American Economic Review 2022, 112(9): 2851–2883
https://doi.org/10.1257/aer.20201248

Belief Elicitation and Behavioral Incentive Compatibility[†]

By David Danz, Lise Vesterlund, and Alistair J. Wilson*

Subjective beliefs are crucial for economic inference, yet behavior can challenge the elicitation. We propose that belief elicitation should be incentive compatible not only theoretically but also in a de facto behavioral sense. To demonstrate, we show that the binarized scoring rule, a state-of-the-art elicitation, violates two weak conditions for behavioral incentive compatibility: (i) within the elicitation, information on the incentives increases deviations from truthful reporting; and (ii) in a pure choice over the set of incentives, most deviate from the theorized maximizer. Moreover, we document that deviations are systematic and center-biased, and that the elicited beliefs substantially distort inference. (*JEL* D83, D91)

Information on individual beliefs is central to our ability to draw inference on economic decisions (Manski 2004). 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 individual beliefs are of clear importance, eliciting them is not straightforward.

Researchers incentivize individuals to truthfully report their beliefs—with incentive compatible rules outperforming incompatible ones (Nelson and Bessler 1989; Palfrey and Wang 2009; Schotter and Trevino 2014) and these in turn dominating unincentivized elicitation (Gächter and Renner 2010; Wang 2011; Trautmann and van de Kuilen 2015).[1] While the empirical evidence demonstrates the advantage of incentive compatible elicitations, it is less evident that the offered incentives succeed in eliciting the participant's true belief. The task at hand is complex. To identify beliefs through revealed choice it must be that for every possible belief there is a corresponding choice in the mechanism that uniquely maximizes the agent's objective.

*Danz: University of Pittsburgh (email: danz@pitt.edu); Vesterlund: University of Pittsburgh and NBER (email: vesterlund@pitt.edu); Wilson: University of Pittsburgh (email: alistair@pitt.edu). Stefano DellaVigna was the coeditor for this article. We thank four thoughtful reviewers for their constructive and very helpful feedback. We also thank Tim Cason, Yoram Halevy, Paul J. Healy, Steffen Huck, Alex Imas, Dorothea Kübler, Ryan Oprea, Isabel Trevino, Emanuel Vespa, and seminar audiences at Chicago-Berkeley, Stanford, Vibes, Vienna University of Economics and Business, 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, and for excellent research assistance Pun Kanatip and Marissa Lepper.

[†]Go to https://doi.org/10.1257/aer.20201248 to visit the article page for additional materials and author disclosure statements.

[1]Performance measures for belief elicitation mechanisms include the distance between reported beliefs and realized outcomes or objective probabilities, assessment of the dispersion, support, and additivity of a report distribution, and internal consistency between beliefs and actions. We focus our analysis on elicited beliefs over an unambiguous objective prior, and assess performance by the distance between reports and the given objective prior.

Critical for designing elicitations is that the agent's objective is accounted for. Early elicitations were incentive compatible only for individuals who were risk-neutral expected-utility (EU) maximizers. However, an observed tendency to report beliefs toward the center was eventually seen as evidence that the assumption of risk neutrality was misguided, and that risk-averse participants were reporting more-conservative beliefs than those they truly held.[2] Recent design efforts have therefore aimed to find elicitations that make truth telling theoretically incentive compatible for a broader set of preferences.

We argue that to secure truthful revelation, elicitation mechanisms need to not only be incentive compatible in a purely theoretical sense, but also in a behavioral one. We propose for assessment two weak conditions for behavioral incentive compatibility, that information on deployed incentives increases truthful revelation; and that most participants, when given a choice over the pure incentives, select the outcome thought to be uniquely maximizing under the mechanism (i.e., a requirement of behavioral incentive compatibility for a representative agent).

To demonstrate we explore a state-of-the-art belief elicitation, the binarized scoring rule (BSR) (Hossain and Okui 2013). The BSR is seen as a particularly promising alternative to elicitations requiring risk neutrality because its incentive compatibility expands to arbitrary EU preferences—in fact, to any decision-maker who maximizes the overall chance of winning a prize. Building on the insights of Roth and Malouf (1979), this is achieved by linking reported beliefs to a pair of state-contingent lotteries, where for each distinct belief, the mechanism provides a lottery pair with a stochastically dominant reduction. That is, decision-makers who maximize their chance of winning are given strict incentivizes under the BSR to reveal their true belief.

In addition to being incentive compatible for a wider set of preferences, initial empirical evidence shows that the BSR outperforms its narrower forerunner, the quadratic scoring rule (Hossain and Okui 2013; Harrison and Phillips 2014). Weakened theoretical assumptions and evidence for superior relative performance has quickly rendered the BSR the preferred elicitation.[3] However, limited evidence exists on whether subjects behave in a truth-telling manner, and the conservative reporting patterns that identified failures in quadratic-scoring elicitations have also been detected in BSR elicitations. For example, in Babcock et al. (2017), despite the qualitative comparative statics for beliefs mirroring behavior, the elicited reports appeared overly conservative.

[2] For recent reviews see Schotter and Trevino (2014); Schlag and van der Weele (2015); and Charness, Gneezy and Rasocha (2021).

[3] Recent applications of the BSR include studies on gender and coordination (Babcock et al. 2017), investment and portfolio choice (Hillenbrand and Schmelzer 2017; Drerup, Enke, and von Gaudecker 2017), coordination (Masiliūnas 2017), matching markets (Chen and He 2022; Dargnies, Hakimov, and Kübler 2019; Sonsino, Lahav, and Levkowitz 2020), biased information processing (Graeber 2021; Hossain and Okui 2019; Erkal, Gangadharan, and Koh 2022; Rafkin, Shreekumar, and Vautrey 2021), cheap talk (Meloso, Nunnari, and Ottaviani 2020), risk taking (Ahrens and Bosch-Rosa 2019), information source choice (Charness, Oprea, and Yuksel 2021), memory and uncertainty (Enke, Schwerter, and Zimmermann 2020; Enke and Graeber 2019), discrimination (Aksoy, Chadd, and Koh 2021; Dianat, Echenique, and Yariv 2022, Koutout 2020), correlated and motivated beliefs (Hossain and Okui 2019; Oprea and Yuksel 2022; Cason, Sharma, and Vadovič 2020), auctions with private information (Corazzini, Galavotti, and Valbonesi 2019); information and market behavior (Filippin and Mantovani 2019; Renes and Visser 2019); institutions and cognitive ability (Choi et al. 2020); overconfidence and bargaining (Colzani and Santos-Pinto 2020); second-order beliefs (Dustan, Koutout, and Leo (2022); beliefs in strategic games (Aoyagi, Fréchette, and Yuksel 2021; Castillo, Cross, and Freer 2019).

We examine whether the incentives offered in the BSR lead to truthful revelation and in doing so demonstrate the need for elicitations to satisfy weak conditions for behavioral incentive compatibility. We start by asking whether the BSR incentives—in particular, participants' knowledge of the precise quantitative incentives being offered—encourage truthful reporting. The main finding is that information on the offered incentives increases false reports and causes systematic bias toward the center. Later, we directly assess the BSR incentives and find that most participants, when given a choice, fail to select the outcome assumed to be uniquely maximizing under the mechanism. Finally, we demonstrate the substantial impact on inference of using the center-biased reports under the mechanism and point to the potential mistakes from using elicitations that violate weak conditions on behavioral incentive compatibility.

We start by eliciting beliefs in four distinct treatments, solely varying participants' information over the BSR incentives. A challenge for examining whether information on the mechanism's quantitative incentives encourages truth telling is that we do not know participants' true beliefs. Hence, to assess the consequences of providing information on the BSR incentives we begin by evaluating reports over an objective prior. While our study also elicits posteriors that exhibit similar patterns, our focus is on eliciting the induced prior probability—where we know what a well-incentivized participant should report.

Our first treatment, a baseline *Information* treatment, provides transparent quantitative information on the incentives. Participants are told that their chance of winning is maximized by truthful reporting, and provided with a description of how the BSR mechanism is implemented, simple numerical information on the offered lotteries for all provisional responses, and end-of-period feedback. The Information treatment reveals frequent and substantial deviations in the reports from the objective prior—where we term any deviation from the induced prior as false.[4] Further, the large rate of false reports is systematically center biased—with false reports more likely for noncentered than centered priors, and the direction of deviations "pulling to center." The remaining three treatments help distinguish whether the center-biased reporting results from confusion over the reporting task or from the BSR incentives used in the elicitation. Cognitive limitations, hedging motives, and flat incentives are all factors that could result in center-biased reporting under the BSR.[5]

To demonstrate, consider a participant in our experiment asked to guess which of two urns was randomly selected (a red urn or a blue urn). Given an objective prior that the red urn is selected with probability π_0, participants are asked to report their belief q, incentivized by a state-contingent lottery pair: a lottery with a $1 - (1 - q)^2$ chance of winning \$8 if the red urn was selected; and a lottery with a $1 - q^2$ chance of winning the \$8 if the blue urn was selected.

Table 1 illustrates a subset of the state-contingent lotteries offered under the BSR. Consider a participant with an induced prior of 0.8. Truthfully reporting $q = 0.8$ yields a 96 percent chance of winning the prize if the selected urn is red, and a

[4] As such, we term reports of the objective prior as truthful. This true/false terminology is chosen for clarity, and does not imply that all participants are assumed to understand that the objective prior is the true likelihood.

[5] We consider hedging in a broad sense, as in Dean and Ortoleva (2017).

Table 1—BSR Incentive Lotteries

Submitted	Percent 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

36 percent chance if the urn is blue. Reporting the induced prior therefore secures an expected chance of winning of 84 percent, the largest feasible percentage. But cognitive limitations or hedging motives can draw participants to deviate. For example, reporting a more-conservative belief of $q = 0.7$, decreases the chance of winning by 5 percentage points conditional on red (from 96 to 91 percent) and increases it by 15 percentage points conditional on blue (36 to 51 percent). By design, for the modeled decision-maker who perfectly reduces lotteries, the trade-offs across the two lotteries do not warrant misreporting. While a movement toward the center from the true chance on red of $\pi_0 = 0.8$ yields a larger improvement on the less likely blue-urn lottery, this increase is for the modeled decision-maker not large enough to compensate for the losses on the more likely red-urn lottery. However, decision-makers motivated by hedging or who struggle to reduce the compound lottery could be drawn to more-centered reports. Further this draw to center could intensify with the certainty of the belief. For example, consider a participant reporting 0.89 instead of a true belief of 0.99 on red. Doing so increases the chance of winning conditional on blue by an order of magnitude (from 2 to 21 percent) but decreases the chance of winning conditional on red by a tiny amount (from 99.99 to 98.79 percent). Importantly, the cost of deviations from the theorized maximizer are trivially small for those tempted by a more-centered report.[6]

To examine whether the BSR incentives give rise to center-biased reporting in the Information treatment, we deploy two treatment modifications: one supplementing the provided quantitative incentive information and the other eliminating it. In the first modification, the *Reduction-of-Compound-Lottery* (RCL) treatment, we provide participants with a calculator to help them reduce the chance of winning on the lottery pair offered. While the RCL treatment helps reduce the rate of false reports and eliminates the pull-to-center effect, false reports continue to arise at high rates and remain more likely on noncentered priors.[7] In sharp contrast, when removing all quantitative information on the BSR incentives in our *No-Information* treatment, we find a substantially lower rate of false reports and no systematic bias in reporting. Our comparison of our Information versus No-Information treatments (henceforth

[6] The cost of reporting a prior other than that given is a 1 percentage point decrease in the chance of winning for a 10 percentage-point deviation in the report. For an induced prior of 0.8 a truthful report secures an 84 percent overall chance of winning. The chance of winning drops to 83 percent with a false report of 0.7 and to 80 percent with a false report of 0.6.

[7] The RCL treatment points to divergence of the reported belief from the known prior, partially resulting from subjects' misconception of the incentives (à la Cason and Plott 2014).

Information-No-Information) shows that participants understand the task at hand—as they report the induced priors at high rates in the absence of quantitative information on incentives—and that center-biased reporting results from the provided information on incentives.

Next, we conduct a *Feedback* treatment to further explore the perverse finding that information on the BSR incentives substantially and systematically biases reports. While our Information-No-Information comparison reveals the effect of incentive-information between subjects, the Feedback treatment instead evaluates the within-subject response using a more incremental revelation of incentive information (through end-of-round feedback on the earned lotteries). We find that false reports start out at the same low rate as in the No-Information treatment, but that the rate increases as participants gradually acquire information on the quantitative incentives, eventually reaching the level of false reports in the Information treatment. Confirming our initial finding, we see that it is information on incentives that leads to center-biased reporting.

