested that risk aversion caused participants to report beliefs that were more conservative than those truly held (Schotter and Trevino, 2014). The binarized quadratic scoring rule (Hossain and Okui, 2013, henceforth BSR) was uncovered as a particularly promising alternative, offering incentives that make truth telling incentive compatible for any decision maker who wishes to maximize the chance of winning a prize—nesting arbitrary EU preferences. Building on the insights of Roth and Malouf (1979) the flexibility is achieved by linking reported beliefs to a pair of state-contingent lotteries. For every specific belief, the mechanism provides a unique stochastically dominant lottery pair, whereby decision makers aiming to maximize the chance of winning are given strict incentivizes to reveal their true belief.

In addition to being theoretically incentive compatible for a wider set of preferences, initial empirical evidence show that the BSR outperforms narrower mechanisms like the standard (non-binarized) quadratic scoring rule (Hossain and Okui, 2013; Harrison and Phillips, 2014), which in turn have been shown to outperform improper scoring rules (Nelson and Bessler, 1989; Palfrey and Wang, 2009) and unincentivized elicitations (Gächter and Renner, 2010; Wang, 2011; Trautmann and van de Kuilen, 2014). Better theoretical properties and evidence of superior performance has quickly rendered the BSR the state-of-the-art in belief elicitation.[1] However, despite the many positives, problems persist. Disturbingly, the conservative reporting patterns that identified problems with risk-neutral quadratic scoring have been shown to arise in BSR elicitations. For example,

[1] Recent applications of the BSR include studies on gender and coordination (Babcock et al., 2017), investment and portfolio choice (Hillenbrand and Schmelzer, 2017; Drerup et al., 2017), coordination (Masiliūnas, 2017), matching markets (Chen and He, 2017; Dargnies et al., 2019), biased information processing (Hossain and Okui, 2019; Erkal et al., 2019), cheap talk (Meloso et al., 2018), and risk taking (Ahrens and Bosch-Rosa, 2018).

Babcock et al. (2017) use the BSR to elicit beliefs over four possible outcomes and while the qualitative comparative statics mirror those for behavior, the reported beliefs are conservative, with a pull toward the center of the elicitation region.

This paper explores whether BSR incentives—in particular, subjects' knowledge of the precise quantitative incentives—encourage truth telling.[2] Challenging for this pursuit is that we generally do not know participants' true beliefs. To identify the response to information on incentives we therefore focus on beliefs elicited when participants are asked to report an induced objective prior probability. Although our experimental design also elicits posterior probabilities, our focus is on the elicited priors, as for these measurements we know that well-incentivized participants should report the given prior. Of interest is when and why participants deviate from reporting the objective prior; whether potential distortions are systematic and of a form that suggests that they arise from the incentives rather than from simple confusion; and finally, whether there are ways of modifying the BSR implementation to better encourage truth telling.

We address these questions by exploring the performance of the BSR in five different treatments using a between-subject design. An initial Information treatment provides participants with precise and transparent quantitative information on the incentives. Participants receive a description of how the BSR mechanism is implemented, and simple numerical information on the offered lotteries for all provisional responses, on top of end-of-period feedback given state realizations. Behavior in this Information treatment reveals a substantial and robust failure to report the objective prior. We show that reported beliefs are more likely to be false when it is of a non-centered prior, than of a centered prior of ½. Moreover, the misreported beliefs for non-centered priors tend to be more conservative than induced—a 'pull-to-center' effect.

Exploring what gives rise to such high rates of false reports we ask whether it results from confusion over the reporting task, or from the BSR incentives used in the elicitation. To address this question, we deploy two modifications to the Information treatment, one further supplementing the quantitative information on the incentives, the other eliminating it. A reduction-of-compound-lottery (RCL) treatment provides a further quantitative aid to participants, a calculator that helps them reduce the compound BSR lottery. While the RCL reduces false reports and eliminates the pull-to-center effect, false reports continue to be observed, and again with greater misreporting for non-centered priors. In contrast, a No-Information treatment that removes all quantitative information on the offered incentives secures a lower overall rate of false reporting, does not show differential rates by the prior location, nor does it exhibit pull-to-center effects. Thus, participants do understand the task at hand and can report the objective prior at high rates. Substantial false reporting and pull-to-center effects arise only when participants are informed of the BSR's quantitative incentives.

[2] For a more thorough review of the scoring-rule literature, see Trevino and Schotter (2014) and citations therein.

To further test the effect of the provided incentives on truth telling we conduct a Feedback treatment where quantitative incentive information is revealed gradually, through feedback at the end of each period. Confirming information on the quantitative incentives as the cause we find that end-of-period feedback gradually erodes truth telling. While false reports start out at the same low-rate as our best-case No-Information treatment, deviations from truth telling increase as the session evolves, eventually reaching the level of the Information treatment.

Finally, we take a pragmatic approach and explore a design for providing more limited information on the incentives. A Description treatment reveals that a high degree of truth telling is preserved when we present participants with a verbal non-quantitative description of how the mechanism prizes are realized. While this treatment is somewhat successful, perversely, the highest degree of truth telling is observed in the No-Information treatment.

The paper is structured as follows. Section 2 presents our design and results from the Information treatment where participants get detailed information on the quantitative incentives. Section 3 contrasts the Information results with treatments where we (i) remove all information about the incentives and (ii) add a calculator to help reduce the compound lottery inherent to the BSR mechanism. Section 4 focuses on gradual release of information on quantitative incentive through end-of-period feedback, while Section 5 examines whether we can add limited information on the mechanism's incentives and still maintain the No-Information truth telling. Section 6 shows that our main results extend to the elicited posterior beliefs and concludes with a discussion of what might be causing the deviations from truth telling under the BSR and the implications of our findings for belief elicitation more generally.

2. Baseline Design and Results

The core elicitation environment is held constant across our five experimental treatments. A total of 299 undergraduate students were recruited to participate in an individual decision-making task at the Pittsburgh Experimental Economics Laboratory (PEEL).[3] Each treatment consisted of three separate sessions (with a recruitment aim of 20 participants per session). The number of periods, the elicited belief scenarios, as well as the offered incentives are all held constant across treatments.[4] In terms of exposition, we describe the common features of the experimental environment as we introduce the Information treatment.

[3] In addition, subjects had to be 18+ years of age to be eligible for participation. Invitations to all sessions were gender balanced.

[4] Within each session, all 20 subjects see the same sequence of 10 scenarios, though in different random orders. While new state and signal realizations are drawn for each scenario across the three sessions, realizations and the sequence of task orders are matched across treatment at a session-participant level. All random realizations are therefore held constant within treatment comparisons.

2.1. Information Treatment

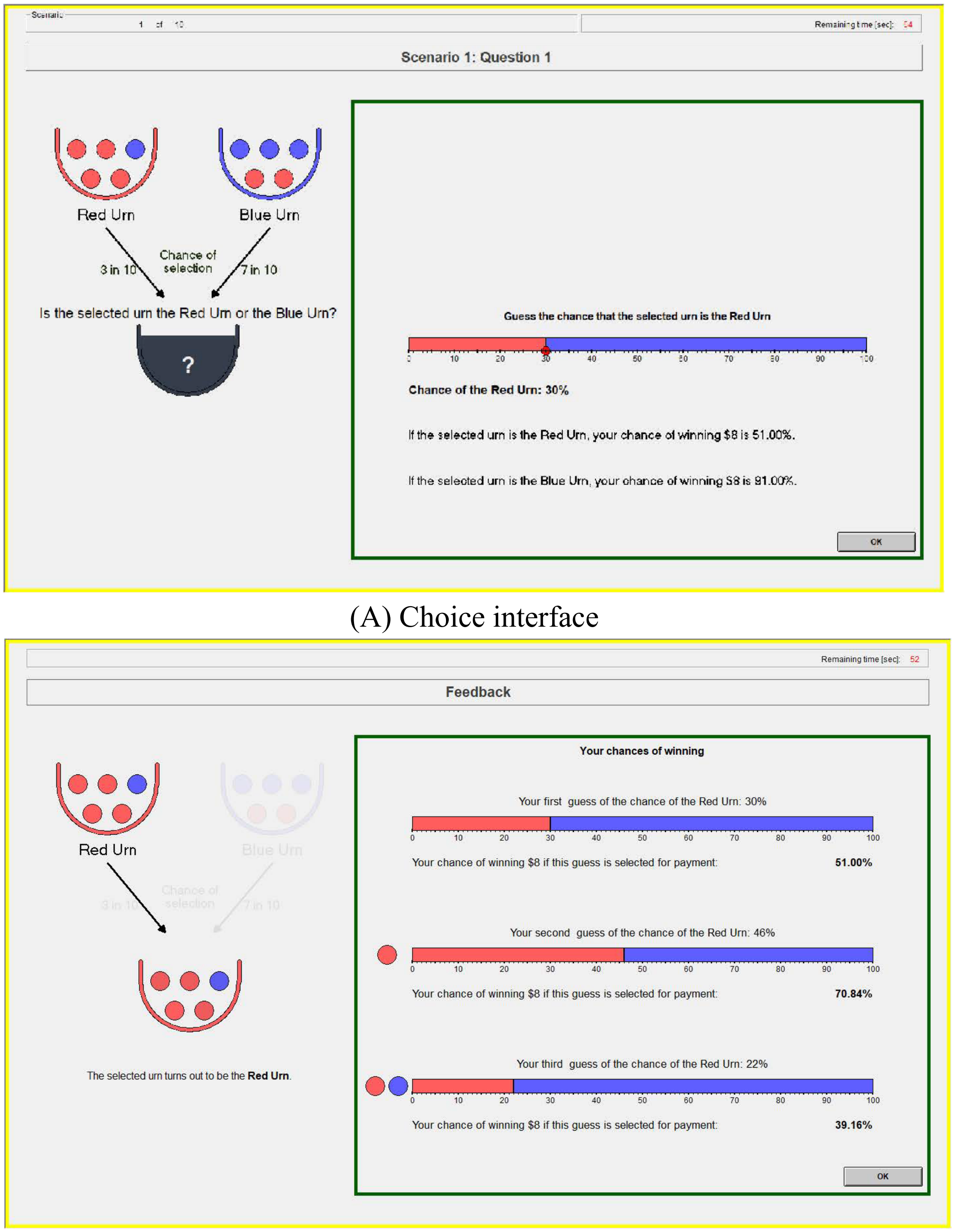
The Information treatment is designed as a baseline: an implementation of the BSR that presents participants with clear information on the quantitative incentives associated with any provisionally considered report on the belief.

