Journal of Economic Perspectives—Volume 38, Number 4—Fall 2024—Pages 131–154

Evaluating Behavioral Incentive Compatibility: Insights from Experiments

David Danz, Lise Vesterlund, and Alistair J. Wilson

In a mechanism design framework, the economist acts as an engineer, choosing the incentives and rules of an environment to shape participants' behavior towards the designer's objective. For example, the task may be to select bidding rules to maximize revenue from bidders in an auction or to construct a matching algorithm to efficiently allocate applicants to a limited number of medical residency positions. A core challenge in designing a mechanism is that the specific outcome the designer wants to achieve depends on agents' "types," where these types are private information, unknown to the designer. An agent's type includes any information relevant to their decision and the designer's objective, such as their willingness to pay for an item in an auction or their personal rankings of medical residency programs. Successful implementation of a mechanism hinges on the ability to acquire information on types, but it may not be in the agent's interest to reveal this information. For example, a bidder in an auction may be reluctant to reveal their true willingness to pay for the item if it adversely affects the price they must pay. In designing a mechanism, the economist aims both to provide agents with incentives that make them want to reveal their type, and to implement the designer's ideal objective given their types.

In modeling this problem, Hurwicz (1972, 1973) cleverly treated the agent's decision in response to the mechanism as simultaneously revealing their type and

■ *David Danz is a Research Assistant Professor, Lise Vesterlund is the Andrew W. Mellon Professor of Economics, and Alistair J. Wilson is a Professor of Economics, all at the University of Pittsburgh, Pittsburgh, Pennsylvania. Lise Vesterlund is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are danz@pitt.edu, vester@pitt.edu, and alistair@pitt.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.38.4.131.

securing the designer's intended outcome. That is, the selected mechanism is one that addresses the designer's objective subject to the agents' *incentive compatibility constraints*, where these constraints ensure that the only valid rules and incentives set by the designer are those that make the agents prefer to reveal their type truthfully.

Under the assumptions that agents are cognitively perfect and rational and that they hold certain preferences, theoretical modeling of the incentive compatibility constraint has led to the development of countless mechanisms. However, research is showing that when human decision-makers are faced with these mechanisms, they often fail to reveal their type, suggesting that the mechanisms are not incentive compatible in a behavioral sense. Individuals faced with mechanisms that are not *behaviorally incentive compatible* will not reveal their type, leading the designer to select outcomes that differ from their objective: auction revenue not being maximized with participants underbidding, or the allocation of applicants to residency programs being inefficient (and unstable) because hospital-resident pairs want to break from the given match.[1]

In using and improving mechanisms, it is critical that we determine whether they are behaviorally incentive compatible. Although mechanisms are designed to be used in the field, it is not possible in a field setting to verify that they succeed in eliciting participants' private "types." Experimental studies allow for such verification and have served a critical role in assessing whether mechanisms are behaviorally incentive compatible. The reason is that we in an experimental study directly can induce a participant's type and observe whether the induced type is revealed under the mechanism (referred to as *truthful revelation*). While the laboratory differs from the field, the structure of the incentives is the same, and mechanisms that fail in the lab are expected to similarly fail in the field (for example, Kagel and Roth 2000; Kessler and Vesterlund 2015).

This paper will review the techniques used in experiments to assess behavioral incentive compatibility. The experimental tests discussed have been applied to a wide set of mechanisms, including auctions, centralized clearinghouses, and others. However, to demonstrate these techniques we use as a running example the conceptually simple mechanism of eliciting beliefs from individuals where the designer's objective is one of truth-telling. As an example, we may want to learn how likely people think it is that a specific event occurs, say, that the Federal Reserve decreases interest rates by 50 basis points. To achieve truth-telling, we can elicit this belief by presenting incentives that depend on the actual realization of the event and make

[1] In formal terms, consider a screening problem where we abstract from strategic interactions and try to identify an individual's private type, $\theta \in \Theta$, which captures their preference over a set of outcomes $\mathcal{A}$, where $x \succ_\theta y$ indicates a strict preference for x over y. The designer asks the individual to report a type q, and in trying to get truthful revelation, selects a direct mechanism, a rule outlining an outcome $\phi(q) \in \mathcal{A}$ for every report q. A direct mechanism ϕ is incentive compatible if $\phi(q = \theta) \succ_\theta \phi(q = \theta')$ for every possible alternative report $\theta' \neq \theta$. In a strategic mechanism, the incentive compatibility condition will be based on a truthful report being an expected best response conditional on equilibrium behavior of all other types in a Bayesian implementation; or for all possible reports by the other players in a dominant-strategy implementation.

it in the respondent's best interest to report accurately their subjective assessment over the likelihood that rates are decreased. In the case of belief elicitation, the individual's private type is the belief that they hold over the event, with the designer's objective merely being one of truth-telling. So in this case, the designer's objective and the incentive compatibility constraint coincide.