To assess the impact of our finding we review prior applications of the BSR and find that the majority provide information on incentives with the potential of eliciting center-biased reports. In understanding the impact on inference of such reporting, we first explore a simple theoretical model, and second examine the magnitude of the inferential effect by replicating the paper by Niederle and Vesterlund (2007, henceforth NV) on gender and competition. We replicate the study using the BSR to elicit subjective beliefs over tournament ranking, using treatments with and without quantitative information on incentives. Through our Information-No-Information design we therefore evaluate the impact of information on incentives and the potential impact of center-biased reporting.

Abstracting from the elicited beliefs, both replications mirror the original NV finding that men compete more than women. However, treatment differences are uncovered when drawing inference that factor in the beliefs. The *NV-No-Information* treatment fully replicates the original findings that men are more overconfident than women and that this gap in confidence helps explain the gender gap in tournament entry. Looking instead at the *NV-Information* treatment, the elicited beliefs appear center biased, and (in opposition to the original results) we find both that there is *no* gender gap in confidence and that confidence *does not* help explain the gender gap in tournament entry. Thus, our replication demonstrates that information on the BSR incentives can lead to a qualitatively distinct conclusion—here in a setting that can affect policies aimed to address the advancement of women.

Our results demonstrate the substantial inferential impacts from using an elicitation that fails a weak condition for behavioral incentive compatibility: that information on the incentives increases truthful reporting. A similarly weak condition for behavioral incentive compatibility is that most participants, when given a direct choice over the set of incentives under the mechanism, will select the assumed maximizing outcome. While our Information-No-Information comparison tests the first condition (compatibility within the elicitation environment), the second condition can be assessed in an *Incentives-Only* treatment, in which subjects face the same incentives as in the BSR treatments—absent the elicitation framing. Participants in our Incentives-Only treatment are shown the set of available lottery pairs under the BSR, mirroring the options in Table 1, and asked to pick their preferred pair under a

fixed probability for the red event. As additional evidence that the elicitation is not behaviorally incentive compatible, we find that the large majority of participants fail to select the outcome assumed to be the unique maximizer. Deviations instead move towards the center, consistent with the center-bias documented in the elicitation, and indicative of incentives as the root cause of center-biased reporting.

The paper is structured as follows: Section I presents our design and results from the Information treatment where participants get detailed information on the quantitative incentives. Section II contrasts the Information results with those of two treatments, one that removes all information about the incentives, and another that adds a calculator to help reduce the compound lottery inherent to the BSR. Section III focuses on gradual release of information on the quantitative incentives through end-of-period feedback. Section IV shows that our main results extend to the elicited posterior beliefs, and Section V examines the impact of biased elicitation on inference and the impact on belief elicitations more generally.

I. Baseline Design and Results

We start by evaluating whether information on the incentives offered under the BSR elicitation increases truthful reporting. We conduct four initial treatments to assess elicited beliefs of an objective prior and to explore the response to information on incentives. The environment is held constant across all four treatments. Undergraduate students were recruited to participate in an individual decision-making task at the Pittsburgh Experimental Economics Laboratory (PEEL).[8] Each treatment consisted of three separate sessions (with a recruitment aim of 20 participants per session). The procedures of the experiment, the number of periods, the elicited belief scenarios, as well as the offered incentives were all held constant across treatments. In terms of exposition, we describe the common features of the experimental environment as we introduce the Information treatment.

A. *Information Treatment*

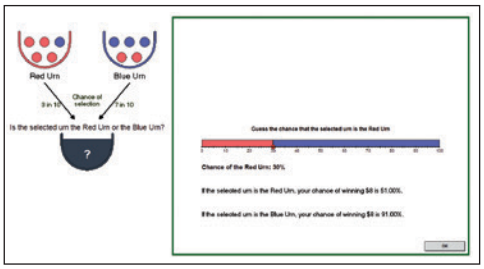
The *Information* treatment is designed as a baseline: an implementation of the BSR that presents participants with clear information on the quantitative incentives associated with any provisionally considered report on the belief. An Information treatment session (like all the treatments we examine) consists of ten periods, where each period has three sequential decisions. Participants are paid for one of the three decisions in two of the ten periods.[9]

Within the session participants make choices in the interfaces shown 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.

[8] In addition, subjects had to be at least 18 years old to be eligible for participation. Invitations to all sessions were gender balanced.

[9] 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, Vesterlund, and Wilson (2019); and Appendix C for the slides and script used in the summary presentation). Across all treatments the average duration of a session was 71 minutes with average earnings of $20.08, including an $8 show-up fee.

Panel A. Choice interface



Panel B. End-of-period feedback

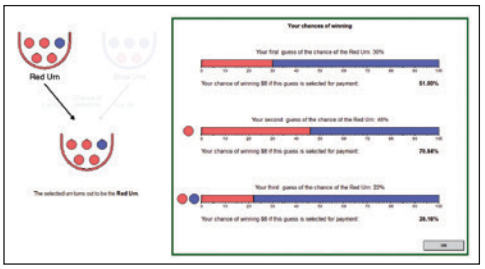


FIGURE 1. INTERFACE SCREENSHOTS

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 and 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-ten 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.[10]

The decision screen in panel A of Figure 1 shows a 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 guess 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. As noted in the introduction, the BSR incentive given a stated probability of q on the red urn offers a $1 - (1 - q)^2$ chance of winning the \$8 if the red urn is drawn, and $1 - q^2$ if the blue urn is drawn. Thus, the chance of winning is maximized by reporting the given likelihood red is selected, which for guess 1 corresponds to the induced prior, π_0, and for guesses 2 and 3 the Bayesian posteriors, is updated in response to the draws from the selected urn.[11]

Later we will introduce (sequentially) three additional treatments that vary the information on the quantitative incentives provided to participants. However, we first outline our results in the 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

[10] Our task resembles the standard Bayesian updating task with the addition of the prior elicitation, see Benjamin (2019) for a recent review of the literature on belief updating.

[11] The evaluated scenarios and random realizations are held constant across treatments. Within each session all 20 participants see the same sequence of ten scenarios, though in different random orders. While the signal realizations vary across participants and sessions of a treatment, the sequencing of scenarios considered (state, signal realizations, and sequence) is held matched across treatments.

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 most-accurate guess." This statement is also emphasized in a slide presentation summarizing the instructions where it is the last thing participants see before making their first decision (see final slide of Appendix C).

(ii) The written instructions provide a concise verbal description of how the state-contingent lotteries determine the prize realization.[12]

(iii) Within the interface, as participants move their provisional belief, the screen is instantly updated to provide clear quantitative information on the state-contingent probabilities of winning. This can be seen in panel A of Figure 1 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, as shown in panel B of Figure 1, 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 through 3.

After the ten periods (30 elicitations total) we elicit risk attitudes (encoded as switch points on price lists) and ask participants to respond to a cognitive reflection test (Frederick 2005).[13] 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.

B. *Information Treatment Results*

In examining whether BSR incentives secure truthful reporting we focus our analyses on guess 1, the elicitation of the induced prior. The 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, independent of the participants ability to Bayesian update.[14] Section III presents a complementary analysis for guesses 2 and 3, showing that the same qualitative results hold.

[12] 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).

[13] 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 $1/2 \cdot \$8 \oplus 1/2 \cdot \$0$ and the sure payoff increased over the rows from \$0 to \$8 in steps of \$0.80 (see Abdellaoui, Driouchi, and L'Haridon 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.

[14] While Offerman et al. (2009) use the elicitation of induced priors for ex post corrections, we use it to assess the elicitation procedure itself; see also Hao and Houser (2012) and Holt and Smith (2016). The simplistic elicitation eliminates belief formation and may help participants focus on the incentives provided—see for example, Avoyan and Schotter (2020) on shared attention.

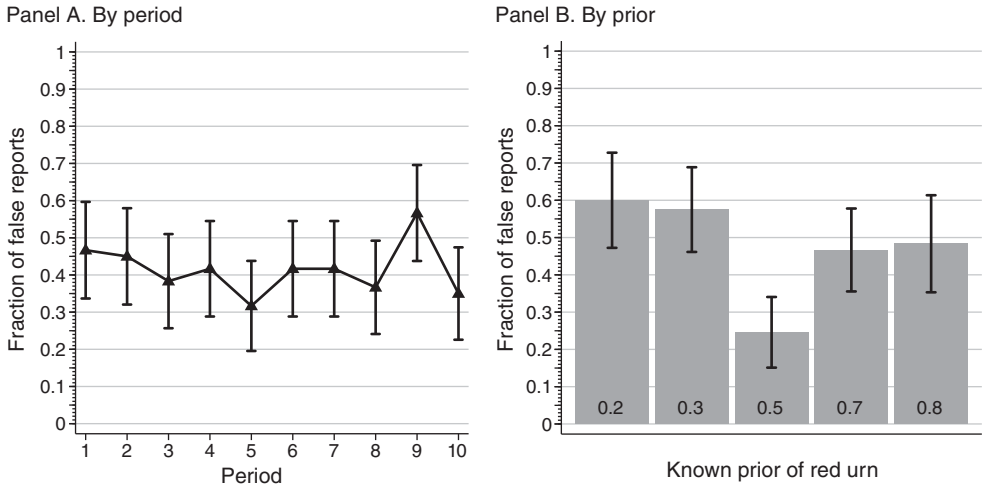


FIGURE 2. FALSE-REPORT RATE IN INFORMATION TREATMENT

Figure 2 illustrates the rate of false reports (any elicited belief q that differs from the induced prior π_0) by period across the session, 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 ($p = 0.842$).[15] 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 particularly concerning considering the incentive compatibility for a general family of underlying preferences, the qualitative statement that truth telling will maximize participants' chances of winning, and the prior evidence on the comparative superiority of the BSR. Panel B illustrates how the rate of false reports varies with the location of the induced prior. For noncentered 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 an ordinary least squares [OLS] regression with participant-clustered standard errors).

Conditional on a false report, the average deviation from the prior is 0.167, and the large proportion of false reports are not driven by small mistakes.[16] Further, false reports tend to pull to center. Among the false reports for noncentered priors we find that 53.7 percent lie between the objective prior and the exact center (a stated report of $q = 1/2$), while only 32.6 percent fall between the objective prior and the nearest extreme (with the remaining 13.7 percent of misreports being somewhere between the exact center and the distant extreme). This reveals a significant

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

[16] The results are the same when shifting our binary definition of a false report to allow for small errors that deviate by no more than 5 percentage points from the prior (Appendix Table A.1). For example, this less strict definition of false reports give rise to a false-reports rate of 40.6 percent for noncentered priors and 17.1 percent for centered priors (different with $p < 0.001$). See Cason and Plott (2014) for a similar approach when assessing the Becker Degroot Marschak mechanism.

pull to center, with greater false reports toward the center than the nearest extreme ($p = 0.058$).[17] In contrast to belief-elicitation mechanisms such as the quadratic scoring rule (QSR), where the "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 conservative reporting tendency.[18]

Why are so many participants misreporting the prior? Is it the task at hand that participants fail to understand or do they simply object to reporting the given prior? Alternatively, are the offered incentives causing participants to distort their reports? We argued in the introduction that along with the minuscule deviation cost, center-biased reporting may result because participants fail to reduce compound lotteries, or because participants hold hedging motives (preferring a substantial increase in winning on the unlikely event in return for a slight decrease in the chance of winning on the likely event).

In examining the drivers of the large degree of false-reporting in our Information treatment we consider three potential channels: (i) false reports inherent to the task of reporting the prior, for example, resulting from confusion about what is being asked; (ii) false reports driven by a failure to reduce the compound lottery inherent to the BSR incentive; (iii) false reports driven by other features of the BSR incentives, for example, through a non-EU hedging motive.[19] While the first channel captures confusion inherent to our belief elicitation, the latter two relate directly to the BSR incentives and thus point to deviations that result from failed behavioral incentive compatibility.

C. *RCL and No-Information Design*

We use two additional treatments to provide insights on the source of false reporting. Both manipulate the participants' information on the incentives. In the first, we provide *additional* information specifically tailored to limit misunderstanding of the compound lottery, providing a calculator that reduces it to a simple lottery (the *RCL* treatment). In the second, we remove all information on the quantitative incentives (the *No-Information* treatment).

[17] Partitioning the noncentered false reports in this manner secures that deviations toward the center and the near extreme have the same width, allowing us to fairly assess the extent to which participants deviate toward the center (as opposed to the near extreme).

[18] 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).

[19] Our after-experiment survey asked participants how they made their decisions, and the responses provide anecdotal evidence 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 (all responses verbatim): (i) "I generally erred on the side of caution when picking the urns. For example, if $x = 5$, I would select 50 percent for the red urn. If $x = 8$ then I would pick the red a little more opportunistically." (ii) "I kept my initial answers at 50 percent because you get a 75 percent chance of getting the \$8 anyways. Then I adjusted as I saw the different outcomes." (iii) "At first, I guessed based on 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 percent 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)." (iv) "I believe that leaving each chance at 75 percent was my best chance of making the most money in the experiment."