Each session of the experiment consists of ten periods, where one decision from two of the ten periods are selected for payment.[5] The participants' interface is represented by the screenshots in Figure 1. Panel A shows the decision screen and Panel B the end-of-period feedback. At the beginning of each period participants are shown two urns, one Red and one Blue. Each urn contains five colored balls (either red or blue) where the Red Urn contains more red balls than the Blue Urn. One of the two urns is selected at random, where the main task for participants is to guess the likelihood that the selected urn is Red. Participants are informed of the composition of both urns and of the prior probability π_0 that the Red Urn is selected, presented as an X-in-10 chance (see panel A). Given this information, participants are asked to submit three sequential guesses on the chance that the selected urn is Red. Guess 1 is made without any additional signals, and Guesses 2 and 3 are made, respectively, after observing the colors of a first and then a second independent draw from the selected urn.

The decision screen in panel A shows a provisionally marked 30 percent guess on the Red Urn, secured by placing a cursor on a slider ranging between 0 and 100 percent (with one percentage-point increments). Each provisionally marked belief leads to an offered state-contingent lottery pair displayed on the screen: one if the selected urn is the Red Urn, another if the selected urn is the Blue Urn. Both lotteries are over a prize of \$8 if won, and \$0 otherwise. In particular, the BSR incentive given a stated probability of q on the Red Urn offers a $1\text{-}(1\text{-}q)^2$ chance of winning the prize if the Red Urn was selected, and $1\text{-}q^2$ if the Blue Urn was selected. Thus, the chance of winning is maximized by reporting the likelihood Red is selected, which for Guess 1 corresponds to the stated prior and for Guesses 2 and 3 the Bayesian posterior, updated in response to the draws from the selected urn.

Later in the paper we will introduce (sequentially) four additional treatments that vary the information on the quantitative incentives provided to participants. However, our presentation begins by outlining our results in the baseline Information treatment, which provides clear incentive information through four channels: (i) The instructions explicitly provide the qualitative information that truthful reporting is a dominant strategy (a common feature to the presentation in all of our treatments) stating that "[t]he payment rule is designed so that you can secure the largest chance of winning the prize by reporting your

[5] The experimental interface is programmed in z-Tree (Fischbacher, 2007). Participants received printed instructions that were read out loud and summarized in a short-scripted presentation at the start of each session (see Appendix B for instructions for the Information Treatment, with exact language-deltas for all other treatments; cf. reporting best practices articulated in De Quidt et al. 2019). Across all treatments the average duration of a session was 71 minutes with average earnings of \$20.08, including an \$8 show-up fee.



Scenario 1: Question 1

Is the selected urn the Red Urn or the Blue Urn?

Guess the chance that the selected urn is the Red Urn

Chance of the Red Urn: 30%

If the selected urn is the Red Urn, your chance of winning $8 is 51.00%.

If the selected urn is the Blue Urn, your chance of winning $8 is 91.00%.

(A) Choice interface

Feedback

The selected urn turns out to be the **Red Urn**.

Your chances of winning

Your first guess of the chance of the Red Urn: 30%

Your chance of winning $8 if this guess is selected for payment: **51.00%**

Your second guess of the chance of the Red Urn: 46%

Your chance of winning $8 if this guess is selected for payment: **70.84%**

Your third guess of the chance of the Red Urn: 22%

Your chance of winning $8 if this guess is selected for payment: **39.16%**

(B) End-of-period feedback

FIGURE 1. INTERFACE SCREENSHOTS

most-accurate guess." (ii) The written instructions provide a concise verbal description of how the state-contingent lotteries determine the prize realization.[6] (iii) Within the interface, as participants move their provisional belief, the screen is updated to provide clear

[6] The explanation of how the chance of winning was determined in the state-contingent lotteries relied on a comparison of the reported guess to that of two (uniform) random numbers, thereby avoiding the presentation of formulas, or the understanding of a squared error (see Wilson and Vespa, 2018).

quantitative information on the state-contingent probabilities of winning. This can be seen in Figure 1 panel A in the two lines below the input slider. With q=0.3 selected, the interface displays the associated chances of winning the prize for each realization of the selected urn; in this case 51 percent if Red, and 91 percent if Blue. (iv) Finally, participants receive feedback information on the selected urn at the end of each period, as well as the realized quantitative chance of winning the \$8 prize given the state realization (the selected urn) and their submitted Guesses 1−3 (panel B).

After the ten periods (30 elicitations total) we measure risk attitudes (encoded as switch points on pricelists) and ask participants to respond to a Cognitive Reflection Test (Frederick, 2005).[7] One participant per session is randomly selected to be paid for these end-of-experiment elicitations. Finally, participants complete a post-experiment questionnaire on demographics, and provide a self-assessment of their comprehension of the incentives and the extent to which they reported their most-accurate guess.

2.2. Information Treatment Results

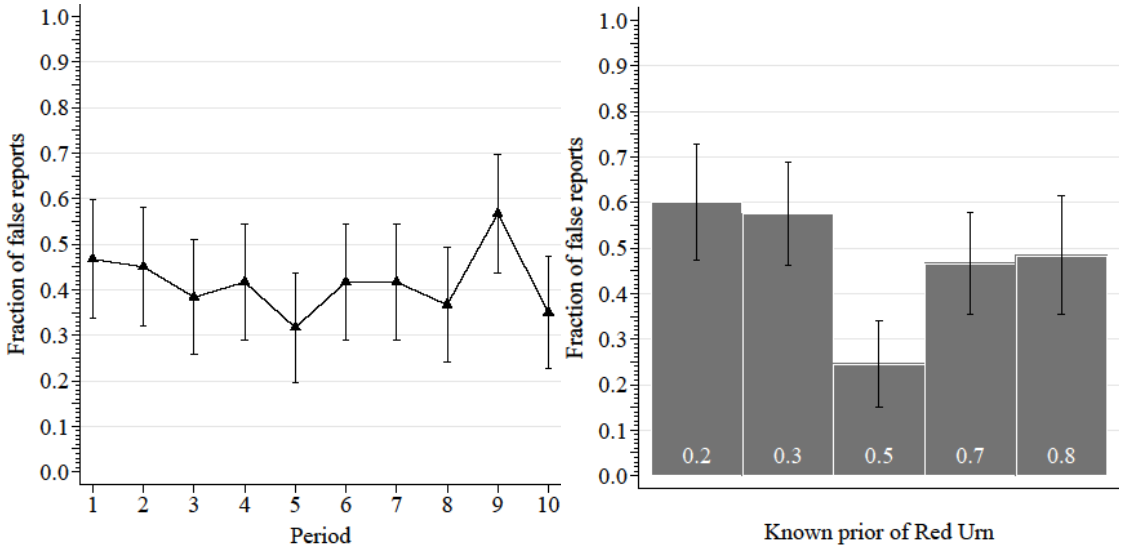
In examining whether BSR incentives secure truthful reporting we focus our analyses on Guess 1, the prior elicitation. A complementary analysis for the posteriors (Guesses 2 and 3) is discussed in Section 6, where we show that the same qualitative results hold. However, the advantage of focusing on the prior elicitations is that it removes confounds with regards to the ability to Bayesian update. The objective induced prior is unambiguous, and should be reported back by every participant who understands the offered incentives and seeks to maximize their chance of winning.[8]

Figure 2 illustrates the rate of false reports—any elicited belief q that differs from the induced prior π_0—by period across the sessions and by the objective prior. Panel A shows a substantial rate of false reporting over the ten periods, averaging 41.5 percent, which is maintained without a time trend across the session (p=0.842).[9] False reports are widespread, with 85 percent of participants failing to report the induced prior in one or more of the ten periods. These deviations from truth telling are surprising considering the incentive compatibility for such a general family of underlying preferences, the qualitative

[7] In each row of the first table, participants chose between a sure payoff of \$4 and a lottery $p \cdot \$8 \oplus (1-p) \cdot \$0$ with p increasing in each row from 0 to 1 in steps of 0.1 (see Bruner, 2009). In the second table the lottery was the same in all rows with $\frac{1}{2} \cdot \$8 \oplus \frac{1}{2} \cdot \$0$ and the sure payoff increased over the rows from \$0 to \$8 in steps of \$0.80 (see Abdellaoui et al., 2011). Tables 3 and 4 were the same as 1 and 2, respectively, except that all prize payoffs were scaled by a factor of 1.25.

[8] This simplistic elicitation eliminates belief formation and may help participants focus on the incentives provided (see for example, Avoyan and Schotter, 2016 on shared attention).

[9] Tests of time trends are based on probit regressions of false reports on period with participant-clustered standard errors.



(A) By Period (B) By Prior

FIGURE 2. FALSE-REPORT RATE IN INFORMATION TREATMENT.

statement that truth telling will maximize participants' chances of winning, and the prior evidence on the comparative superiority of the BSR.

Interestingly, the frequency of false reports varies with the location of the induced prior. Panel B illustrates how the rate of false reports varies across the five distinct prior probabilities induced within a session. For non-centered priors ($\pi_0 \in \{0.2, 0.3, 0.7, 0.8\}$) we find that false reports are the norm (52.8 percent), while they are significantly less likely to occur for the exact-centered prior ($\pi_0 = 0.5$, with a 24.6 percent false report rate, $p<0.001$ from a participant-clustered t-test).[10]

Adding to the high frequency of false reporting is the magnitude of the deviations. Conditional on a false report the average deviation from the prior is 0.167. For non-centered priors the absolute deviation does not vary with the prior ($p=0.507$), and false reports tend to be conservative with a pull-to-center effect. On non-centered priors we find that 28.3 percent of reported beliefs are false and lie between the objective prior and the exact center (a stated report of $q=½$), while only 17.2 percent fall between the objective prior and the nearest extreme (with the remaining 7.2 percent of misreports being somewhere between the exact center and the distant extreme). The pull-to-center effect shows up with a greater rate of false reports moving toward the center than the nearest extreme ($p=0.051$). In contrast to belief elicitation mechanisms such as the QSR, where the

[10] Shifting our binary definition of a false report to allow for small errors does not substantially change our results. For example, using the definition of a mistake from Cason and Plott (2014) as a difference of more than 5 percentage points from prior, we find a false-reports rate of 40.6 percent for non-centered priors and 17.1 percent for centered priors (different with $p<0.001$).

TABLE 1. INCENTIVES UNDER THE BSR

Submitted	Chance of receiving $8 if:	
Belief on Red	Urn is Red	Urn is Blue
1.0	100%	0%
0.9	99%	19%
0.8	96%	36%
0.7	91%	51%
0.6	84%	64%
0.5	75%	75%

'pull-to-center' effect was interpreted as resulting from risk aversion, the effect here is unexpected as the BSR is incentive compatible for arbitrary risk preferences. Indeed, we find no evidence that risk aversion is the culprit: Individual risk attitudes do not predict the propensity to deviate from the true prior, nor the tendency to be conservative in the report.[11]

Why then are so many participants misreporting the prior? Is it that participants fail to understand the task at hand? Or alternatively, are the offered incentives causing participants to distort their stated beliefs?

Table 1 indicates the state-contingent lottery pairs offered under the BSR for reports from 1 to 0.5 in 10 percentage-point increments and makes clear how failure to reduce the compound lottery/non-EU preferences might cause distorted reporting. Consider a participant with an induced prior of 0.8. Truthful reporting leads to a 96 percent chance of winning the $8 prize if the selected urn is Red, and a 36 percent chance if Blue. By instead (mis)reporting a more-conservative belief of 0.7, the chance of winning decreases by 5 percentage points conditional on Red and increases by 15 percentage points conditional on Blue.

While the large increase in the chance of winning on the less likely event will not provide enough of an offset for an EU-decision maker who reduces the lottery using the true likelihood of each state (here 0.8/0.2), this need not be the case for decision makers who either struggle to reduce the compound lottery and/or who make decisions in a non-EU way.[12] Moreover, by construction, the asymmetry in the gain/loss in the chance of

[11] Individual false-report rates and the extent to which these move toward the center are not significantly correlated with an individual being risk averse or loving (identified by whether willingness to pay for a 50 percent chance of winning $8 is below or above the certainty equivalent of $4). Tobits on the number of false reports made by each participant show no significant effect of being risk averse or risk loving (p=0.681 and p=0.883). Both risk averse and loving individuals are less likely to pull-to-center, however these relative effects are not significant (p=0.184 and p=0.602).

[12] Deviations can be interpreted as resulting from preferences, from calculation mistakes/ misinterpretations of the probabilities, from failures in critical thinking, from a misunderstanding of the task,

winning on the less/more likely event intensifies with the certainty in the belief. For example, consider a participant reporting 0.89 instead of a true belief of 0.99 on Red. The deviation increases the chance of winning conditional on Blue from 2 to 21 percent but only decreases the chance of winning conditional on Red from 99.99 to 98.79 percent.

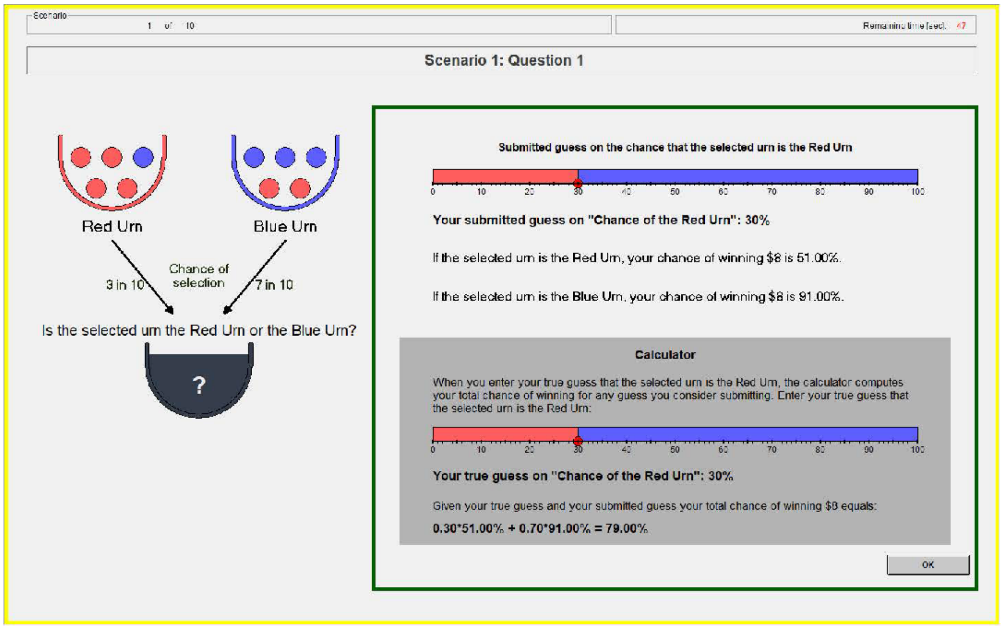
To understand the source of the deviations from truthful reporting we next explore the effects from changing the available information on incentives. In a treatment that reduces information, we examine behavior when participants receive no quantitative information on the incentives (the No-Information treatment). In a treatment that instead increases the available information, we continue to provide the state-contingent probabilities of winning, but additionally provide a calculator to help participants reduce the compound lotteries offered (the RCL treatment). The No-Information treatment thereby serves as a test for whether the failure to report the stated prior results from knowledge of the incentives or from a general misperception of the task. The RCL treatment instead offers a channel to improve participants' ability to reduce compound lotteries and to assess the extent to which false reports result from an inability to parse the offered incentives.

3. No-Information and RCL Treatments

The No-Information treatment subtracts all information on the quantitative incentives from the Information treatment. Participants are still given the qualitative information (common to all treatments) that the procedure was designed so that truthful reporting will maximize their chances of winning, but in terms of the quantitative incentives are only told that an $8 prize is at stake.[13] In addition to removing the description of the mechanism in the instructions, the No-Information interface removes the numerical information on the state-contingent lotteries at each provisionally selected belief (the figures below the input

an unwillingness to undertake the costly computation or a combination. Our after-experiment survey asked participants how they made their decisions, and the responses provide anecdotal evidence (all responses verbatim) that they purposefully distort reports to secure a higher chance of winning on the less likely event, and were aware of the incentives for doing so: "I generally erred on the side of caution when picking the urns. For example, if x=5, I would select 50% for the red urn. If sat x=8 then I would pick the red a little more opportunistically."; "I kept my initial answers at 50% because you get a 75% chance of getting the $8 anyways. Then I adjusted as I saw the different outcomes."; "at first, I guessed based on probability probability the urn was picked based on the dice roll and then considered the balls that were drawn from the bag; however, I quickly realized that since I am pretty risk adverse, sticking to a 50-50 chance would result in being paid the $8 75% of the time regardless of which urn was selected. I mostly stuck to that model as I proceeded through the experiment. When i felt daring, I would move my guesses a little bit around the 50-50 mark (but never very far)."; "I believe that leaving each chance at 75% was my best chance of making the most money in the experiment."

[13] Participants are informed that "[t]he precise payment rule details are available by request at the end of the experiment." However, of the 60 subjects in the treatment, only one requested this information at the session's end.



Scenario 1: Question 1

Is the selected urn the Red Urn or the Blue Urn?

Submitted guess on the chance that the selected urn is the Red Urn

Your submitted guess on "Chance of the Red Urn": 30%

If the selected urn is the Red Urn, your chance of winning $8 is 51.00%.

If the selected urn is the **Blue Urn**, your chance of winning $8 is 91.00%.

Calculator

When you enter your true guess that the selected urn is the Red Urn, the calculator computes your total chance of winning for any guess you consider submitting. Enter your true guess that the selected urn is the Red Urn:

Your true guess on "Chance of the Red Urn": 30%

Given your true guess and your submitted guess your total chance of winning $8 equals:

0.30*51.00% + 0.70*91.00% = 79.00%

FIGURE 3. RCL TREATMENT SCREENSHOT

bar in Figure 1 panel A) and removes the end-of-period feedback on the earned chance of winning for each Guess (the three ex post probabilities in Figure 1 panel B).[14]

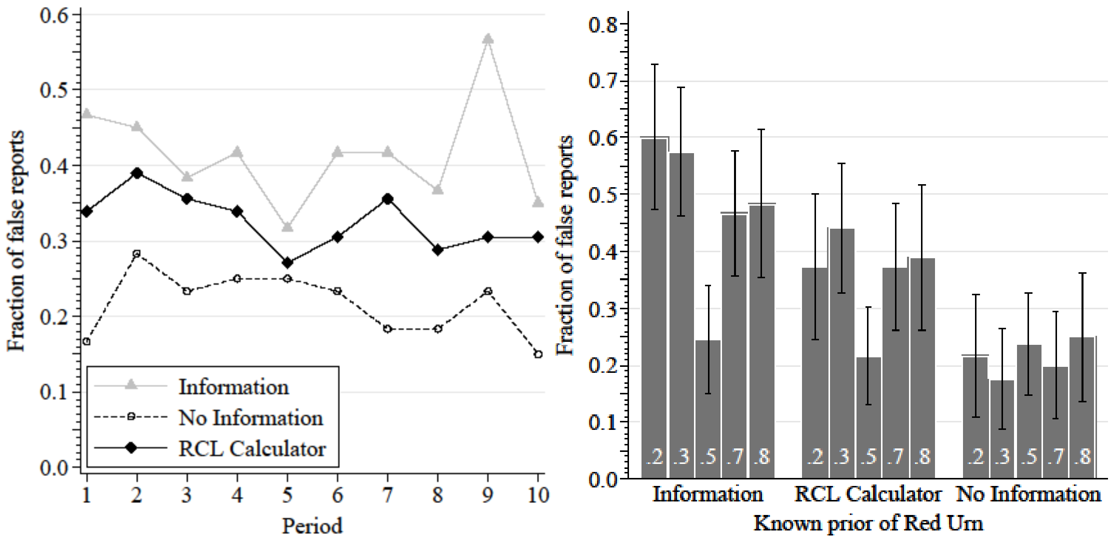
In contrast, the RCL treatment increases information provision relative to the Information treatment. Participants are provided with a Reduction of Compound Lotteries (RCL) calculator that allows them to compute the total chance of winning the prize (Figure 3 provides a screenshot of the decision screen). The RCL calculator helps participants verify that truth telling maximizes the chance of winning by performing the reduction of the lottery pair for any true and stated belief.[15] Beyond the addition of the RCL calculator (and some supplementary instructions on how to use it) the treatment is otherwise identical to the Information treatment.

Three sessions were run for each treatment, with 60 participants in No Information and 59 in RCL (one RCL session under-recruited and ran with 19 participants instead of 20). Paralleling our data presentation for the Information treatment, Figure 4 reports the false-report rate by session period (panel A) and by the objective prior (panel B).[16] We

[14] The end-of-period feedback screen in No Information instead provides feedback only on the realized selected urn.