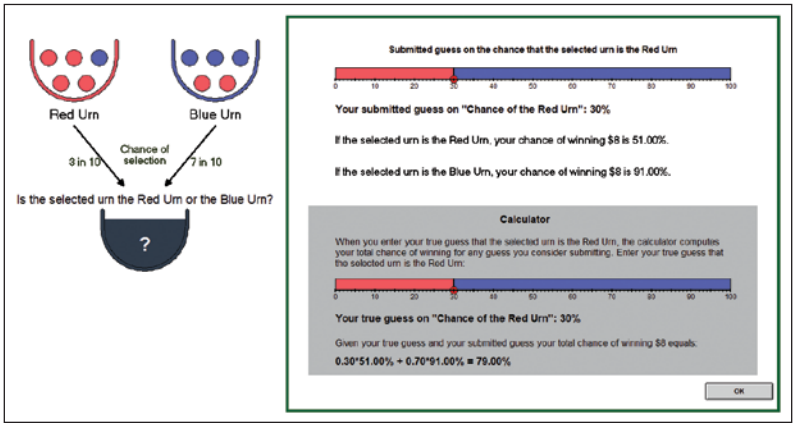


FIGURE 3. RCL TREATMENT SCREENSHOT

Our RCL calculator provides participants with a tool to compute the total chance of winning the $8 prize for any report (see Figure 3 for the interface). The RCL calculator asks participants to enter their true belief and determines the overall chance of winning for each potential report. The calculator therefore helps participants verify that truth telling maximizes their chance of winning by reducing the offered lottery pair for any true and stated belief.[20] Beyond the addition of the RCL calculator (and supplementary instructions on how to use it) the treatment is otherwise identical to the Information treatment.

In contrast, our No-Information design holds constant all qualitative details from the Information treatment but removes all quantitative information on the incentives. Participants are still told that the procedure was designed so that truthful reporting will maximize their chances of winning and that $8 is at stake as a prize, but they are uninformed on the chances of winning the prize.[21] In addition to removing the description of the mechanism in the instructions, the No-Information interface also removes the numerical information on the state-contingent lotteries at each provisionally selected belief (the two lines below the input bar panel A of Figure 1) and the end-of-period feedback on the earned chance of winning for each guess (the three ex post probabilities in panel B of Figure 1).[22]

The RCL treatment offers a channel to assess the extent to which an inability to reduce compound lotteries is driving the Information results. Given the lack of *any* incentive information in No Information, the level of false reports in this treatment serves to identify factors other than the incentives (for example, confusion with the task). By removing all quantitative incentive information, any difference in false

[20] 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 (see, for example, Burfurd and Wilkening 2018), because belief elicitations are typically secondary measures in experiments, we opted for an aid that would not substantially increase the length of the instructions.

[21] Participants are informed that *"[t]he precise payment rule details are available by request at the end of the experiment."* Of the 60 subjects in the treatment, only one requested this information.

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

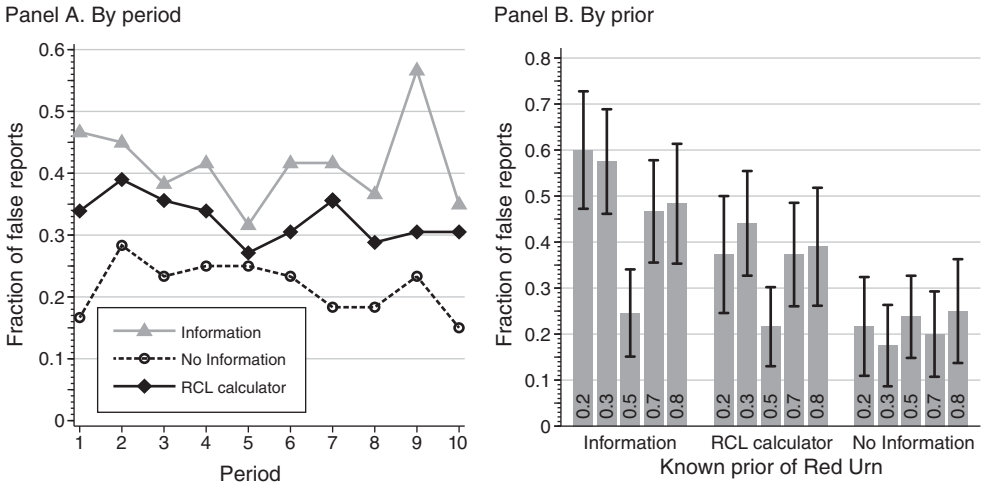


Figure 4. False-Report Rate in Information, No-Information, and RCL Treatments

reports with the Information treatment is identified as coming from some feature of the BSR incentives. The relative differences between the three treatments can therefore help to decompose the effects on false reporting.

D. *RCL and No-Information Results*

Three sessions were run for each treatment, with 60 participants in No Information and 59 in RCL.[23] 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).[24] We note that while the RCL treatment reduces the frequency of false reports, the reduction is even greater in the No-Information treatment.[25] We also see that the pattern of greater false reporting for noncentered than centered priors is reduced but not eliminated in the RCL treatment, while it disappears entirely in No Information.[26] Thus, although an improved ability to reduce compound lotteries decreases false reports, the best results are obtained by eliminating quantitative information on the BSR incentives.

As evidence of center-biased reporting we illustrate in Figure 5 the distributions of reports at asymmetric priors of 0.2 and 0.3 (panels A and B, respectively). On the left of each panel, we indicate the overall rates of truthful reporting for the asymmetric priors by treatment. On the right of each panel, we illustrate the distributions of

[23] One RCL session underrecruited and ran with 19 participants.

[24] 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 no significant differences in the magnitude of the deviation from truth telling, with treatment-average deviations for false reports ranging between 0.156 and 0.183.

[25] For comparison we note that Holt and Smith (2016) use a similar two-urn-guessing paradigm to elicit beliefs under the QSR, the BDM, and the lottery choice method. Evaluating only one prior of 0.5 they find false report rates on the prior elicitation ranging between 20 to 33 percent.

[26] Further the patterns in the results continue to hold when eliminating small mistakes within 5 percent of the induced prior (see online Appendix Table A.1).

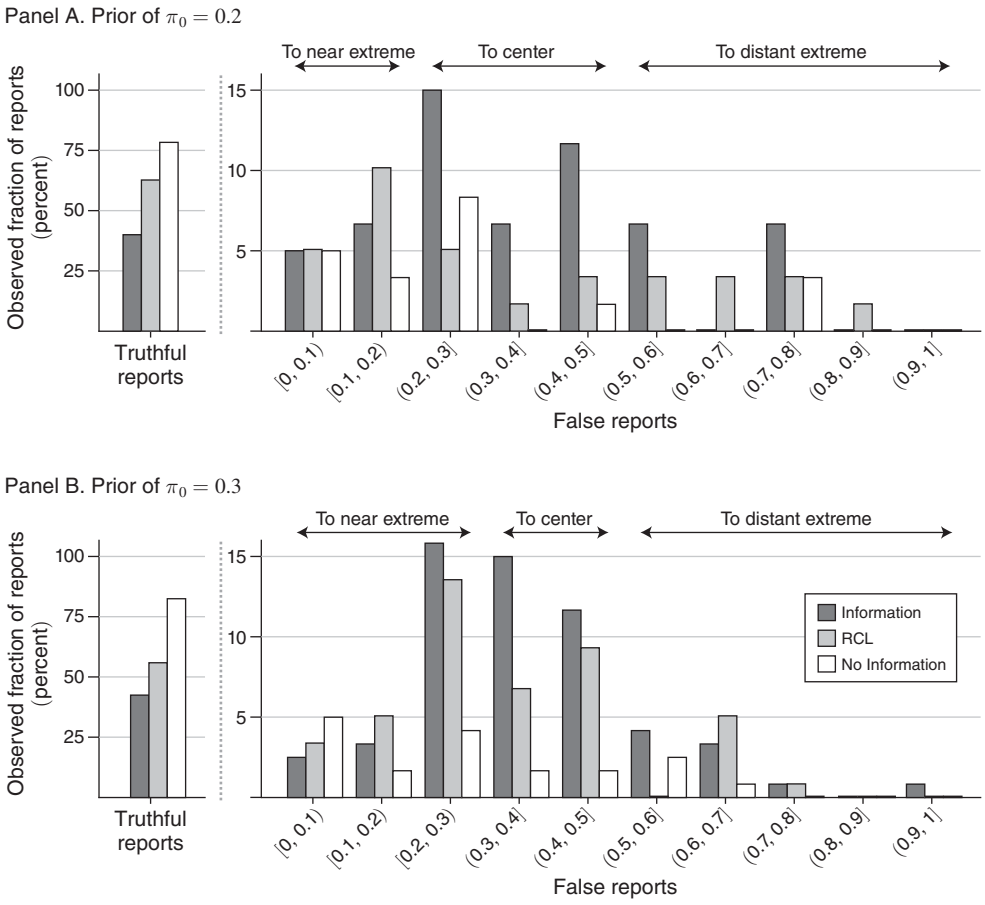


FIGURE 5. PROPORTION TRUTHFUL, AND DISTRIBUTION OF RESPONSE FOR FALSE REPORTS

false reports. Compared to the RCL and No-Information treatments we see for the Information treatment not only the low rate of truthful reporting, but also the center bias with greater false reporting toward the center than the near extreme.

The center-bias effects illustrated in the figures are quantified in Table 2 which shows the false-report rates with participant-clustered standard errors, where the first three rows confirm that the treatment effects illustrated in Figures 4 and 5 are significant.[27] The first data column in Table 2 indicates treatment-level false-report rates across all prior elicitations (pooling centered and noncentered priors).[28] The two *by prior* columns separate the false-report assessment into two subcategories: those occurring when the induced prior is centered ($\pi_0 = 0.5$), and those when it is noncentered ($\pi_0 \neq 0.5$). Finally, the last three columns decompose the false reports for noncentered priors into three regions, mirroring those labeled at the top of panels A and B in Figure 5, to assess the extent to which reported beliefs

[27] Unless otherwise stated, all treatment comparison *p*-values are obtained from two-sided tests (OLS with standard errors corrected for clusters at the individual level). Probit estimates indicate almost identical quantitative marginal effects and qualitative inference, so we focus here on the easier to interpret measures.

[28] Table 2 also reports the average treatment levels from two additional treatments (details below).

TABLE 2—FALSE REPORTS BY TREATMENT

Treatment	False reports			False-report type ($\pi_0 \neq 0.5$)		
		By prior				
	All priors	$\pi_0 = 0.5$	$\pi_0 \neq 0.5$	Center	Near extreme	Distant extreme
Information	0.415	0.246	0.528	0.283	0.172	0.072
	(0.042)	(0.047)	(0.048)	(0.042)	(0.032)	(0.014)
RCL	0.325	0.216	0.398	0.169	0.164	0.065
	(0.041)	(0.043)	(0.048)	(0.033)	(0.031)	(0.014)
No Information	0.217	0.237	0.203	0.058	0.108	0.036
	(0.039)	(0.040)	(0.041)	(0.018)	(0.027)	(0.014)
Feedback$_{(t=1,2)}$	0.217	0.236	0.200	0.031	0.154	0.015
	(0.045)	(0.060)	(0.055)	(0.021)	(0.047)	(0.015)
Feedback$_{(t=9,10)}$	0.342	0.255	0.406	0.087	0.275	0.043
	(0.053)	(0.064)	(0.071)	(0.039)	(0.062)	(0.024)
Description	0.245	0.196	0.278	0.108	0.131	0.039
	(0.040)	(0.039)	(0.046)	(0.027)	(0.029)	(0.012)
Observations	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/noncentered prior location; and (iii) *False-report type* column triple, treatment-level estimation for the division of noncentered false reports into three mutually exclusive regions: *Center* (between the π_0 and $1/2$), *Near extreme* (between the closer extreme $0/1$ and π_0), and *Distant extreme* (the further of $0/1$ and $1/2$).

are skewed toward the center. We examine the proportion of noncentered priors for which a false report (i) moves toward the center (false reports of $q \in \left(\pi_0, \frac{1}{2}\right]$ when $\pi_0 < \frac{1}{2}$, and of $q \in \left[\frac{1}{2}, \pi_0\right)$ 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 $q \in \left(\frac{1}{2}, 1\right]$ and $q \in \left[0, \frac{1}{2}\right)$, respectively). The results for the baseline Information treatment (the first data row) mirror Figure 2: more than 40 percent of the reports do not equal the objective prior, where the false-report rate is significantly greater for noncentered than centered priors ($p < 0.001$), and that false reports on non-centered priors are more likely to pull to 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 elicited priors the RCL treatment leads to a 9.0 percentage point reduction in false reports. While this reduction is not significantly different from zero ($p = 0.130$), it is significant when we focus solely on the noncentered priors (a 12.9 percentage point reduction, $p = 0.056$).[29] Adding to the reduction in false reports is that we no longer see false reports that pull to center.[30] The greater

[29] While 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 for consistency.