[15] Like the more-involved explanation of the BDM in Healy (2017, 2018), the hope was that this would help participants understand that the mechanism was incentive compatible. While more substantial explanations and training may have enhanced participants' comprehension of the mechanism's incentive compatibility (and their propensity to reduce compound lotteries), because belief elicitations are typically secondary measures in experiments we opted for an aid that would not substantially increase the length of the instructions.

[16] While there are significant differences in the average absolute error across treatments, the difference is driven by the rate of false reports, hence our focus on this measure. Conditional on a false report there are



(A) By Period

(B) By Prior

FIGURE 4. FALSE-REPORT RATE IN NO-INFORMATION AND RCL TREATMENTS

note that while the RCL treatment reduces the frequency of false reports, the reduction is even greater in the No-Information treatment. Further, the pattern of submitting more false reports for non-centered than centered priors is reduced but remains in the RCL treatment, where this pattern disappears entirely in No Information. Thus, although an improved ability to reduce compound lotteries reduces false reports, the best results are obtained by eliminating quantitative information on the BSR incentives.

Table 2 reports on the rates of false reports (with participant-clustered standard errors), where the first three rows confirm that the treatment effects illustrated in Figure 4 are significant.[17] The first data column in Table 2 indicates the treatment-level false-report rate across all prior elicitations (pooled across centered and non-centered priors).[18] The two *By-Prior* columns complement the aggregate results by separating the false-reports assessment into two subcategories: those occurring when the induced prior is centered ($\pi_0 = 0.5$), and those when it is non-centered ($\pi_0 \neq 0.5$). Finally, the last three columns decompose the non-centered prior false reports into three regions to assess the extent to which agents' reported beliefs are skewed toward the center. We examine the proportion of non-centered priors

no significant differences in the magnitude of the deviation, with treatment-average deviations for false reports ranging between 0.150 and 0.169.

[17] Unless otherwise stated, all treatment comparison p-values are obtained from two-sided t-tests derived from the difference in estimates of treatment false reports rates (with standard errors clustered over the 299 participants). As all conditioning variables are indicators that fully partition the data (so the constant term is excluded), the results are equivalent to comparisons of regression coefficients obtained from a (well-specified) LPM model. Probit estimates indicate almost identical quantitative coefficients and qualitative inference, so we focus here on the easiest to interpret measures.

[18] Table 2 also reports the average treatment levels from two additional treatments introduced below.

TABLE 2. FALSE REPORTS AND TYPE BY TREATMENT

Treatment	False Reports			False-Report Type		
	All Priors	By Prior		$(\pi_0 \neq 0.5)$		
		$\pi_0 = 0.5$	$\pi_0 \neq 0.5$	Center	Near Extreme	Distant Extreme
Information	0.415 (0.042)	0.246 (0.047)	0.528 (0.048)	0.253 (0.043)	0.172 (0.032)	0.072 (0.042)
RCL	0.325 (0.041)	0.216 (0.043)	**0.398** (0.048)	**0.153** (0.036)	0.164 (0.031)	0.065 (0.014)
No Information	**0.206** (0.036)	0.179 (0.040)	**0.225** (0.048)	**0.061** (0.021)	0.133 (0.032)	**0.019** (0.009)
Feedback $_{(t=1,2)}$	**0.217** (0.045)	0.236 (0.060)	**0.200** (0.055)	**0.031** (0.021)	0.154 (0.047)	**0.015** (0.015)
Feedback $_{(t=9,10)}$	0.341 (0.053)	0.255 (0.064)	0.405 (0.071)	**0.116** (0.042)	0.275 (0.062)	0.043 (0.024)
Description	**0.245** (0.040)	0.196 (0.039)	**0.278** (0.046)	**0.106** (0.029)	0.131 (0.029)	0.038 (0.012)
N	2,630	2,630		1,568		

Note: Standard errors in parentheses clustered by participant (299 clusters) recovered from three separate joint estimates on the false report proportion in the prior elicitations: (i) *All priors*, dependent variable an indicator for $q \neq \pi_0$, with treatment level estimation; (ii) *By Prior* column pair, same dependent variable as All priors, but with separate treatment estimates for centered/non-centered prior location; and (iii) *False-Report type* column triple, treatment-level estimation for the division of non-centered false reports into three mutually exclusive regions: center (between the π_0 and ½), near extreme (between the closer extreme 0/1 and π_0), and distant extreme (the further of 0/1 and ½). Bold face coefficients are significantly different from the relevant Information coefficient with $p < 0.1$ (two-sided test).

for which a false report: (i) moves toward the center (false reports of $q \in (\pi_0, \frac{1}{2}]$ when $\pi_0 < \frac{1}{2}$, and of $q \in [\frac{1}{2}, \pi_0)$ when $\pi_0 > \frac{1}{2}$, respectively); (ii) moves to the nearest extreme (false reports of $q \in [0, \pi_0)$ and $q \in (\pi_0, 1\,]$, respectively); and (iii) moves between the exact center and the distant extreme (false reports of $q \in (\frac{1}{2}, 1]$ and $q \in [0, \frac{1}{2})$, respectively). Partitioning the non-centered false reports in this simple manner allows us to assess the extent to which participants deviate toward the center (as opposed to the near extreme).

The results for the baseline Information treatment (the first data row) mirror Figure 2: more than 40 percent of the submitted beliefs misreport the prior, where the false-report rate is significantly greater for non-centered than centered priors (p<0.001), and false reports on non-centered priors are more likely to be pulled toward the center than the nearest extreme (p=0.058).

Relative to Information we note that the RCL calculator reduces the rate of false reports. Across all prior elicitations the RCL treatment leads to a 9.0 percentage point reduction in false reports, while this reduction is not significantly different from zero (p=0.131), it is when we focus solely on the non-centered priors (a 12.9 percentage point reduction, p=0.057).[19] Despite the reduction in false reports under RCL, we still find that false reports are more likely for non-centered than centered priors (an 18.2 percentage point difference, p<0.001). However, false reports on the non-centered priors no longer pull toward the center. For non-centered elicitations, the 39.8 percent false report rate in RCL is more evenly distributed between those that move toward the center versus the nearest extreme (p=0.903). Deviations made toward the nearest extreme remain at the same rate as the Information treatment (16.4 vs 17.2 percent, p=0.851), but there is a significant drop in false reports toward the center (17.0 vs. 28.3 percent, p=0.034).

While helping participants reduce compound lotteries decreases false reports and the pull-to-center effect, we find that greater truth telling is obtained through the elimination of quantitative information on the incentives in the No-Information treatment. The proportion of false reports when participants have no quantitative information on the incentives is half that in the Information treatment and 60 percent lower than in the RCL treatment (p<0.001 and p=0.031, respectively).[20] Further false-reports rates are no greater on non-centered than centered priors (p=0.254) and there is no evidence of false reports on non-centered priors pulling to the center (p=0.112 for a two-sided test, where the difference has the opposite sign).[21]

With No Information substantially reducing the rate of false reports and eliminating the pattern of false reports being more common on non-centered priors, we infer that these effects are causally linked to knowledge of the quantitative incentives.[22] The data demonstrates that false reporting does not arise from a failure to understand the task at

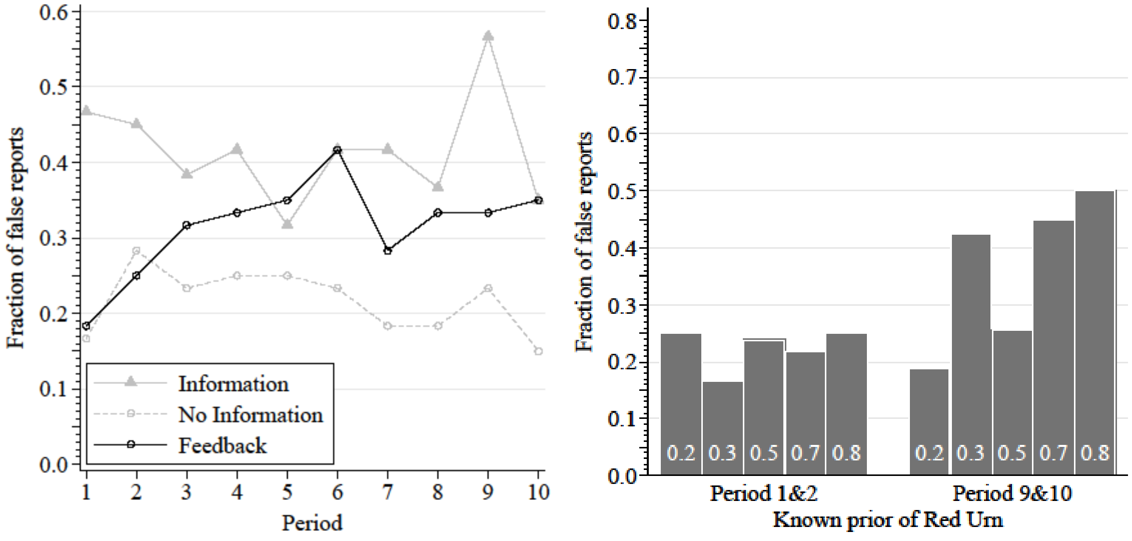
[19] Although the working hypothesis motivating the RCL treatment was one-sided—that reducing the lottery will help participants understand the incentive compatibility—we report two-sided tests throughout to be consistent.

[20] The rate of false reports across the No-Information session decreases by 5 percentage points between the last and first two periods of the treatment, though the effect is insignificant (p=0.261).

[21] Intriguingly the reduction in false reports is only seen for reports that pull-to-center (28.3 in Information vs. 7.2 percent in No Information, p<0.001) and not in those toward the nearest extreme (17.2 vs. 13.3 percent, p=0.394).