[30] The RCL treatment's 39.8 percent false-report rate for noncentered elicitations is more evenly distributed between those that move toward the center, and those that move to the nearest extreme ($p = 0.903$). Specifically, while deviations made toward the nearest extreme in RCL continue to occur at a similar rate to Information (16.4 versus 17.2 percent, $p = 0.851$), we find a significant reduction in the false reports moving toward the center (16.9 versus 28.3 percent, $p = 0.034$).

frequency of false reports on noncentered than centered priors however remains (an 18.2 percentage point difference, $p < 0.001$).

The proportion of false reports decreases even further when participants have no quantitative information on the BSR incentives (significantly lower than Information and RCL, $p < 0.001$ and $p = 0.056$, respectively).[31] Further, there is no evidence of center-biased reporting. False-report rates are no greater on noncentered than centered priors ($p = 0.317$) and there is no pull to center ($p = 0.175$ for a two-sided test, where the difference has the opposite sign).[32] With No Information substantially reducing the rate of false reports and eliminating center-biased reporting, we infer that both effects are causally linked to knowledge of the quantitative BSR incentives.[33] The No-Information treatment therefore demonstrates that false reporting in the Information treatment does not simply arise from confusion over the task. Participants report the objective prior at high rates, independent of its location, provided they are *uninformed* of the quantitative incentives from doing so.[34]

Comparing the three treatments we get a sense of what drives participants to falsely report noncentered priors in the Information treatment. The false report rate on noncentered priors is 52.8 percent in Information treatment and 20.3 percent in the No-Information treatment, which suggests that 38 percent ($= 0.203/0.528$) of the false reports can be attributed to the task itself (e.g., confusion). The remaining 62 percent are directly linked to the BSR incentives, whether through an inability to reduce compound lotteries or another feature of the offered incentives. The estimated 39.8 percent false-report rate in the RCL treatment roughly suggests that failure to reduce compound lottery accounts for approximately 25 percent ($= (0.528 - 0.398)/0.528$) while the remaining 37 percent results from some other aspect of the incentives ($= (0.398 - 0.203)/0.528$).

We find in our three treatments little evidence that false reporting is correlated with individual characteristics. Individual risk attitudes are not predictive of the rate of false reports or of center-biased reporting in any treatment, nor are individual cognitive reflection scores predictive of false or center-biased reporting in the Information and No-Information treatments. Only in the RCL treatment do we see evidence that cognitive reflection impacts behavior, with higher levels of cognitive reflection being predictive of fewer false reports and less center-biased reporting.

[31] The rate of false reports across the No-Information sessions decreases by 4 percentage points between the last and first two periods of the treatment, though the effect is insignificant ($p = 0.420$).

[32] Intriguingly, the reduction in false reports is only seen for reports that pull to center (28.3 in Information versus 5.8 percent in No Information, $p < 0.001$) and not in those toward the nearest extreme (17.2 versus 10.8 percent, $p = 0.160$).

[33] The difference in the rate of false reports across centered/noncentered priors in RCL is highly significant ($p < 0.001$), where the frequency of pull-to-center reports is significantly larger than in No Information ($p = 0.004$).

[34] The data in our postexperimental 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 five-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 A.1 in the online Appendix for further details). Self-assessment of truthful reports is (insignificantly) higher in RCL than Information (81 versus 70 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 (72 percent) and how the submitted belief affected their pay (70 percent) than participants in the Information and RCL treatments (80 and 86 percent on pay, and 83 and 86 percent on beliefs, respectively).

This suggests that the RCL calculator is particularly helpful to those with high levels of cognitive reflection, and that even with a tool tailored to simplify decision-making under transparent incentives, cognitive skills and effort are required for subjects to benefit from such aids.

II. Feedback Treatment

Our Information-No-Information comparison presents between-subject evidence that information on the quantitative incentives drives center-biased reporting under the BSR, which is indicative of the elicitation lacking behavioral incentive compatibility. To further explore and identify the effect as coming from information on the quantitative incentives we conduct a *Feedback* treatment with 60 additional participants. Incentive information in this treatment is gradually revealed through end-of-period feedback screen (see panel B of Figure 1). That is, in the Feedback treatment we replicate the No-Information instructions and main decision screen, but after each period's elicitations we provide participants with the Information treatment's end-of-period feedback, which informs them on the earned probability of winning as a simple lottery (given the realized state). The quantitative incentive information provided in the Feedback treatment is therefore acquired gradually as the session proceeds and is limited to the reported beliefs and realized state.

Panel A of Figure 6 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, eventually 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 $Feedback_{(t=1,2)}$ row) which grows significantly ($p = 0.003$) to 34.2 percent for the final two periods (the $Feedback_{(t=9,10)}$ 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. The false-report rate in the first two periods of Feedback is statistically inseparable from the overall No-Information rate ($p = 1.000$) but significantly different from Information ($p = 0.001$). Conversely, the false-report rate in Feedback's final two periods is significantly different from the No-Information treatment ($p = 0.060$) but inseparable from Information ($p = 0.282$).[35]

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 the lower power when focusing on the first and last two periods, panel B of Figure 6 suggests that participants also begin to respond differentially to noncentered and centered priors. Comparing false reports in the last two Feedback periods we find a rate of

[35] Notably, there is no significant time trend at the 5 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 percent because you get a 75 percent chance of getting the $8 anyways. Then I adjusted as I saw the different outcomes").

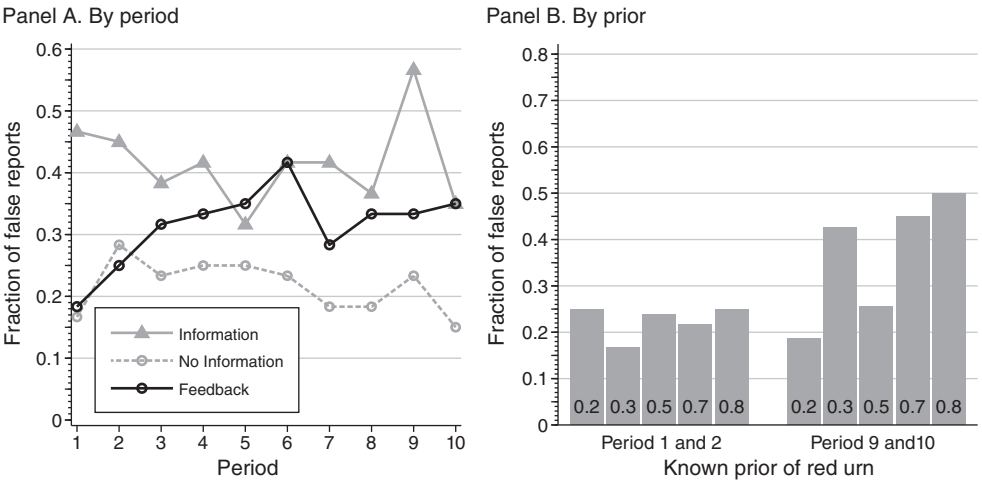


FIGURE 6. FALSE-REPORT RATE IN FEEDBACK TREATMENT

25.5 percent for the exact-center priors, and of 40.6 percent for the noncentered priors ($p = 0.089$ from a two-sided test of difference).[36]

The Feedback treatment shows that center-biased reporting increases significantly, as participants gradually learn about the BSR incentives. These results are consistent with our Information-No-Information finding that information on incentives decreases truthful reporting—and serves as further evidence that the BSR is not incentive compatible in a behavioral sense.

III. Posterior Reports

Our analysis has focused on elicitations of the induced prior—as this provides the cleanest measure on truthful reporting, in that there is no concern over participants' ability to perform Bayesian updating. However, it may be that quantitative-incentive information encourages accurate reporting when we are eliciting beliefs that are not objectively known, or that require effort to form.[37] Mirroring our analysis on the priors we examine the frequency and patterns of distortion in the reports for guesses 2 and 3, where participants receive signals on the state and are asked to report a posterior belief.

After observing the setup for a given scenario—the composition of the two urns and the prior probability—participants report their belief on the objective prior in guess 1. They are then sequentially shown two independent draws from the selected urn, and are asked to report an updated posterior belief after each draw. While the objective Bayesian posterior is easily determined by the analyst from the provided

[36] Using the last three periods of data instead of the last two, the difference between noncentered and centered priors (with participant-clustered errors) is significantly different with $p = 0.046$.

[37] Note that a perhaps more intuitive counterargument could be made: when additional cognitive efforts are needed incentives could become less salient or less compensating for the additional effort required. Further, even if participants are taking the incentives more seriously for harder tasks, it is unclear why this should decrease rather than increase the hedged pull-to-center reporting.

details, such inference is nontrivial and requires probabilistic sophistication.[38] As such, we expect the elicited posterior beliefs to deviate from the Bayesian posteriors. Indeed, the number of cases in which participants exactly report nonboundary Bayesian posteriors is just 6.9 percent across all treatments. Focusing on "truthful" reports would therefore only capture a tiny fraction of participant decisions, and it would exclude reports of truly held, but non-Bayesian, posterior beliefs.[39]

To assess false reporting on elicited posterior beliefs, we instead characterize reported beliefs by whether they are *distant* from the objective Bayesian posterior. We then assess the pattern of distant reports across treatments. Although the distant-posterior treatment effects are smaller than those for false reports on the priors, the qualitative patterns mirror the previous results—in terms of the total rates, the sensitivity to location, and the pull-to-center effect. We classify distant reports as those that differ from the Bayesian posterior by more than 15 percentage points (the approximate average deviation for a false report in our prior elicitations).[40] Such reports comprise 33 percent of the data in the Information treatment, where distant reports make up 25–28 percent for the other treatments (except for the last rounds of the Feedback treatment).

Further, there is evidence of center-biased reporting when participants are informed of the quantitative incentives.[41] We can further explore this by evaluating how the likelihood of reporting perfectly centered beliefs varies by treatment. In the Information treatment 11.6 percent of the elicitations with an intermediate Bayesian posterior have participants report a perfectly centered belief, where the rate in No Information is a quarter this size at 2.9 percent ($p < 0.001$ for the comparison). To illustrate the differential response between Information and No Information, Figure 7 indicates the fraction of exact-center reports in each as a function of the Bayesian posterior π (where we additionally plot the exact-center report rates for the prior elicitations). The figure makes clear that for both the prior and posterior beliefs, exact-center false reports are less likely under No Information than Information.

In the online Appendix (Figure A.3) we further show that information on incentives moves the range of reports toward the center. Looking across our 60 elicitation scenarios (the priors, the urn compositions, and the realized signal draws) we find a significantly greater tendency for the distribution of reports to move towards the center in the Information than the No-Information treatment.[42]

[38] See Benjamin (2019) for a survey of the literature on behavior in Bayesian updating tasks.

[39] For boundary posteriors, the realized signal perfectly reveals the state through a simple inference without the need for calculation. For boundary cases where the Bayesian posterior is either 0 or 1, 84.5 percent of the elicited posteriors report the true boundary belief.

[40] In the online Appendix, Figure A.2 shows that the results are not sensitive to what we define as "distant."

[41] For evidence of center-biased reporting we parallel in online Appendix Table A.2 the analysis presented in Table 2. First, we show that treatment differences over distant reports are driven by elicitations where the true Bayesian posterior is in an intermediate range ($0.15 \leq \pi \leq 0.35$ or $0.65 \leq \pi \leq 0.85$): with a 41 percent distant-report rate for such posteriors in Information, decreasing to 31 percent in RCL (different from Information with $p = 0.024$) and 32 percent in No Information ($p = 0.040$). By contrast there are no treatment differences for posteriors in the central region ($0.35 < \pi < 0.65$). Second, we show pull to center by examining three distinct regions when the true posterior is *intermediate*: (i) posterior beliefs at the exact center ($q = 1/2$), (ii) posterior beliefs as the nearest extreme belief of $q = 0$ or $q = 1$ to the Bayesian posterior, and (iii) posteriors beliefs in the wrong half of the elicitation interval (so any $q > 1/2$ if $\pi < 1/2$ and vice versa). While we find no significant differences in the rates of near extreme or wrong-half posterior reports between Information and No Information/RCL, there are strong significant differences in the frequency of exact-center reports ($p < 0.014$).

[42] Using the upper (lower) quartiles for the elicited beliefs in the 58 matched scenarios where the Bayesian posterior was below (above) 0.5, we form a Wilcoxon signed-rank test for equivalence between Information and

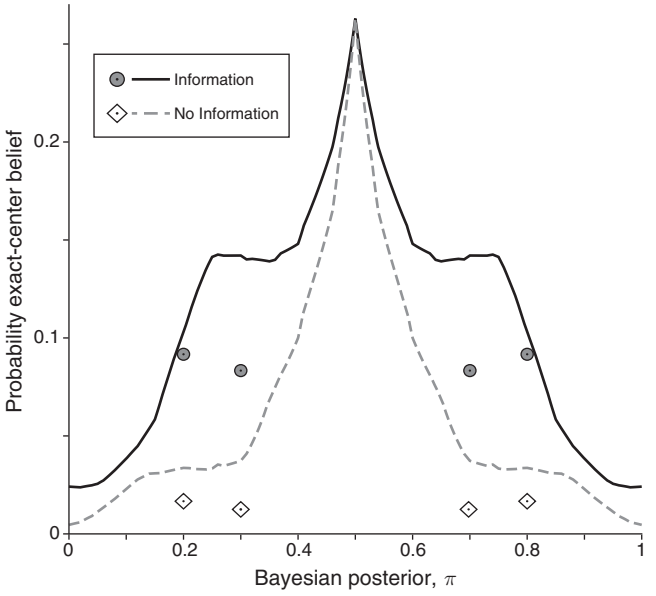


FIGURE 7. EXACT-CENTER REPORTS BY BAYESIAN POSTERIOR

Gaining clear identification of when a particular posterior report is distorted is certainly more challenging than for the priors, as we do not observe the participant's true updating rule. While this focuses our analysis on more-specific measures of false reporting, the general patterns mirror our findings when eliciting the induced priors. Information on the quantitative BSR incentives distorts the posterior reports toward the center.[43]

IV. Discussion: Impact and Implication

Our results show a large rate of false reports when providing participants with clear quantitative information on the BSR incentives. Only 15 percent of participants consistently report the induced prior under Information, in comparison to 50 percent under No Information. Moreover, the distortions are systematically center biased, with false reports being more likely for noncentered than centered priors and pulling to center. By contrast, false reports are not center biased in the absence of information on incentives.

Our consistent finding that information on incentives increases false reports points to the BSR failing a weak condition for behavioral incentive compatibility.

No Information. Consistent with greater movement to the center the test shows that both quartiles are significantly closer to the center in the Information treatment than in the No-Information treatment ($p = 0.007$ and $p = 0.031$, respectively).

[43] Our data show that information increases false (distant) reports both for prior and for (the more challenging) posterior reports. Further, as seen in Figure 7, the change in the exact-center reports between the Information-No-Information treatments are very similar for the prior and posterior reports. Thus, information on the BSR results in center-biased reporting—whether participants are asked to report on the (easy) prior or on the (harder) posterior.

We use this section to first discuss the potential inferential impact of such a failure, before moving on to a broader examination of the implications for belief elicitation.

A. *Impact of Center-Biased Reporting*

Clarity of measurement focused our assessment of truth telling in the BSR on elicited reports over an objective induced prior. However, existing BSR studies mostly focus on eliciting subjective beliefs, attempting to measure a private perception where the analyst has no guidance on what the true belief might be. Should we expect our findings over the objective priors to translate to this growing number of studies using the BSR, and if so, what might the inferential impact be for those studies?

To explore these questions, we first assess whether our implementation of the elicitation mirrors those used in existing studies. Are scholars, as in our study, eliciting the likelihood of events and providing information on the quantitative incentives? Table 3 reports by study the information participants were provided, and whether an event likelihood was elicited. We find that most studies provide detailed information on the incentives, with many using two or more quantitative examples and thereby demonstrating the trade-offs associated with center-biased reporting.

While the types of elicitations and information provided in existing BSR-founded studies suggest the potential for center-biased reporting, the impact on the inferences drawn will depend on how the elicited beliefs are used. In particular, we distinguish in Table 3 between elicited beliefs used as (i) a left-hand-side dependent variable and (ii) a right-hand-side independent control variable. When used as a left-hand-side variable, the effects of center bias are predictable, attenuating any estimated treatment response. In contrast, when used as a right-hand-side variable, the inferential distortions depend on the precise specification and on the relationship between beliefs and the other variables, with the bias potentially magnifying or reducing the estimated effects (see online Appendix B.1).

To demonstrate the magnitude of the potential inferential distortions, we replicate an experimental study where elicited beliefs play a core role for inference. Specifically, we examine the Niederle and Vesterlund (2007, or NV) study of gender and competition. We mirror the study's examination of an individual's decision to perform under a noncompetitive piece rate or a competitive tournament. However, we replace the belief elicitation at the end of the study with one of two treatments: a BSR elicitation with information on the quantitative incentives, and a BSR elicitation without that information—mirroring our Information-No-Information design.

The original NV study finds that men more than women prefer to perform under a competitive tournament compensation, rather than under the noncompetitive piece rate, and that gender differences in confidence help explain this gender gap in competition. Our two treatments, which we label *NV Information* and *NV No Information*, therefore serve as a testbed to evaluate first whether there is evidence of center-bias when eliciting subjective beliefs in the NV-Information treatment, and second the potential inferential impact of using the elicited beliefs for analysis. Importantly, we can do this both when the beliefs are used as a dependent measure (evaluating the gender gap in confidence) and as a control (evaluating the gender gap in competition when controlling for confidence).

Table 3—Papers Using the BSR

Paper	Full instructions	Likelihood	Dominant	Description	Quantitative information	2+ quant. example	Inference
Published							
Hossain and Okui (2013)		✓		✓	✓	✓	*LHS*
Babcock et al. (2017)	✓	✓	✓	✓	✓	✓	*LHS*
Hillenbrand and Schmelzer (2017)	✓	✓	✓	✓	✓		Both
Drerup et al. (2017)		✓	✓	✓	✓	✓	*RHS*
Masiliūnas (2017)	✓	✓			✓	✓	*LHS*
Castillo et al. (2019)	✓	✓	✓	✓	✓	✓	
Corazzini et al. (2019)	✓	✓	✓	✓	✓	✓	*LHS*
Dargnies et al. (2019)	✓		✓	✓	✓		*RHS*
Erkal et al. (2020)	✓	✓		✓	✓	✓	*LHS*
Sonsino et al. (2020)	✓	✓	✓	✓	✓	✓	*RHS*
Charness et al. (2021)	✓	✓	✓	✓			*LHS*
Rafkin et al. (2021)			✓	–	✓	–	*LHS*
Chen and He (2022)	✓		✓	✓	✓		*RHS*
Oprea and Yuksel (2022)	✓	✓	✓	✓			*LHS*
Working Papers							
Hossain and Okui (2019)	✓		✓	✓			Both
Ahrens and Bosch-Rosa (2019)	✓						*LHS*
Dianat et al. (2022)		✓	✓	✓	✓	✓	*LHS*
Filippin and Mantovani (2019)		✓					*RHS*
Renes and Visser (2019)	✓	✓	✓	✓	✓	✓	*LHS*
Choi et al. (2020)	✓	✓		✓	✓	✓	*LHS*
Colzani and Santos-Pinto (2020)	✓		✓	✓	✓	✓	*LHS*
Dustan et al. (2022)	✓		✓	✓	✓	✓	*LHS*
Enke et al. (2020)	✓		✓	✓			Both
Koutout (2020)	✓		✓	✓	✓	✓	*RHS*
Meloso et al. (2020)	✓	✓	✓	✓	✓	✓	*LHS*
Aksoy et al. (2021)	✓	✓	✓	✓			*RHS*
Enke and Graeber (2019)	✓	✓	✓	✓			*LHS*
Erkal et al. (2022)	✓	✓	✓	✓	✓	✓	*LHS*
Aoyagi et al. (2021)	✓	✓	✓	✓			*LHS*
Graeber (2021)	✓	✓	✓	✓	✓	✓	*LHS*
Zimpelmann (2021)		✓	–	–	–	–	*RHS*
Totals	81%	71%	79%	90%	70%	62%	

Notes: Full instructions—instructions available; *Likelihood*—elicit likelihood of an event occurring; *Dominant*—participants given information to reveal that truthful revelation is a dominant strategy; *Description*—payoffs described; *Quantitative information*—participants provided with some quantitative information on the incentives; *2+ quant. example*—participants provided with two or more quantitative information on incentives; *Inference*—whether for inference the elicited belief is used as a left- or right-hand-side (LHS or RHS) variable.

Design Details, and Elicited Beliefs.—Participants in our online replication of the NV study were given two minutes to correctly add as many sets of two two-digit numbers as possible under three different performance incentives: (task 1) a \$0.50 piece rate, (task 2) a four-person tournament with \$2.00 per problem solved for the winner, and (task 3) under the participant's preferred payment scheme (piece rate or tournament).[44] Beliefs on relative performance were then elicited at the end of the study using the BSR, with all participants given the qualitative dominance statement that the chance of winning a \$4 prize is maximized by accurately reporting the likelihood of being ranked first, second, third, and fourth in their group of four. Further, participants in the NV-Information treatment were given the precise conditional

[44] Standard in-person lab procedures were modified to an online format that closely mirrored those of the lab (the experimenter and other participants were visible, instructions read out loud, clarifying questions addressed in real time; see Danz et al. 2021 for the complete protocol). Participants who entered the task 3 tournament were identified as winners if their performance in task 3 exceeded that of the other group members in the task 2 tournament. Participants also faced a task 4 decision where they decided whether to submit their task 1 performance to a piece rate or a tournament.

likelihoods of winning the $4 prize associated with any prospective report, corresponding to our Information treatment (see instructions in online Appendix C.3).

With the exception of the belief elicitation, our two NV treatments mirror the NV (2007) experimental design. In the original NV study participants were instead rewarded a $1 prize for correctly guessing their rank in their group of four (first, second, third, or fourth). By eliciting the modal rank the NV (2007) elicitation may be seen as less informative on the distribution of beliefs; however the coarse elicitation holds advantages that nonetheless may improve reporting over a BSR-Information treatment. The simple elicitation is more natural, and compared to the BSR the incentives are easier to explain (see also Abeler, Nosenzo, and Raymond 2019) and do not provide hedging opportunities (see Table 1). Further, inference on incentive compatibility does not hinge on the participant's ability to reduce a compound lottery. Thus, we anticipated that the coarser NV (2007) elicitation would better reflect the participants' true subjective priors, and that inference would be similar to those in the NV-No-Information treatment, while distinct from the NV-Information treatment. Indeed, as we show below, this is precisely what we find. The NV (2007) findings are qualitatively similar to the NV-No-Information treatment, while more centered reports in the NV-Information treatment result in qualitatively different inference.

Looking at results in the NV-Information and NV-No-Information treatments, and ignoring beliefs, both replications match the key result of the original study, that conditional on performance, men more than women enter the tournament. Our focus though is on how the elicited beliefs on relative rank move across treatment and how this affects inference.

Focusing first on beliefs, Figure 8 reports the average weight attached to each rank by treatment and gender. Panel A shows the beliefs elicited in the NV-No-Information treatment and reveals, as in NV, that men are more confident than women, reporting a substantially greater likelihood of being ranked first and a corresponding lower likelihood of being ranked third and fourth.

Our results on eliciting objective priors revealed that information on incentives biased reports on noncentered, but not on centered priors.[45] With men's beliefs less centered than women's in the NV-No-Information treatment we therefore expect that quantitative information on incentives will lead to more center-biased reporting for men than for women. Panel B confirms this expectation in the data. Relative to NV No Information the reports by men pull-to-center, leading to a much smaller gender gap in confidence for NV Information.

Inferential Impact of Center Bias.—Next, we explore the impact of using center-biased belief elicitations for inference. Seeing our NV-No-Information sample as capturing the true effect, we can predict and sign the theoretical consequences of center bias in the two inferential exercises from the NV study: first, when the

[45] In explaining the substantial change in reports by men, we also find in our elicitation of objective priors a greater center bias for priors further from center (a response that is also consistent with the intuition provided for intensified hedging in Table 1). Using a simple center-biased model, where subject i's report q_i is given by $q_i = (1 - \alpha) \cdot \pi_i + \alpha \cdot c$, which is an α-weighted average of the centered belief c and her true belief π_i, we estimate significantly greater values of α on priors of 0.2 and 0.8 than on 0.3 and 0.7 ($\hat{\alpha} = 0.289$ vs. $\hat{\alpha} = 0.149$, respectively, $p = 0.011$).