[22] The difference in the rate of false reports across centered/non-centered priors in RCL is highly significant (p<0.001), where the frequency of pull-to-center reports is significantly larger than in No Information (p=0.015). Curiously, RCL is the only treatment where false reports and pull-to-center correlates with CRT scores and attitudes toward risk. A Tobit on the number of false reports shows that risk-averse individuals are more likely to falsely report (p=0.023) but not to deviate toward the center (p=0.209); risk seeking individuals are no more likely to falsely report (p=0.620) but are more likely to pull-to-center (p=0.098). High CRT scores are predictive of fewer mistakes (0.003) and less pull-to-center (p=0.078). Risk attitudes and CRT scores are not significantly predictive of false reports in any other treatment (nor systematic in the assessed directions).

hand. Participants are fully capable of reporting the objective prior independent of its location, provided they are uninformed of the quantitative incentives of doing so.[23]



(A) By Period

(B) By Prior

FIGURE 5. FALSE-REPORT RATE IN FEEDBACK TREATMENT

4. FEEDBACK TREATMENT

The evidence thus far suggests that knowledge of the quantitative BSR incentives causes an increase in the frequency of false reports, in particular when the objective prior is non-centered. To further explore and identify the effect as coming from information on the quantitative incentives we conducted a Feedback treatment with 60 further participants, with incentive information gradually revealed through the end-of-period feedback screen (Figure 1 panel B). That is, in the Feedback treatment we replicate the No-Information instructions and decision screen, but after each period's elicitations we provide participants with the Information treatment's end-of-period feedback screen. This feedback informs the

[23] The data in our post-experimental questionnaire further bolsters the case that it is the incentives that drives the false reports. Participants are asked to rate their agreement with “I always reported my most-accurate guess on the Red urn being the selected urn.” Responses were collected on a 5-point Likert scale. Looking at the fraction of answers in the Strongly Agree/Agree categories, we find, 70 percent of respondents claiming they always reported their most-accurate guess in the Information treatment, and 85 percent in No-Information (p=0.049, χ^2-test of independence; see Figure 9 in the Appendix for further details). Self-assessment of truthful reports is (insignificantly) higher in RCL than Information (81 vs. 71 percent, p=0.149). Further, while there are no differences in comprehension of mechanism between RCL and Information, participants in No-Information are less likely to report that they understood how their pay would be calculated (70 percent) and how the submitted belief affected their pay (72 percent) than participants in the Information and RCL treatments (80 and 86 percent on pay, and 83 and 86 percent on beliefs, respectively).

participants on the earned probability of winning (given the realized state). The quantitative incentive information provided in the Feedback treatment is therefore limited to the reported beliefs, and the information is acquired slowly as the session progresses.

Figure 5 panel A indicates the false-report rate by period across the Feedback sessions. While false reports start out at the same rate as No Information, over time the fraction of false reports increases, reaching a level that is indistinguishable from that of the Information treatment.

Referring to Table 2 for comparisons and inference, we find a false-report rate of 21.7 percent for the first two periods (the $\text{Feedback}_{(t=1,2)}$ row) which grows significantly (p=0.003) to 34.2 percent for the final two periods (the $\text{Feedback}_{(t=9,10)}$ row). Thus, feedback on the quantitative incentives increases the frequency of false reports over the session, where the starting and ending points provide a strong match to the No-Information and Information treatments, respectively. Comparing overall rates of false reports in the first column of Table 2 we find that the false-report rate in the first two periods of Feedback is statistically inseparable from the overall No-Information rate (p=0.862) but significantly different from Information (p=0.001). Conversely, the final false-report rate in Feedback's last two periods is significantly different from No Information (p=0.037) but inseparable from Information (p=0.283).[24]

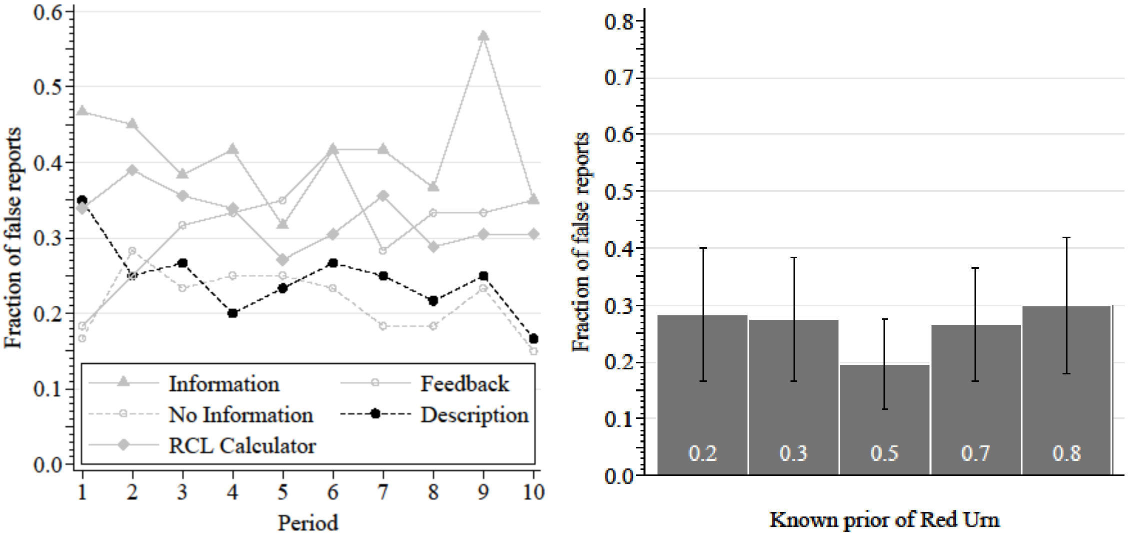
Though only provided with three ex post measures per period on the quantitative incentive (one for each submitted belief) the fraction of false reports in Feedback reaches the Information-treatment level within four periods. Despite a loss of power with our focus on the first and last two periods, the evidence suggests that participants also begin to respond differentially to non-centered and centered priors. Comparing false reports in the last two Feedback periods we find a 25.4 percent rate for the exact-center priors, and 40.6 percent for the non-centered priors (p=0.089 from a two-sided test for differences).[25] The Feedback treatment illustrates that a within-subject difference emerges even with a slow and partial release of information on the quantitative incentives, with increased false reports. As participants learn about the incentives the rate of false reports significantly increases, where the effect is most pronounced for the non-centered priors.[26]

[24] Notably, there is no significant time trend at the 10 percent level in any treatment except for the Feedback. Responses to the exit survey indicate that participants learn the incentives over time (for example, "I kept my initial answers at 50% because you get a 75% chance of getting the $8 anyways. Then I adjusted as I saw the different outcomes"), and point to some participants being drawn to the extremes "Sometimes I took extreme risks (100) because I thought they would pay off. It wasn't like I was gambiling, I wasn't going to lose my own money so I thought, might as well go all in."

[25] Using the last five periods of data instead of the last two, the difference between non-centered and centered priors is significantly different with p=0.005.

[26] Similar to the RCL treatment, we do not detect pull-to-center effects in Feedback at any point. As the feedback provided is over irreducible binary lotteries, this provides further evidence that pull-to-center is connected to the inability to reduce the compound lottery pair.

5. Description Treatment



(A) By Period (B) By Prior

Figure 6. false-report rate in Description treatment

Considering the above, a dilemma emerges for belief elicitation. On one hand we want participants to be fully apprised of the offered incentives, as knowing this is a necessary condition for incentivizing truthful reporting. On the other, the point of the mechanism is to measure beliefs with minimal noise, and information on the quantitative incentives distorts the reported beliefs. Can we inform participants partially on their incentives without distorting the reports? One option is to simply rely on the (truthful) qualitative statement that the mechanism is incentive compatible. The data from our No-Information treatment certainly suggests that this is the better option in terms of the accuracy of collected belief data. However, advocating for what amounts to a black box from the point of participants is jarring to the general philosophy of incentivized decision making.

An intermediate option is to add a description of the mechanism's implementation rule without providing the more-precise quantitative details provided in our less-successful treatments. The point of the description would be to make the quantitative incentives ostensibly calculable by subjects, and to provide a skeleton structure to participants on the rules for how their earnings will be calculated. This approach is frequently used in other mechanisms; for example, consider the non-technical description of how a second-price bidding rule works (and equivalently, how many strategy methods like the BDM function), or how a complicated matching algorithm like top-trading cycles would be described to parents in school choice.[27]

[27] Our results (with a fixed mechanism) dovetail with Holt &Smith (2016) who find evidence across mechanisms for the superiority of a BDM-based crossover elicitation. Similar to our Description treatment,

Our Description treatment pursues this approach with a further 60 unique participants. Participants in the treatment are provided with the same statement on the dominance of truth telling, but this is augmented by the short non-quantitative description of how the mechanism determines prize realizations. Mathematically inclined participants are thus informed on the mechanism's quantitative incentives, while the less mathematically inclined are provided with a procedure for how their reported beliefs are mapped into final earnings.

Figure 6 summarizes the false-report rate by period and prior for the Description treatment. Panel A reveals a moderate rate of false reports (24.5 percent) which does not differ significantly from the false-report rate in No Information (p=0.474, cf. All Prior column in Table 2) and is significantly lower than the Information treatment (p=0.004). Further, panel B shows that the rate of false reports does not vary with the prior (p=0.211), and we find no evidence that false reports are more likely to be toward the center than the nearest extreme (p=0.564).

As such, the complete ambiguity over the chosen incentive compatible mechanism in No Information can be relaxed without damaging reports. However, given the distinctly different reporting behavior under the same incentives in Information—where participants are provided the precise incentives offered—it is unlikely that participants comprehend the offered quantitative incentives in Description.[28]

6. Discussion and Conclusion

In a treatment providing participants with precise quantitative information on the BSR incentivizes we document large rates of false reports over an objective prior. These false reports are significantly more likely for non-centered than centered priors, and the deviations for the non-centered priors are more likely to move toward the center. While a calculator that helps reduce compound lotteries improves truth telling, the rate of false reports remains high. Manipulations on the available information, both between and within subjects, reveal that truth telling in the BSR is distorted by knowledge of the quantitative incentives. Overall, the highest degree of truthful reporting (and the absence of other systematic patterns in the distortions) is found in a treatment where participants have no information at all on the quantitative incentives.

their crossover mechanism does not spell out the marginal effects on the probability of winning, and focuses on the qualitative compatibility. In comparison, their QSR elicitation uses a table to make clear the marginal effects on the monetary prize .

[28] Looking at the response to survey questions and coding these as agreeing to the statement on understanding how payoffs were calculated, how a stated belief affected pay, and whether they truthfully reported we find that the Description and the No-Information treatments are statistically indistinguishable from one another in participants' self-reported understanding of the mechanism (72 vs 77 percent, p=0.532, χ^2-test) but that there are differences in understanding how beliefs affected pay (70 vs 83 percent, p=0.084), and indications of differences in self-reported inclination to report truthfully (85 vs 75 percent, p=0.171).

Our analysis focuses on elicitations over an objective prior—as this provides the cleanest test on truthful reporting. However, it could be that quantitative information on the incentives is necessary to improve elicitations where more careful thought or effort is required to form the belief. Complementing our analysis of the priors we therefore begin this section by examining the frequency and pattern of distorted reports in Guesses 2 and 3, where participants receive signals on the state and are asked to report a posterior belief. After outlining how the patterns for truth telling on the posterior elicitations dovetail with those for the prior, we then turn the discussion toward a consideration of the likely drivers for the high frequency of false reports. We conclude the paper by discussing implications for belief elicitation more generally.