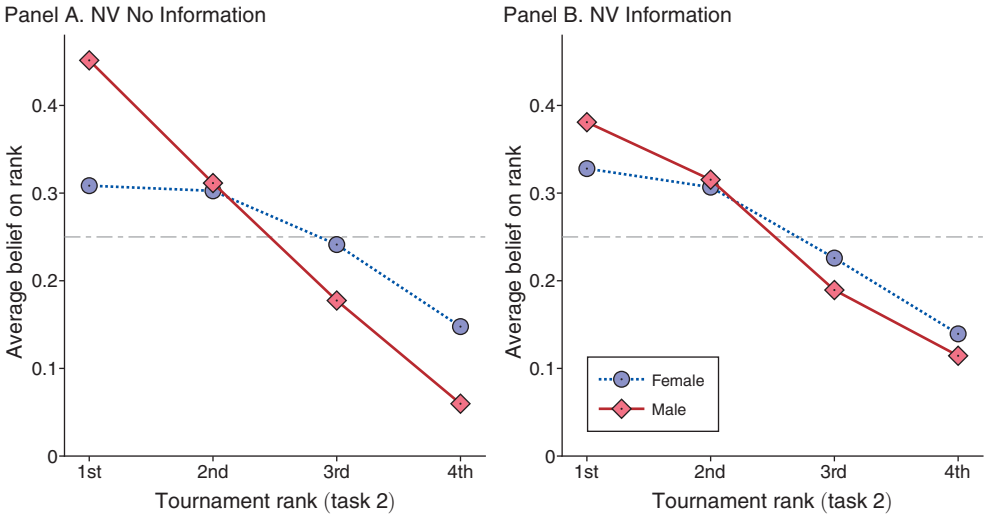


FIGURE 8. ELICITED LIKELIHOOD OF PERFORMANCE RANK IN TOURNAMENT

elicited belief q_i is used as a left-hand-side dependent variable, where we are looking for a difference in means across gender; second as a right-hand-side independent control variable, where we are looking to draw inference on how the participant's tournament entry decision y_i differs between men and women, after controlling for confidence differences via the elicited beliefs. Each exercise seeks to make inference over a difference between women and men, δ, and our theoretical analysis examines how center-biased reporting might distort inference over δ in the NV-Information treatment.

The first key inferential regression in the NV study uses measured beliefs to examine differences in confidence between men and women:

$$q_i = \mu_q + \delta_q \cdot \mathit{Female}_i + \epsilon_i, \tag{1}$$

where the estimated gender-gap in confidence $\hat{\delta}_q$ is the focus.[46] The second inferential exercise examines the tournament-entry decision y_i, conditional on confidence, where the modeled regression is

$$y_i = \mu_y + \delta_y \cdot \mathit{Female}_i + \beta_q \cdot q_i + \nu_i. \tag{2}$$

Here the focus is the estimated gender gap in tournament entry, $\hat{\delta}_y$, after controlling for the estimated confidence effect on entry (the $\hat{\beta}_q \cdot q_i$ term).

To predict the inferential distortions we use a simple model of center bias, where the elicited belief q_i is modeled as a random variable: with probability α the belief is the center value c (a constant) and with probability $1 - \alpha$ we observe the true belief $q_i^\star$. Using this model of center bias, we can predict how inference over $\hat{\delta}_q$ and $\hat{\delta}_y$ will be affected. To facilitate the prediction, we make the following simplifying

[46] While the actual regressions controls for ability, for simplicity of presentation we focus on the simplest specification to outline the main effect.

assumptions: (i) the sample is balanced with N men and N women; and (ii) the econometric errors ϵ_i and ν_i are independent mean zero errors, with finite variance (given by σ_ϵ^2 and σ_ν^2, respectively).

What then is the effect on inference for center-biased distortions in the beliefs? We here focus on statements of the results, where more detailed derivations are in online Appendix B.1. When there is no center bias ($\alpha = 0$) the OLS coefficients are unbiased, consistent estimators of the true gender effects δ_q and δ_y; however, both estimators are biased when measured beliefs are center biased ($\alpha > 0$).

When beliefs are the dependent variable in (1), center bias moves average beliefs for both men and women to the same point c. As such, the expected *difference* between the two populations is directly attenuated to $(1 - \alpha) \cdot \delta_q$,[47] leading to an underestimate of the true magnitude for the gender gap under center bias.[48]

When center-biased beliefs are instead used as an independent variable to control for the belief in (2), the asymptotic biases of the OLS estimator are again increasing in the intensity of center bias α (though here in a nonlinear manner). However, the bias size and direction also depend on the true population parameters that we are attempting to estimate. In particular, for estimation equation (2) we show that the asymptotic bias is proportional to $\beta_q \cdot \delta_q$, the product of the true marginal effect of confidence on tournament entry and the true gender difference in confidence.[49] If as in the original NV study men are more overconfident than women (so $\delta_q < 0$) and confidence has a positive effect on tournament entry ($\beta_q > 0$), then center-biased beliefs are predicted to generate a more negative estimate for the tournament entry differences between women and men after controlling for beliefs. That is, we predict that $\hat{\delta}_y < \delta_y < 0$. Center-biased beliefs would therefore *overestimate* the size of the gender gap in tournament entry after controlling for beliefs.

Inference from the NV-No-Information and NV-Information treatments, provided side by side in Table 4, are fully consistent with the predicted effects from center bias. Examining first the case where beliefs are used as the dependent variable in columns 1 and 2, we find that whether or not information is provided on the BSR incentives has a clear qualitative effect on the inference made. The results from the NV-No-Information treatment replicate the original NV finding that men conditional on performance are more confident than women (a 15 percentage-point gap in the believed likelihood of winning the tournament, $p = 0.005$). In contrast, there is no significant gender gap in confidence for NV Information (a 4 percentage-point gap, $p = 0.523$).

Next, exploring the consequences of using center-biased beliefs as an independent variable we examine the gender gap in tournament entry after controlling for confidence. Again, the results of the NV-No-Information study mirror those of the original. Conditional on performance, men more than women enter the tournament

[47] The predicted effect of attenuation at rate $1 - \alpha$ would be identical if we were trying to estimate the marginal effect on beliefs for a continuous right-hand-side treatment.

[48] Where the coefficient is attenuated in proportion to $(1 - \alpha)$, the T-statistic on the $\hat{\delta}_q$ coefficient is more complicated as the variance can be bigger or smaller with the center-biased reports, depending on the location of the center point c relative to the true belief means μ_q and $\mu_q + \delta_q$. However, the variance effect can be bounded, so that the effective t-statistic on the difference in means variable is attenuated by at least $\sqrt{1 - \alpha}$ under the model of distortion.

[49] In general, when beliefs are on the right-hand side, asymptotic biases will depend on the covariance between the unobserved belief mismeasurement, the other variables, and the true parameters.

Table 4—Gender Differences in Confidence and Tournament Entry: NV Replication Results with and without Information

	Dependent variable (cf. Table V in NV 2007)		Independent variable (cf. Table II and VI in NV 2007)			
	Belief on first rank (OLS)		Tournament entry (Probit)			
	No Information	Information	No Information		Information	
	(1)	(2)	(3)	(4)	(5)	(6)
Female	–0.148	–0.038	–0.258	–0.146	–0.357	–0.382
	(0.051)	(0.059)	(0.080)	(0.115)	(0.118)	(0.124)
Tournament	0.008	0.018	0.022	0.017	0.006	–0.011
	(0.004)	(0.005)	(0.012)	(0.014)	(0.011)	(0.014)
Tournament—piece rate	0.017	–0.021	0.001	–0.015	–0.023	–0.004
	(0.008)	(0.010)	(0.017)	(0.021)	(0.021)	(0.022)
Constant	0.305	0.059				
	(0.097)	(0.101)				
Belief weight on first rank				1.275		0.994
				(0.432)		(0.329)
Observations	74	68	74	74	68	68
R^2/adj. R^2	0.273	0.187	0.157	0.303	0.093	0.208

Note: Columns 1–2: Tobit regressions yield the same qualitative results. Columns 3–6: marginal effects (estimated constants omitted). Regressions including the average reported rank instead of the reported weight on rank 1 yield the same qualitative results.

(column 3), with the gender gap in tournament entry being partly explained by the gender gap in confidence (column 4). Participants that think they will win the tournament are more likely to enter it, and so controlling for beliefs substantially reduces the gender gap in entry. The results for the same estimation procedure in our NV-Information sample (columns 5 and 6) lead to a very different conclusion.[50] While conditional on performance men are more likely to select the tournament (column 5), controlling for beliefs does not reduce the gap in tournament entry (column 6).[51]

Our experimental results match what our theory predicts, that center-biased reporting in NV Information can lead to the false inference that there is no gender gap in confidence and that confidence plays no role in explaining the gender gap in competition.[52]

[50] Fully replicating the NV (2007) design with the original coarse belief elicitation a recent study by Recalde and Vesterlund (2022), with 263 participants at the University of Pittsburgh, finds that women were 25.2 percent less likely than men to report a belief that they were first in their group (corresponding to Table 4 columns 1 and 2). Further, an initial gender gap in tournament entry of 27.8 percent was shown to decrease to 18.9 percent when controlling for the coarse belief measure of being ranked first (corresponding to Table 4 columns 3 and 4 versus 5 and 6). All gender differences were significant at the 0.001 level.

[51] To secure no prior exposure to similar studies we conducted the study with 147 students from the University of California, Santa Barbara. Adding an "other" option to the demographic question on gender we have five students in the NV-No-Information treatment who do not identify as male or female. Mirroring the NV analysis we examine responses only for the 142 participants who identify as male or female. Including all participants, and assessing the difference between the either male/not male or not female/female yields the same effects for both Information and No Information. While not female/female yields the same qualitative effects as those shown in Table columns 1 and 4, the not-male coefficient is significant in the equivalent of column 4 ($p = 0.041$).

[52] Considering the broader impact of the NV study, this false inference could well impact policies aimed to address gender differences in advancement (e.g., the study was in a submission to the Senate's Employment, Workplace Relations and Education Legislation Committee's *Inquiry into the Workplace Relations Amendment* (*WorkChoices*) *Bill 2005* in Australia: http://www.hreoc.gov.au/legal/submissions/workplace_relations_amendment_2005.html.

Further, the results from the NV-Information treatment mirror the comparative statics when we simulate center-biased beliefs using the data from the NV-No-Information treatment (distorting each belief q_i in the data with probability α to the center value c). The simulations also indicate a smaller gender gap in confidence and an exaggerated assessment of the gender gap in tournament entry after controlling for the beliefs. By contrast, attempts to reconstruct the NV-No-Information results using the NV-Information data are not successful. Once the center bias masks the gender gap in confidence, it is not easily uncovered (see online Appendix B.2. for details).[53]

Our Information-No-Information replication of the NV study confirms that information on incentives leads to centered reporting, here over subjective beliefs. Moreover, the direct comparison demonstrates the substantial impact of using biased beliefs for inference, with center-biased reporting causing us to underestimate the gender gap in confidence and overestimate the gender gap in preferences for competition.

B. *Implication for Belief Elicitations*

Our NV replication demonstrates the serious inferential impact of using an elicitation that fails a very weak condition for behavioral incentive compatibility. What is less clear is how scholars should respond to the finding. If incentives lead to distorted reports under this state-of-the-art belief elicitation, then how do we elicit beliefs?

Needless to say, it is problematic to use any elicitation where information on the incentives underlying that mechanism increases the rate of false reports. But perhaps information only distorts reports under BSR? To demonstrate the diagnostic merit of our Information-No-Information comparison we apply it to an examination of the QSR.[54] The results mirror those from the BSR. Information on the QSR incentives increases the rate of false reports—persistently over time and particularly for non-centered priors (39.2 percent false reports in *QSR Information* versus 25.6 percent in *QSR No Information;* see Figure A.5 in the online Appendix). The QSR's failure reflects the more widely accepted idea that this elicitation is not incentive compatible for risk-averse agents. However, it also corroborates the diagnostic power of our Information-No-Information treatment comparison in determining whether an elicitation violates weak conditions over behavioral incentive compatibility.[55]

When considering improved elicitations, the finding that information on the incentives distort reports, tempts the somewhat perverse option of reducing the information provided. One option here is to simply rely on the (truthful) qualitative statement that accurate reporting maximizes the chance of winning. Certainly, the data from our No-Information treatment where only this statement is given, shows that

[53] Assessment of potential inferential bias in prior work therefore calls for Information-No-Information replications of the study of interest.

[54] The experiments were conducted online using 60 University of Pittsburgh participants in each treatment using our online protocol.

[55] We cannot rule out that the center-biased reporting under the QSR results from risk aversion, however we see no evidence that risk aversion drives the reports. Individual false-report rates and center-biased reporting are not significantly correlated with attitudes toward risk (whether measured as the certainty equivalent of a gamble, or as a probability equivalent switch point from gamble to a certain option, all $p > 0.284$).

this is the better option in terms of the accuracy of collected belief data. However, advocating for what amounts to a fully black-box mechanism from the point of participants is jarring to the general philosophy of incentivized decision-making.

A less extreme option is to also add a description of the mechanism's implementation rule, without providing precise quantitative details on the incentives. As seen in Table 3 this procedure has been used in recent implementations.[56] We pursue this approach in what we refer to as our *Description* treatment. We augment the limited information in No Information (a statement on dominance), with the short nonquantitative description, used in the Information treatment, of how prize realizations are determined—but with no other information provided.[57] Mathematically inclined participants are thus informed on the mechanism's quantitative incentives, while the less mathematically inclined learn that a concrete procedure is used to map reported beliefs into final earnings.

Results reported at the bottom of Table 2 reveal that the Description treatment largely mirrors that of the No-Information treatment, with a moderate rate of false reports (24.5 percent). This rate is not significantly different from the false-report rate in No Information ($p = 0.610$) but is significantly lower than the Information treatment ($p = 0.004$). As in the No-Information treatment, there is no evidence that false reports pull to center ($p = 0.564$), though false-report rates are slightly, but significantly, higher for noncentered than for centered priors ($p = 0.019$).[58]

As such, the complete ambiguity over the elicitation's incentives in No Information can be relaxed without severely damaging reports. However, given the distinctly different reporting behavior when participants receive detailed information on the quantitative incentives (Information treatment), it is unlikely that participants comprehend the offered incentives in Description.[59]

[56] This approach is frequently used in other mechanisms. For example, consider the nontechnical description of how a second-price bidding rule works (and equivalently, how strategy methods like the BDM function), or how a complicated matching algorithm like top-trading cycles would be described to parents providing school-choice rankings. Our results (with a fixed mechanism) dovetail with Holt and Smith (2016) who find evidence across mechanisms for the superiority of a BDM-based crossover elicitation. Similar to our Description treatment, their crossover mechanism does not spell out the marginal effects on the probability of winning, focusing on the qualitative compatibility. In comparison, their QSR elicitation uses a table to make clear the marginal effects on the monetary prize.

[57] Participants received no quantitative information on the incentives prior or after submitting a report (so No Information on the lottery pair at each choice, and no feedback after the round). Information on the incentives was only provided at the start of the experiment, augmenting the instructions for the No-Information treatment with a compact description of the payment rule. Specifically, we make use of two uniform draws, and tell participants that they will win the \$8 if (i) the event happens and their report is greater than the smaller of the two random draws, or (ii) the event does not happen and their reported belief is lower than the larger of the two random draws. As shown in Wilson and Vespa (2017) this rule is payoff equivalent to the BSR. The experiment was conducted in person at PEEL with 60 unique participants.

[58] Figures of the rates of false reports by prior and by period are shown in Figure A.6 in the online Appendix. Intriguingly, while there is no time trend in our Information, No-Information, or RCL treatments, the rate of false reports decreases over time in the Description treatment ($p = 0.077$). This decrease in false reports is consistent with the information on incentives becoming less salient in the Description treatment, where participants only learn the incentives up front. This salience argument can be mirrored in the Feedback treatment where false reports increase as participants get repeated feedback on the incentives.

[59] 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 (77 versus 72 percent, $p = 0.532$, χ^2-test) but that there are differences in understanding how beliefs affected pay (83 versus 70 percent, $p = 0.084$), and indications of differences in self-reported inclination to report truthfully (75 versus 85 percent, $p = 0.171$). Reporting an understanding of

While we may be able to limit false reporting by presenting the barest possible information on incentives, this is hardly what we have in mind when designing incentive compatible elicitations. The intent is for the incentives to drive participants toward truth telling rather than away from it.

We see our consistent finding that information on the BSR incentives increases false reports as a violation of a very weak condition for behavioral incentive compatibility. A related but more direct assessment of the incentives evaluates another weak condition for behavioral incentive compatibility, that most participants select the outcome assumed to be maximizing under the mechanism.

Fixing again the BSR incentives as the example, we explore this condition by conducting an *Incentives-Only* treatment, where participants are given a choice over the set of lottery pairs underlying the BSR, stripped of the elicitation framing. The set of 11 lotteries mirrors those shown in Table 1, with each pair of tickets consisting of a red- and a blue-lottery ticket. One lottery ticket in the participant's chosen pair is implemented for payment, where participants are informed that the ticket-color counting for payment is determined by a random draw with a predetermined chance (20 or 30 percent) on the red-ticket lottery counting.[60] Truthful revelation under the BSR relies on participants perceiving one unique lottery pair as the maximizer for each possible probability on the elicited event.

In Figure 9 we illustrate the lottery choices in the Incentives-Only treatment. On the left of each panel we present the proportion choosing the theorized maximizer, while on the right we report the deviations (here identifying the chosen lottery with the corresponding prior belief under the BSR rather than its lottery label in the experiment). We see in violation of our weak condition for behavioral incentive compatibility that the majority of participants fail to select the lottery thought to be maximizing. In fact, the distribution shows wider dispersion than observed under the comparable prior elicitations (cf. Figure 5).[61] Common to both sets of choices though is a tendency to select lottery pairs that have similar chances of winning, reflecting the hedging motives alluded to in the introduction, and consistent with the center-biased reporting documented in our Information-No-Information comparisons.[62]

What then are the implications of these findings for belief elicitation more generally? As the state-of-the-art elicitation, the BSR conveniently allows us to truthfully state to participants that their chances of winning are maximized by truthful reporting. However, the evidence for violations of two weak conditions—that provision of quantitative information on the incentives increases false reporting, along with the finding that participants when presented with the bare incentives mostly

how beliefs affected pay may result from understanding that truthful reporting maximized the chance of winning the prize.

[60] The exact choices are made over lottery pairs A through K, with the chance of winning for lottery A being 100 percent on the red lottery ticket and 0 percent on the blue, for B it is 99 percent on red and 19 percent on blue, etc. This short study was administered to 162 participants as a module at the end of another PEEL study examining public-goods provision, similar to the implementation of Fischbacher and Follmi-Heusi (2013). One of the two decisions was selected for payment, with an $8 prize in the lottery.

[61] We inform participants when eliciting beliefs that accurate reporting maximizes the chance of winning, but in contrast participants in the Incentives-Only treatment are given no guidance on what the "right" choice is.

[62] For the Incentives-Only treatment 50.6 percent (38.3 percent) of the participants in the Pr(red) $=$ 0.20 (0.3) condition move towards the center, the illustrated choice range in Figure 9. In contrast the proportion moving toward the nearest extreme is 19.8 percent (18.5 percent).

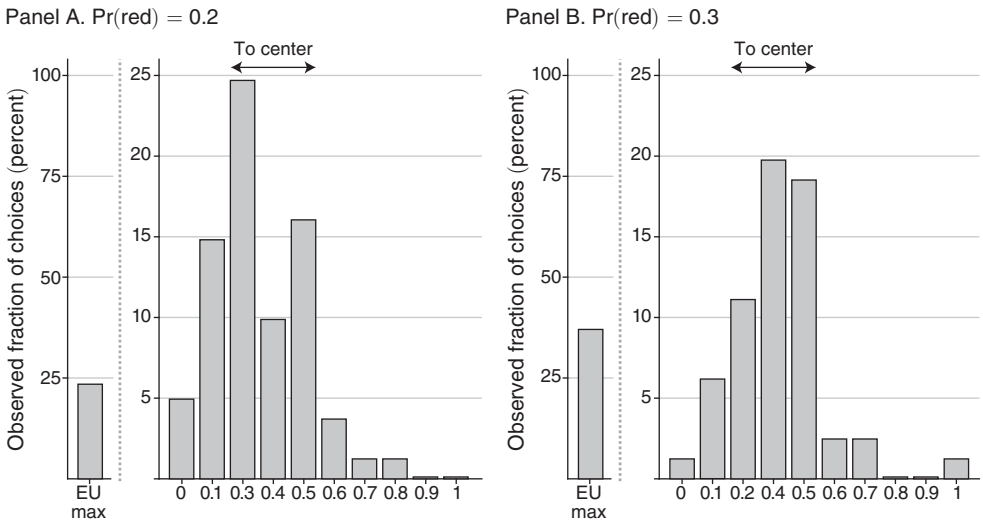


FIGURE 9. CHOSEN OUTCOMES UNDER INCENTIVES-ONLY TREATMENT

fail to select the presumed maximizing outcome—shows that the elicitation is not behaviorally incentive compatible. Adding to this concerning evidence is the clear impact that the resulting center-biased reporting has on inference.

In pursuing improved elicitations, we need to be cognizant that we are designing mechanisms for behavioral agents. In this respect, our findings and proposed tests for behavioral incentive compatibility relate to Li's (2017) concepts of obvious dominance and obvious strategy proofness. Both our work and Li (2017) stress the importance of considering cognitive limitations (in addition to a broader set of preferences) when designing incentive compatible mechanisms. However, while Li (2017) provides a theoretical criterion of a mechanism's incentive compatibility for a class of cognitively limited agents, our work stresses the importance of, and provides means to, testing whether a theoretically incentive compatible mechanism is behaviorally incentive compatible in an empirical sense. As in the BSR, relatively weak-seeming theoretical assumptions permit the design of fully separating mechanisms, to measure beliefs at arbitrary precision. But such precision may well be costly—where we need to empirically test that the assumptions put in place hold, and that behavioral agents actually perceive truthful revelation as beneficial.

Our study has proposed weak conditions for behaviorally incentive compatible elicitations and provided diagnostic tools for checking them. The hope is that new elicitations will be assessed against and succeed in passing these standards. Given the challenges associated with this task though, we caution that it may be time to question whether it is reasonable to assume that participants in our studies hold exact probabilistic beliefs, let alone our ability to use monetary incentives to elicit such beliefs at arbitrary precision. Instead of taking our results as a call for the development of mechanisms that are incentive compatible for an ever-more-general class of decision-maker, we might instead ask whether the necessary economic inferences could be drawn with less-precise measurements, where the incentives for truthful

reporting can be simpler and starker.[63] For example, in discrete settings it may be sufficient to elicit the event the participants deem most likely and incentivize the elicitation by offering compensation only in the event that the report is correct.[64] In continuous settings, the same can be achieved by paying participants if the true population outcome falls within some bounds around their guess.[65] Alternatively, it may be sufficient to determine whether a belief lies within a certain *fixed* interval. This allows for deviations between the potential intervals to come at a higher perceived cost and may still provide the information necessary for inference.[66] For example, suppose that in understanding individual behavior we wish to elicit the belief that an opponent will select action *A* or *B*, and that the individual's predicted behavior theoretically depends on the belief on *A* exceeding a 30 percent cutoff. Rather than eliciting the precise belief that action *A* is chosen, it may secure more reliable and truthful reporting to instead focus the elicitation on whether or not the belief on *A* exceeds the theoretical cutoff. If elicited beliefs are collected primarily as controls or for auxiliary tests of a behavioral mechanic, inference may be improved with starker incentives over coarser elicitations.

While there are many paths to improve belief elicitation, we propose two simple assessments: that information on the incentives increases truthful reporting, and that most participants when given a choice over the pure set of incentives select the theorized maximizer. In demonstrating the very substantial inferential consequences from using biased elicitations, our results serve as a call for elicitations to be incentive compatible both theoretically and behaviorally, but also as a strong caution against elicitations that rely on incentives that decrease truthful reporting.

REFERENCES

Abdellaoui, Mohammed, Ahmed Driouchi, and Olivier L'Haridon. 2011. "Risk Aversion Elicitation: Reconciling Tractability and Bias Minimization." *Theory and Decision* 71 (1): 63–80.

Abeler, Johannes, Danielle Nosenzo, and Collin Raymond. 2019. "Preferences for Truth-Telling." *Econometrica* 87 (4): 1115–53.

Ahrens, Steffen, and Ciril Bosch-Rosa. 2019. "The Motivated Beliefs of Investors under Limited Liability." Unpublished.

Aksoy, Billur, Ian Chadd, and Boon Han Koh. 2021. "Hidden Identity and Social Preferences: Evidence from Sexual Minorities." Unpublished.

Aoyagi, Masaki, Guillaume R. Fréchette, and Sevgi Yuksel. 2021. "Beliefs in Repeated Games." Unpublished.

[63] Ex post corrections of beliefs à la Offerman et al. (2009) similarly rely on individuals holding exact beliefs. If precise beliefs are a requirement, then dynamic mechanisms that elicit the same belief through multiple (adaptive) coarse elicitations may be used to make incentives appear stark (see Schmidt and Zankiewicz 2016). With starker incentives we mean mechanisms that provide steeper marginal incentives, especially in the neighborhood of the theorized maximizer.

[64] See, for example, Bhatt and Camerer (2005); Hurley and Shogren (2005); Niederle and Vesterlund (2007); Vanberg (2008); Blanco et al. (2010); Dargnies (2012); Di Tella et al. (2015); LeCoq, Tremewan, and Wagner (2015); Toussaert (2018); Bordalo et al. (2019); Cantoni et al. (2019); Wilcox and Feltovich (2000); and Huffman, Raymond, and Shvets (forthcoming).

[65] For example, Charness and Dufwenberg (2006); Abeler, Nosenzo, and Raymond (2019); Danz, Madarász, and Wang (2018). As Abeler, Nosenzo, and Raymond (2019, p. 1140) point out, "[t]his mechanism is very simple and easier to explain and understand than proper scoring rules. It elicits in an incentive-compatible way the mode (or more precisely, the mid-point of the [x]-percentage point interval with the highest likelihood) of a subject's distribution of estimates."