After seeing the period's scenario—the composition of the two urns and the prior probability—participants first report their belief on the objective prior; they are then shown two independent draws from the selected urn and report an updated posterior belief after each signal realization. While the objective Bayesian posterior is easy for the analyst to calculate from the details given, such inference requires both probabilistic sophistication and non-trivial calculation on the part of participants. As such, the elicited posterior belief is expected to be less precise. Indeed, the number of participants who exactly report non-boundary Bayesian posteriors is just 6.9 percent across all treatments; focusing on 'truthful' reports would therefore only capture a small fraction of participants' decisions.[29] To assess false reporting when eliciting posterior beliefs, we instead characterize reported beliefs by whether they are 'distant' from the objective Bayesian posterior and assess the pattern of such 'distant' deviations across treatments.

Comparing treatments, we find that the pattern of distant-posterior reports mirrors that of false-prior reports, whether in terms of the overall rate, the sensitivity to the location of the true Bayesian posterior, and the pull-to-center effect.

Classifying distant reports as those that differ from the Bayesian posterior by more than 0.15 (the approximate average size of the deviation for false report in our prior elicitations) we find that such reports account for 37 percent of observations in the Information treatment (see Table 4 in the appendix). While the RCL treatment reduces this rate to 30.4 percent (p=0.067), distant reports are found to be least likely in the No Information treatment (27.3 percent, different from Information with p=0.008), in the early stages of the Feedback treatment (p=0.058), and in the Description treatment (p=0.025).

For the sensitivity to the location of the elicited posterior, we examine the rate of distant reports when the true posterior is at the extreme, the center or in between (intermediate). In the Information treatment distant-posterior reports are most likely when the posterior is in the intermediate region, where the distant-posterior reports in this intermediate region are more likely to pull-to-center. Reducing the quantitative information

[29] For boundary cases, the realized signal perfectly reveals the state without further calculation. For these boundary cases the Bayesian posterior is either 0 or 1, where 84.5 percent of the elicited posteriors report the true boundary belief.

on incentives reduces these patterns substantially, with decreases in distant-posterior reports across treatment primarily coming from a decrease in pull-to-center reports when the Bayesian posterior lies in the intermediate region.

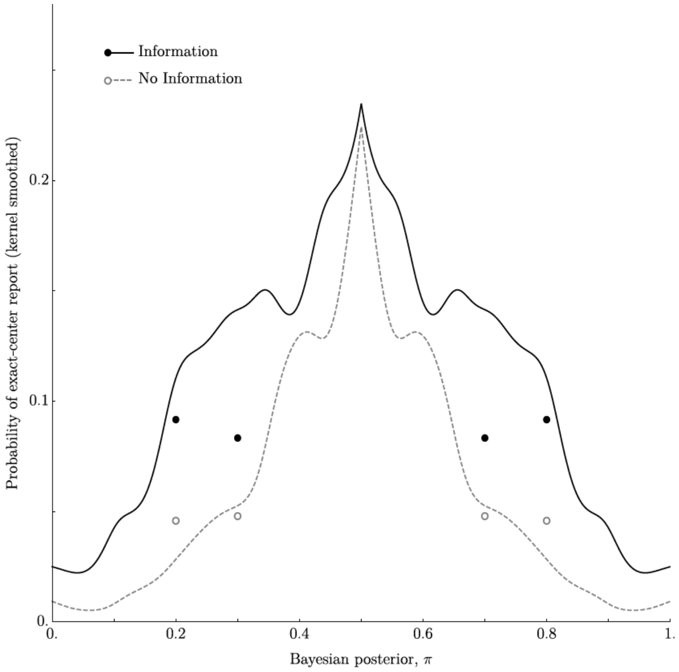


FIGURE 7. FREQUENCY OF EXACT-CENTER REPORTS BY OBJECTIVE PROBABILITY

Note: Results from a kernel-smoothed estimation of rate of exact-center reports in Guesses 2 and 3, conditional on the Bayesian posterior, enforcing symmetry around 0.5. Circles show exact-center report rates in the prior elicitations of 0.2/0.8 and 0.3/0.7. Exact center reports for centered prior elicitations not shown due to scale (0.754 for Information, 0.742 for No-Information).

Mirroring the results for the objective priors we find that information on quantitative incentives distorts truthful reporting of the posteriors. Participants appear drawn to the higher minimum chance of winning across the two states. Faced with a 0.8 chance that the selected urn is Red, a substantial minority of participants opt to report an exact-center belief of 0.50 to secure a 75 percent chance of winning independent of the state (where truthfully reporting secures a 96 percent chance of winning if Red, 36 percent if Blue, see Table 1). To illustrate this effect, Figure 7 provides a smoothed plot of the likelihood of exact-center reports at each Bayesian posterior, comparing the response between the Information and No-Information treatments. The Information treatment retains a high likelihood of exact center reports across a large range of Bayesian posteriors (approximately 0.2 to 0.8). While such reports are always less likely in No-Information, the distortive effect of information on the quantitative incentives is largest when the elicited probability is in the intermediate range (not centered or extreme).[30]

The presentation of quantitative information on the incentives therefore leads to distorted reports for both the prior and posterior elicitations, pushing reported beliefs to the center. While the phenomenon over the reported beliefs is evident, it is less clear *why* the BSR incentives gives rise to these types of deviations.[31]

TABLE 3. EXPECTED PAYOFFS AND BELIEF VARIABILITY

Treatment	Priors		Posteriors	
	Payoffs	MSE ratio	Payoffs	MSE ratio
Data:				
Information	\$6.11	1	\$6.46	1
No Information	\$6.21	0.403	\$6.53	0.807
RCL	\$6.16	0.702	\$6.52	0.820
Feedback	\$6.16	0.685	\$6.52	0.815
Description	\$6.18	0.596	\$6.51	0.840
Benchmark Models:				
Truth telling ($q = \pi$)	\$6.27	0.0	\$6.80	0.0
Middle ($q = 0.5$)	\$6.00	1.684	\$6.00	2.360
Random ($q \sim U[0,1]$)	\$5.33	5.813	\$5.33	4.331

Certainly, failure to reduce the BSR compound lottery contributes to the pull-to-center effect.[32] Truth telling increases when participants are prompted on this dimension and provided with a reduction tool. However, despite the improvement in RCL, false reports remain high and are still higher on non-centered than centered priors, a pattern that is only eliminated when quantitative information on the incentives is removed.

[30] The frequency of exact-center beliefs is larger in the Information treatment than any other treatment. For priors the Information has exact-center beliefs reported in 8.6 percent of the prior elicitations (when the true prior was non-centered) and 9.3 percent for the posteriors. Pooling the other treatments, exact center beliefs are found in 3.2 percent of priors and 4.3 percent of posteriors. Information is significantly different from the other treatments in this regard for both the prior and posterior elicitations ($p < 0.001$).

[31] Importantly, information on the quantitative incentives does not cause participants to report that they are more confused about their offered incentives. Participants were asked if they understood how their experimental payments were calculated, and if they understood how their stated beliefs were used in that calculation (Likert scale). In both questions the Information treatment has the lowest fraction of participants disagreeing (5 percent and 1.7 percent, in comparison to the average across the other treatments of 8.9 percent and 7.1 percent). A third survey question asked whether they always reported their most-accurate guess on the elicitation. Here the results mirror the belief data, with the most disagreements found in Information (10 percent), and the least in No Information (6.7 percent).

[32] Note that the distortions here are not due to an inability to perform Bayesian updating to assess the likelihood of the state. Rather, they stem from a biased assessment of the incentives used in the elicitation.

A reason for the high rate of false reports may be that deviations from truth telling are not very costly. As with many other mechanisms that attempt to fully separate a continuous type space, the BSR suffers from a flat maximum problem. That is, the marginal changes in payoffs that generate strict incentive compatibility are weak in economic terms, where small mistakes produce almost negligible effects on final earnings. Table 3 provides the expected earnings across our five treatments (given a single $8 prize), where we examine responses for the priors and posteriors separately. In addition, the table also provides the expected earnings generated under three hypothetical benchmark reporting models.

By construction, truth telling (reporting the objective probability, q = π) secures the maximum attainable expected earnings. Focusing on the elicitation of priors, truthful reporting yields and expected payoff of $6.27. In comparison, participants in the Information and No-Information treatments prior belief reports yield expected payoffs of $6.11 and $6.21, respectively.[33] Participants in our worst and best treatments therefore generate 97.7 and 99.0 percent payoff efficiencies, relative to the first-best payoff. Although these very high figures could indicate accurate responses, the counterfactual benchmark models make clear that very imprecise reports also secure high expected payoffs. For example, two reporting strategies that are entirely orthogonal to the true beliefs are reporting the exact-center belief ($q = 0.5$) and randomly choosing the reported belief (uniform noise). These reporting strategies generate expected payoffs of $6 or $5.33, respectively, corresponding to 95.7 or 85.0 percent payoff efficiencies.[34] Relative to the truth-telling first best, the expected payoffs for our BSR prior elicitations drop by less than a dollar under complete randomness and by less than 30 cents when the center is chosen in every elicitation.[35]

Though deviations in the reported belief only incur a small financial cost to the participant, the cost to the analyst in terms of precision is large. To quantify the reduction in precision for the elicited beliefs, Table 3 provides a measure for the proportional change in belief variability. The MSE ratio column provides the mean-squared error (MSE= $\frac{1}{N}\sum(q-\pi)^2$) for each treatment measured as a proportion of the MSE from the Information treatment. Examining the shift from Information to No Information, the increase in expected payoffs of ten cents for the prior elicitations shows up as a 60 percent decrease in belief variability. For the posteriors, the seven-cent change in participant payments corresponds to a decrease in the belief variability of almost 20 percent.[36]

[33] Note that the variance of the payoff is strictly increasing in the size of the deviation from truth telling.

[34] Taken to an extreme, a sophisticated agent attempting to minimize their payoff in our prior elicitations can achieve an expected payoff of $2.88 at the lowest (45.9 percent of the truth telling amount).

[35] The expected payoff for posterior beliefs shows similarly limited returns. We also note that our incentives are on the high end for belief elicitation. Many experiments purposefully keep the expected belief payment low to reduce potential interference with the main object of study.

[36] In terms of inference, a 20 percent reduction in the variance of the elicited posteriors can be the difference between a p-value of 0.075 and 0.043 in hypothesis testing; where the 60 percent reduction found for the prior shifts the same marginally significant p-value of 0.075 to highly significant (p=0.005).

Whatever the precise preference over the offered BSR incentives, Table 3 makes clear that the negative effects on participant's final payoffs from misreporting are not large, but that the decrease in precision is substantial.

Although the BSR allows us to truthfully state to participants that their chances of winning the prize are maximized by truthfully reporting, it is disturbing that despite this statement, quantitative information on the incentives causes participants to distort their reports. With the BSR being the current state of the art what does this imply for belief elicitation more generally? The underlying theoretical requirements for fully separating incentive compatible mechanisms makes the decision to collect beliefs at any resolution seem costless. However, when faced with flat incentives (and known to be flat in our treatments with quantitative information) the experiments show that participants opt not to reveal their underlying belief. Moreover, once individual rationality constraints and plausible incentive budgets are factored in, it is hard to conceive of any fully separating incentive compatible mechanism that does not suffer from a similar flatness problem.

As an alternative to pursuing ever-more sophisticated belief elicitations, it may be time to ask whether it is reasonable to expect participants in our studies to hold precise probabilistic beliefs. That is, instead of taking our results as a call for the development of elicitation mechanisms that are incentive compatible for ever-more-general classes of decision maker, we might alternatively ask whether the necessary inferences could be drawn with less-precise measurements. For example, the costs of making a false report can be increased substantially when eliciting beliefs at lower resolution.[37] Focusing on whether a belief lies within a certain interval rather than asking for an exact value allows for deviations between the offered ranges to have greater cost and may still provide the information necessary for inference. For example, suppose Policy A is predicted to be preferred when beliefs on the state are above 30 percent, and Policy B preferred otherwise. In testing the beliefs underlying policy preferences, we may secure more reliable and truthful beliefs if we simply elicited whether the belief on the state is above or below the theoretical cutoff.[38]

If elicited beliefs are collected primarily as controls or for auxiliary tests of a behavioral mechanic, it may be preferable to collect simpler measures with starker

[37] Moreover, if precise beliefs are a requirement, then dynamic mechanisms that elicit the same belief through multiple (adaptive) coarse elicitations may be a preferable route (see Schmidt and Zankiewicz, 2016).

[38] By way of example, consider a stag-hunt game where the decision to cooperate requires a belief that the other player is cooperating 80 percent of the time. A simple binary mechanism provides two state-dependent lotteries: A “not confident” option (framed as [0, 0.8]) which provides an X percent chance of winning a monetary prize if they did not cooperate, and a “confident” option (framed as [0.8, 1.0]) which provides an X/5 percent chance of winning the prize if y they did. Another starker incentive channel would be to use multinomial realizations (as opposed to the binary realizations of BSR). For example, given M independent realizations of an event, a simple elicitation asks participant to guess the number of positive realizations, with a large reward if they guess the correct number and a smaller reward if they guess within a small tolerance. For example, with ten other participants in the stag-hunt example, a prize of $10 for guessing the correct number of cooperators and $5 if guessing within two of the correct number of cooperators is incentive compatible for a division of the probability into eleven distinct reporting regions.

incentives. Moreover, the data collection should be tailored to the domain of analytical interest, rather than a one-size fits all technique. While there are many paths to improve belief elicitation, we should clearly be hesitant to adopt mechanisms with incentives that when clear to the participant, distort truth telling.

References

Abdellaoui, Mohammed, Driouchi, Ahmed, and L'Haridon, Olivier, “Risk aversion elicitation: reconciling tractability and bias minimization,” *Theory and Decision* 71, 1 (2011), pp. 63--80.

Ahrens, Steffen and Bosch-Rosa, Ciril, “Heads we both win, tails only you lose: the effect of limited liability on risk-taking in financial decision making” (2018).

Avoyan, Ala and Schotter, Andrew, “Attention in Games: An Experimental Study” (2016).

Babcock, Linda, Recalde, Maria P, Vesterlund, Lise, and Weingart, Laurie, “Gender differences in accepting and receiving requests for tasks with low promotability,” *American Economic Review* 107, 3 (2017), pp. 714--47.

Bruner, David M, “Changing the probability versus changing the reward,” *Experimental Economics* 12, 4 (2009), pp. 367--385.

Cason, Timothy N. and Plott, Charles R. “Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing,” *Journal of Political Economy*, 122 (2014), pp. 1235--1270.

Chen, Yan and He, Yinghua, “Information Acquisition and Provision in School Choice: An Experimental Study.” (2017).

Dargnies, Marie-Pierre, Hakimov, Rustamdjan, and Kübler, Dorothea, “Self-Confidence and Unraveling in Matching Markets,” *Management Science* 65, 12 (2019), pp. 5603--18.

Drerup, Tilman, Enke, Benjamin, and Von Gaudecker, Hans-Martin, “The precision of subjective data and the explanatory power of economic models,” *Journal of Econometrics* 200, 2 (2017), pp. 378--389.

DeQuidt, Jonathan, Vesterlund, Lise, and Wilson, Alistair, “Experimenter Demand Effects,” *Handbook of Research Methods and Applications in Experimental Economics* (Edward Elgar Pub., edited by Arthur Schram and Aljaz Ule).

Erkal, Nisvan, Gangadharan, Lata, and Koh, Boon Han, “Attribution biases in leadership: Is it effort or luck?” (2019).

Fischbacher, Urs, “z-Tree: Zurich toolbox for ready-made economic experiments,” *Experimental Economics* 10, 2 (2007), pp. 171--178.

Frederick, Shane, “Cognitive reflection and decision making,” *Journal of Economic Perspectives* 19, 4 (2005), pp. 25--42.

Gächter, Simon and Renner, Elke, “The effects of (incentivized) belief elicitation in public goods experiments,” *Experimental Economics* 13, 3 (2010), pp. 364--377.

Harrison, Glenn W and Phillips, Richard D, “Subjective beliefs and statistical forecasts of financial risks: The chief risk officer project,” in *Contemporary Challenges in Risk Management* (Springer, 2014), pp. 163--202.

Healy, Paul J, “Epistemic Game Theory Experiments: Utility Elicitation and Irrational Play” mimeo, (2017).

Healy, Paul J, “Explaining the BDM—Or any Random Binary Choice Elicitation Mechanism—To Subjects”, mimeo, (2018).

Hillenbrand, Adrian and Schmelzer, André, “Beyond information: Disclosure, distracted attention, and investor behavior,” *Journal of Behavioral and Experimental Finance* 16 (2017), pp. 14--21.

Holt, Charles A. and Smith, Angela M., “Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes,” *American Economic Journal: Microeconomics* 8, 1 (2016), pp. 110--39.

Hossain, Tanjim and Okui, Ryo, “The binarized scoring rule,” *Review of Economic Studies* 80, 3 (2013), pp. 984--1001.

Hossain, Tanjim and Okui, Ryo, “Belief formation under signal correlation” (2019).

Masiliūnas, Aidas, “Overcoming coordination failure in a critical mass game: strategic motives and action disclosure,” *Journal of Economic Behavior & Organization* 139 (2017), pp. 214--251.

Meloso, Debrah, Nunnari, Salvatore, and Ottaviani, Marco, “Looking into crystal balls: a laboratory experiment on reputational cheap talk” (2018).

Nelson, Robert G and Bessler, David A, “Subjective probabilities and scoring rules: Experimental evidence,” *American Journal of Agricultural Economics* 71, 2 (1989), pp. 363--69.

Palfrey, Thomas R and Wang, Stephanie W, “On eliciting beliefs in strategic games,” *Journal of Economic Behavior & Organization* 71, 2 (2009), pp. 98--109.

Roth, Alvin E and Malouf, Michael W, “Game-theoretic models and the role of information in bargaining.,” *Psychological Review* 86, 6 (1979), pp. 574.

Schmidt, Tobias and Zankiewicz, Christian, “Binary Choice Belief Elicitation: An Adaptively Optimal Design” (2016).

Schotter, Andrew and Trevino, Isabel, “Belief Elicitation in the Laboratory,” Annual Review of Economics 6,1 (2014) pp. 103--28,

Trautmann, Stefan T and van de Kuilen, Gijs, “Belief elicitation: A horse race among truth serums,” *Economic Journal* 125, 589 (2014), pp. 2116--35.

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

Wilson, Alistair J. and Vespa, Emanuel, “Paired-uniform scoring: Implementing a binarized scoring rule with non-mathematical language” (2018).

APPENDIX A: ADDITIONAL TABLES AND FIGURES

TABLE 4. POSTERIOR INFERENCE (GUESS 2 AND 3): FALSE REPORTS AND TYPE BY TREATMENT

Treatment	Distant Reports				Distant Report Movement		
	All	By Posterior Location			$\pi \in [0.15,0.35] \cup [0.65,0.85]$		
		Center	Extreme	Interm.	Center	Near Extreme	Distant Extreme
Information	0.370	0.275	0.285	0.471	0.255	0.075	0.141
	(0.026)	(0.039)	(0.040)	(0.033)	(0.027)	(0.015)	(0.022)
RCL	**0.304**	0.212	0.258	**0.377**	0.187	0.092	0.098
	(0.025)	(0.035)	(0.036)	(0.033)	(0.025)	(0.021)	(0.020)
No Information	**0.273**	0.279	**0.183**	**0.336**	0.136	0.084	0.116
	(0.025)	(0.033)	(0.029)	(0.034)	(0.019)	(0.026)	(0.020)
Feedback $_{(t=1,2)}$	**0.279**	0.300	**0.159**	**0.361**	0.194	0.111	**0.056**
	(0.040)	(0.076)	(0.052)	(0.052)	(0.042)	(0.031)	(0.024)
Feedback $_{(t=9,10)}$	0.313	0.396	0.216	**0.339**	0.127	0.110	0.102
	(0.039)	(0.066)	(0.057)	(0.047)	(0.031)	(0.031)	(0.032)
Description	**0.287**	0.254	0.235	**0.338**	0.166	0.054	0.118
	(0.026)	(0.035)	(0.033)	(0.034)	(0.019)	(0.016)	(0.019)
N	5,260		5,260			2,458	

Note: Standard errors in parentheses clustered by participant (299 clusters) recovered from three separate joint estimates on the distant-report rate: (i) *Distant Reports*, *All*, proportion of distant reports ($|q - \pi| \geq 0.15$) over treatment; (ii) *Distant Reports, By Posterior Location*, proportion of distant reports over treatment and prior location (centered, $\pi \in (0.35,0.65)$; extreme, $\pi \in [0, 0.15) \cup (0.85, 1]$; or intermediate, all other); and (iii) *Distant Report Movement* proportion of distant reports by location of the movement, conditioning on an intermediate posterior. Types in (iii) for $\pi < 0.5$ (with symmetric definition for $\pi > 0.5$) defined as: movements to the center ($q \in (\pi, 0.5]$); movements to the near extreme ($q \in [0, \pi)$); and to the distant extreme ($q \in (0.5, 1]$). Bold face coefficients are different from the relevant Information coefficient with $p < 0.1$ (two-sided test).