[66] Asking participants to select one of a several fixed ranges of probabilities acknowledges both the ambiguity associated with providing an exact probabilistic belief (Manski 2004) and makes it possible to provide starker incentives for doing so.

Avoyan, Ala, and Andrew Schotter. 2020. "Attention in Games: An Experimental Study." *European Economic Review* 124: 1–28.

Babcock, Linda, Maria P. Recalde, Lise Vesterlund, and Laurie Weingart. 2017. "Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability." *American Economic Review* 107 (3): 714–47.

Benjamin, Daniel. 2019. "Errors in Probabilistic Reasoning and Judgment Biases." In *Handbook of Behavioral Economics: Applications and Foundations 1*, Vol. 2, edited by Douglas Bernheim, Stefano DellaVigna, and David Laibson, 69–186. North-Holland: Elsevier.

Bhatt, Meghana, and Colin F. Camerer. 2005. "Self-Referential Thinking and Equilibrium as States of Mind in Games: fMRI Evidence." *Games and Economic Behavior* 52 (2): 424–59.

Blanco, Mariana, Dirk Engelmann, Alexander K. Koch, and Hans-Theo Normann. 2010. "Belief Elicitation in Experiments: Is There a Hedging Problem?" *Experimental Economics* 13 (4): 412–38.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Shleifer Andrei. 2019. "Beliefs about Gender." *American Economic Review* 109 (3): 739–73.

Bruner, David M. 2009. "Changing the Probability versus Changing the Reward." *Experimental Economics* 12 (4): 367–85.

Burfurd, Ingrid, and Tom Wilkening. 2018. "Experimental Guidance for Eliciting Beliefs with the Stochastic Becker–DeGroot–Marschak Mechanism." *Journal of the Economic Science Association* 4 (1): 15–28.

Cantoni, Davide, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang. 2019. "Protests as Strategic Games: Experimental Evidence from Hong Kong's Antiauthoritarian Movement." *Quarterly Journal of Economics* 134 (2): 1021–77.

Cason, Timothy N., and Charles R. Plott. 2014."Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing." *Journal of Political Economy* 122 (6): 1235–70.

Cason, Timothy N., Tridib Sharma, and Radovan Vadovič. 2020. "Correlated Beliefs: Predicting Outcomes in 2×2 Games." *Games and Economic Behavior* 122: 256–76.

Castillo, Marco E., Philip J. Cross, and Mikhail Freer. 2019. "Nonparametric Utility Theory in Strategic Settings: Revealing Preferences and Beliefs from Proposal–Response Games." *Games and Economic Behavior* 115: 60–82.

Charness, Gary, and Martin Dufwenberg. 2006. "Promises and Partnership." *Econometrica* 74 (6): 1579–1601.

Charness, Gary, Uri Gneezy, and Vlastimil Rasocha. 2021. "Experimental Methods: Eliciting Beliefs." *Journal of Economic Behavior and Organization* 189: 234–56

Charness, Gary, Ryan Oprea, and Sevgi Yuksel. 2021. "How Do People Choose between Biased Information Sources? Evidence from a Laboratory Experiment." *Journal of the European Economic Association* 19 (3): 1656–91.

Chen, Yan, and YingHua He. 2022. "Information Acquisition and Provision in School Choice: A Theoretical Investigation." *Economic Theory* 74: 293–327.

Choi, Syngjoo, Byung-Yeon Kim, Jungmin Lee, and Sokbae Lee. 2020. "Institutions, Competitiveness and Cognitive Ability." Unpublished.

Colzani, Paola, and Luis Santos-Pinto. 2020. "Does Overconfidence Lead to Bargaining Failures?" Unpublished.

Corazzini, Luca, Stefano Galavotti, and Paola Valbonesi. 2019. "An Experimental Study on Sequential Auctions with Privately Known Capacities." *Games and Economic Behavior* 117: 289–315.

Danz, David, Neeraja Gupta, Marissa Lepper, Lise Vesterlund, and K. Pun Winichakul. 2021. "Going Virtual: A Step-by-Step Guide to Taking the In-Person Experimental Lab Online." Unpublished.

Danz, David, Kristóf Madarász, and Stephanie Wang. 2018. "The Biases of Others: Projection Equilibrium in an Agency Setting," CEPR Discussion Paper 12867.

Danz, David, Lise Vesterlund, and Alistair Wilson. 2022. "Replication Data for: Belief Elicitation and Behavioral Incentive Compatibility." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E157161V1.

Dargnies, Marie-Pierre. 2012. "Men Too Sometimes Shy Away from Competition: The Case of Team Competition." *Management Science* 58 (11): 1982–2000.

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

Dean, Mark, and Pietro Ortoleva. 2017. "Allais, Ellsberg, and Preferences for Hedging." *Theoretical Economics* 12 (1): 377–424.

de Quidt, Jonathan, Lise Vesterlund, and Alistair J. Wilson. 2019. "Experimenter Demand Effects." In *Handbook of Research Methods and Applications in Experimental Economics,* edited by Arthur Schram and Aljaz Ule, 384–400. Northampton, MA: Edward Elgar Publishing.

Dianat, Ahrash, Federico Echenique, and Leeat Yariv. 2022. "Statistical Discrimination and Affirmative Action in the Lab." *Games and Economic Behavior* 132: 41–58.
Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman. 2015. "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism." *American Economic Review* 105 (11): 3416–42.
Drerup, Tilman, Benjamin Enke, and Hans-Martin Von Gaudecker. 2017. "The Precision of Subjective Data and the Explanatory Power of Economic Models." *Journal of Econometrics* 200 (2): 378–89.
Dustan, Andrew, Kristine Koutout, and Greg Leo. 2022. "Second-Order Beliefs and Gender." *Journal of Economic Behavior and Organization* 200: 752–81.
Enke, Benjamin, and Thomas Graeber. 2019. "Cognitive Uncertainty." NBER Working Paper 26518.
Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann. 2020. "Associative Memory and Belief Formation." NBER Working Paper 26664.
Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh. 2020. "Replication: Belief Elicitation with Quadratic and Binarized Scoring Rules." *Journal of Economic Psychology* 81: 102315.
Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh. 2022. "By Chance or by Choice? Biased Attribution of Others' Outcomes When Social Preferences Matter." *Experimental Economics* 25 (2): 413–43.
Filippin, Antonio, and Marco Mantovani. 2019. "Risk Aversion and Information Aggregation in Asset Markets." Unpublished.
Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10 (2): 171–78.
Fischbacher, Urs, and Franziska Föllmi-Heusi. 2013. "Lies in Disguise—An Experimental Study on Cheating." *Journal of the European Economic Association* 11 (3): 525–47.
Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives* 19 (4): 25–42.
Gächter, Simon, and Elke Renner. 2010. "The Effects of (Incentivized) Belief Elicitation in Public Goods Experiments." *Experimental Economics* 13 (3): 364–77.
Graeber, Thomas. 2021. "Inattentive Inference." Unpublished.
Hao, Li, and Daniel Houser. 2012. "Belief Elicitation in the Presence of Naïve Respondents: An Experimental Study." *Journal of Risk and Uncertainty* 44 (2): 161–80.
Harrison, Glenn W., and Richard D. Phillips. 2014. "Subjective Beliefs and Statistical Forecasts of Financial Risks: The Chief Risk Officer Project." In *Contemporary Challenges in Risk Management*, edited by Torben Juul Andersen, 163–202. New York, NY: Springer.
Healy, Paul J. 2017. "Epistemic Game Theory Experiments: Utility Elicitation and Irrational Play." Unpublished.
Healy, Paul J. 2018. "Explaining the BDM—or Any Random Binary Choice Elicitation Mechanism—to Subjects." Unpublished.
Hillenbrand, Adrian, and André Schmelzer. 2017. "Beyond Information: Disclosure, Distracted Attention, and Investor Behavior." *Journal of Behavioral and Experimental Finance* 16 (1): 14–21.
Holt, Charles A., and Angela M. Smith. 2016. "Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes." *American Economic Journal: Microeconomics* 8 (1): 110–39.
Holt, Charles A., and Angela M. Smith. 2016. "Replication data for: Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E114330V1.
Hossain, Tanjim, and Ryo Okui. 2013. "The Binarized Scoring Rule." *Review of Economic Studies* 80 (3): 984–1001.
Hossain, Tanjim, and Ryo Okui. 2019. "Belief Formation under Signal Correlation." Unpublished.
Huffman, David, Collin Raymond, and Julia Shvets. Forthcoming. "Persistent Overconfidence and Biased Memory: Evidence from Managers." *American Economic Review.*
Hurley, Terrance M., and Jason F. Shogren. 2005. "An Experimental Comparison of Induced and Elicited Beliefs." *Journal of Risk and Uncertainty* 30 (2): 169–88.
Koutout, Kristine. 2020. "Gendered Beliefs and the Job Application Decision: Evidence from a Large-Scale Field and Lab Experiment." Unpublished.
LeCoq, Chloe, James Tremewan, and Alexander K. Wagner. 2015. "On the Effects of Group Identity in Strategic Environments." *European Economic Review* 76: 239–52.
Li, Shengwu. 2017. "Obviously Strategy-Proof Mechanisms." *American Economic Review* 107 (11): 3257–87.
Manski, Charles F. 2004. "Measuring Expectations." *Econometrica* 72 (5): 1329–76.

Masiliūnas, Aidas. 2017. "Overcoming Coordination Failure in a Critical Mass Game: Strategic Motives and Action Disclosure." *Journal of Economic Behavior and Organization* 139: 214–51.

Meloso, Debrah, Salvatore Nunnari, and Marco Ottaviani. 2020. "Looking into Crystal Balls: A Laboratory Experiment on Reputational Cheap Talk." Unpublished.

Nelson, Robert G., and David A. Bessler. 1989. "Subjective Probabilities and Scoring Rules: Experimental Evidence." *American Journal of Agricultural Economics* 71 (2): 363–69.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics* 122 (3): 1067–1101.

Offerman, Theo, Joep Sonnemans, Gijs Van De Kuilen, and Peter P. Wakker. 2009. "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes." *Review of Economic Studies* 76 (4): 1461–89.

Oprea, Ryan, and Sevgi Yuksel. 2022. "Social Exchange of Motivated Beliefs." *Journal of the European Economic Association* 20 (2): 667–99.

Palfrey, Thomas R., and Stephanie W. Wang. 2009. "On Eliciting Beliefs in Strategic Games." *Journal of Economic Behavior and Organization* 71 (2): 98–109.

Rafkin, Charlie, Advik Shreekumar, and Pierre-Luc Vautrey. 2021. "When Guidance Changes: Government Stances and Public Beliefs." *Journal of Public Economics* 196: 104319.

Recalde, Maria, and Lise Vesterlund. 2022. "Gender Differences." Unpublished.

Renes, Sander, and Bauke Visser. 2019. "Markets Assessing Decision Makers and Decision Makers Impressing Markets: A Lab Experiment." Unpublished.

Roth, Alvin E., and Michael W. Malouf. 1979. "Game-Theoretic Models and the Role of Information in Bargaining." *Psychological Review* 86 (6): 574–94.

Schlag, Karl H., and Joël J. van der Weele. 2015. "A Method to Elicit Beliefs as Most Likely Intervals." *Judgment and Decision Making* 10 (5): 456–68.

Schmidt, Tobias, and Christian Zankiewicz. 2016. "Binary Choice Belief Elicitation: An Adaptively Optimal Design." Unpublished.

Schotter, Andrew, and Isabel Trevino. 2014. "Belief Elicitation in the Laboratory." *Annual Review of Economics* 6 (1): 103–28.

Sonsino, Doron, Yaron Lahav, and Amir Levkowitz. 2020. "The Conflicting Links between Forecast-Confidence and Trading Propensity." *Journal of Behavioral Finance* 22 (4): 443–60.

Toussaert Severine. 2018. "Eliciting Temptation and Self-Control through Menu Choices: A Lab Experiment." *Econometrica* 86 (3): 859–89.

Trautmann, Stefan T., and Gijs van de Kuilen. 2015. "Belief Elicitation: A Horse Race Among Truth Serums." *Economic Journal* 125 (589): 2116–35.

Vanberg, Christoph. 2008. "Why Do People Keep Their Promises? An Experimental Test of Two Explanations." *Econometrica* 76 (6): 1467–80.

Wang, Stephanie W. 2011. "Incentive Effects: The Case of Belief Elicitation from Individuals in Groups." *Economics Letters* 111 (1): 30–33.

Wilcox Nathaniel T., and Nick Feltovich. 2000. "Thinking Like a Game Theorist: Comment." Unpublished.

Wilson, Alistair J., and Emanuel Vespa. 2018. "Paired-Uniform Scoring: Implementing a Binarized Scoring Rule with Non-mathematical Language." Unpublished.

Zimpelmann, Christian. 2021. "Stock Market Beliefs and Portfolio Choice in the General Population." Unpublished.