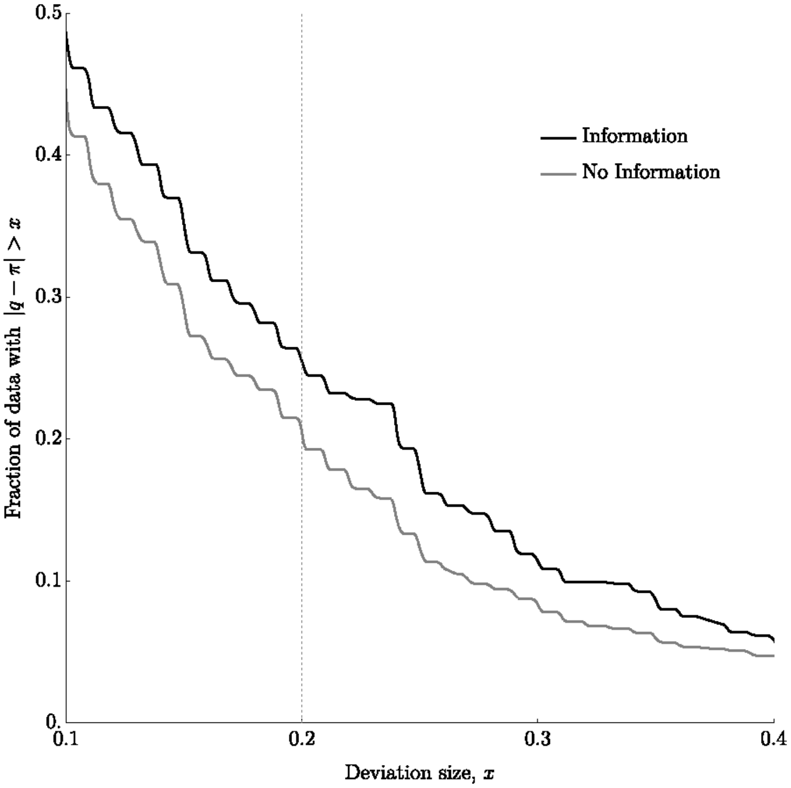


FIGURE 8. PROPORTION OF POSTERIOR REPORTS BY DISTANCE FROM BAYESIAN POSTERIOR

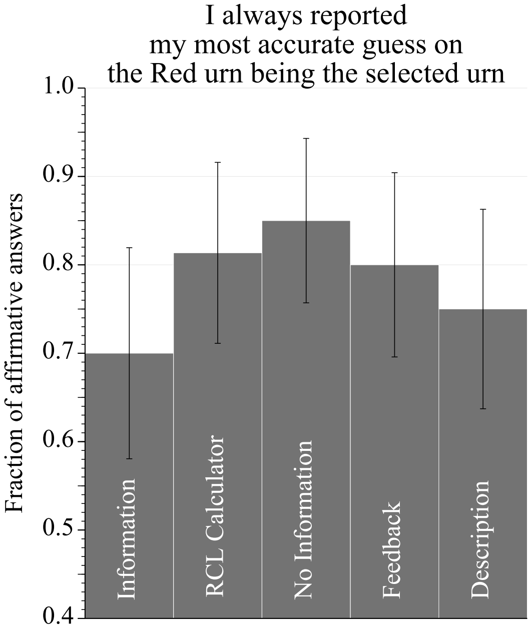


FIGURE 9. RESPONSES TO POST-EXPERIMENTAL QUESTIONNAIRE

APPENDIX B: EXPERIMENT INSTRUCTIONS[39]

Instructions

Thank you for participating in our study. This is an experiment on decision making. The other people in this room are also participating in the experiment, and you may not talk to them. If you have a question, please raise your hand and an experimenter will come and answer you in private.

You will receive $8 for participating in this experiment, but the decisions you make can further increase these earnings. Any money you make will be paid privately and in cash at the end of the experiment.

Explanation of your task

The experiment will consist of **ten scenarios**. In each scenario the computer will fill two urns with five balls, either red or blue. We call the urn with more red balls the **Red** urn, and the one with more blue balls the **Blue** urn. One of these two urns is selected to be used in the scenario. Your task is to guess how likely it is that the selected urn is the **Red** urn. Within each scenario you will make a total of three guesses.

Each scenario proceeds as follows:

Computer Fills the Urns: The two urns are filled with five balls each, some blue, some red. You will always see the exact number of blue and red balls in the two urns.

Computer Selects an Urn: The computer selects the Red or the Blue urn by rolling a fair 10-sided die and comparing it to a number **X** between 1 and 10. The selected urn is determined as follows:

- If the die roll is less than or equal to **X** then the Red urn is selected.
- If the die roll is greater than **X** then the Blue urn is selected.

Once the computer selects an urn it is fixed and stays the same for the entire scenario. The die-roll selection rule **X** means that the chance the computer selects the Red urn is **X**-in-10. For example, suppose **X**=6, then there is a six-in-ten chance (60 percent) that the computer selects the Red urn, and a four-in-ten chance (40 percent) that the computer selects the Blue urn.

The number **X** will vary across the 10 scenarios. After the computer has filled the two urns and rolled the 10-sided die to determine which urn is selected, you will be asked to make your guesses. At the beginning of each scenario you will learn how many red and blue balls there are in each urn, and the rule the computer used to select an urn (the number **X**).

[39] Treatment differences are indicated with square brackets. In addition, the RCL treatment uses the term *Your Submitted Guess* instead of *Your Guess* throughout.

However, you will not learn which of the two urns has been selected until after you have made your guesses.

You are asked to provide your best guess that the computer has selected the Red urn for the scenario. The three questions are ordered as follows:

Guess 1 Knowing only the rule **X** that the computer used to select an urn, you provide your first guess that the selected urn is the Red one.

Guess 2 The computer fairly draws one of the five balls from the selected urn. After seeing the color of this ball, you provide your second guess that the selected urn is the Red one.

Guess 3 After replacing the first-drawn ball back into the selected urn and mixing it, the computer fairly draws a second ball from the five. After seeing the color of the second ball you provide your third guess that the selected urn is the Red one.

Note that the draws from the selected urn in questions 2 and 3 are independent from one another: After the first draw is made, it is as if the ball is returned to the selected urn before the next draw is made. The contents of the selected urn are therefore always the same when a draw is made, and each of the five balls has the same chance of being drawn in each question.

[Information, No-Information, RCL, and Feedback treatment:
Feedback After you have answered the scenario's three questions you learn which urn the computer selected and drew balls from. Your three guesses will be used to determine your chances of winning an $8 prize. Your chance of winning the prize is set so that more-accurate guesses lead to a higher chance of winning.]

Your Guess

For each question you have to guess the chance that the selected urn is the Red one. Your guess is a percentage probability from 0 to 100—with 0 indicating a 0-out-of-100 chance that the selected urn is the Red urn, and 100 indicating a 100-out-of-100 chance. The number you provide is called *Your Guess*.

You choose *Your Guess* by clicking the response bar on your screen. The width of the red part of the bar indicates your guess that the Red urn was selected.

- Larger values of *Your Guess* represent a greater chance that the Red urn was selected and a smaller chance that the Blue urn was selected
- Smaller values of *Your Guess* represent a smaller chance that the Red urn was selected and a greater chance that the Blue urn was selected

The width of the blue part of the bar is *100–Your Guess*, and represents your guess that the Blue urn was selected.

[Information, RCL, and Description treatment:

Payment Rule

We now explain how *Your Guess* is used to determine whether you win the $8 prize.

- The computer chooses two numbers between 1 and 100, where each number is equally likely, as if rolling two 100-sided dice. These numbers are called *Computer Number A* and *Computer Number B*.
- The computer determines whether you win the $8 prize according to which urn was selected:

The selected urn is the Red urn: You will win the $8 prize if *Your Guess* is **greater than or equal** to **either** of the two *Computer Numbers*.

The selected urn is the Blue urn: You will win the $8 prize if *Your Guess* is **less than either** of the two *Computer Numbers*.]

[Information and RCL treatment:

To help you understand the payment rule, as you move *Your Guess* the computer will inform you of:

- The probability of winning the $8 if the Red urn was selected
- The probability of winning the $8 if the Blue urn was selected]

[RCL treatment:

As mentioned above, we designed the payment rule to make sure that your greatest **total** chance of winning is secured by letting *Your Submitted Guess* equal to your most-accurate guess that the urn is Red (what we will call *Your True Guess* on Red). We provide a calculator to help you determine your total chance of winning the prize given any *True* and *Submitted Guesses*.

The calculator will appear in a gray box on the bottom of your screen. When you have entered *Your True Guess* that the urn is Red the calculator will use *Your Submitted Guess* to compute your total chance of winning. The formula used to calculate your total chance of winning is given by:

(True Guess on Red) × (Prob. of Winning if Red given Submitted Guess)

+

(True Guess on Blue) × (Prob. of Winning if Blue given Submitted Guess).]

Final Payment

The payment rule is designed so that you can secure the largest chance of winning the prize by reporting your most-accurate guess. [No-Information and Feedback treatment: The precise payment rule details are available by request at the end of the experiment.]

At the end of the experiment, the computer will randomly choose two of the ten scenarios for payment. From each of these two scenarios, one of the three guesses will be randomly chosen for payment. Every guess has the same chance of being selected for payment. [Information, RCL, Feedback, and Description Treatment: At the end of each scenario you find out which urn was actually selected, and learn your chance of winning the $8 if the guess is selected for payment.]

For the selected questions we will use *Your Guess* and whether the selected urn was the Red urn to determine your chance of winning $8. After determining your chance of winning, the computer will conduct the lottery for the prize to see if you won the $8.

Your payment for this experiment will therefore be:

- $8 if you do not win the $8 on either guess.
- $16 if you win the $8 prize on one of the two selected guesses.
- $24 if you win the $8 prize on both selected guesses

Summary

For a brief summary please take a look at the presentation at the front of the lab.