The advantage of using individual belief elicitations to demonstrate experimental tests of behavioral incentive compatibility is that we can ignore specifics of the designer's objective (which here coincides with truth-telling) and any speculation on the behavior of others (as the elicitation is an individual-decision problem, not a strategic game). As such, we can focus squarely on whether participants under the mechanism see it as in their interest to reveal an induced belief. For example, we can in an experiment directly induce a belief of say 30 percent for the participant—by rolling a ten-sided die and asking participants for reports on the likelihood that a 1, 2, or 3 will appear; or by drawing a ball from an urn with 100 balls, of which 30 are blue, and asking for a report on the likelihood that a drawn ball is blue. After inducing the given type (the belief of 30 percent), we can then assess whether a particular belief elicitation mechanism succeeds in incentivizing reports on the induced belief.

In this paper, we begin by motivating the need for incentive-compatible mechanisms to elicit beliefs. We then use belief elicitations to present the techniques used to explore truthful revelation. First, we review tests centered on evaluating behavior under the mechanism of interest. While these tests can demonstrate failure to reveal the induced type, they do not reveal whether the failure results from the mechanism's incentives or from some other aspect of the mechanism. We therefore refer to these as *indirect* assessments of behavioral incentive compatibility. Tests include evaluations within a mechanism of whether participants reveal an induced type, comparisons between mechanisms to evaluate which comes closer to truthful revelation, as well as assessments of what might cause deviations. Second, we report on more recent *direct* assessments of behavioral incentive compatibility. These assessments directly evaluate the mechanism's incentives by asking whether participants prefer the designed incentive for their type to the other alternatives, and by testing whether full and easily understood information on the incentives increases truthful revelation. Throughout the discussion, we will provide evidence suggesting that although some of the most-used belief-elicitation mechanisms are theoretically incentive compatible, because of either failed modeling of the individual's preferences or cognitive abilities they are not behaviorally incentive compatible. Indeed, the incentives used are shown to distort reports, and researchers will often fare better if instead of explaining the mechanism or the incentives to the participants, they just tell them "you will maximize your expected earnings if you give your best estimate."

Why Elicit Beliefs with Mechanisms?

Getting information on people's beliefs is important for assessing collective expectations and for understanding human behavior. In many situations, researchers will be interested in understanding the extent to which beliefs affect the choices that people make (Manski 2004). Do differences in college attendance result from differences in aptitude or from differences in the expected return from education? Do workers differ in their propensity to apply for promotion because of differences in risk aversion or because of differences in perceptions of how talented they think they are? Is the fact that some people have a greater reluctance to bargain driven by a concern for their counterpart, by the belief that bargaining will result in backlash, or by a belief that they are "not good" at it? Assessing and controlling for beliefs helps us understand behavior and formulate effective policy interventions.

In these and other settings, why not just ask people about their beliefs? Indeed, surveys about beliefs are a common technique used by social scientists. For example, participants could be asked in a survey to report whether they agree with the statement that their relative performance on a test will be in the top quarter of their cohort, perhaps using a five-point scale ranging from "strongly disagree" to "strongly agree." While easy to understand, the reports given may mean different things to different people. One person's "*disagree*" could be another's "*strongly disagree*." To put things on a common scale, we may instead ask participants to report the likelihood that they are ranked in the top quarter of the performance distribution.

But while we might fine-tune the questions we ask, it is harder to encourage the honest and reflective answers we are hoping for. Participants may have a sense that it is likely that they are in the top performance quarter, but find it is difficult to determine how likely. It takes effort to provide a probabilistic assessment of an event occurring: effort to understand the question through this quantitative lens, and perhaps effort to not brag and tell others that you are certain you are in the top quarter (Ewers and Zimmerman 2015), or to be humble and report that you are unlikely to be top-ranked (Thoma 2016).

To encourage truthful reporting, economists have resorted to paying participants. These payments differ from common flat-fee payments for completing a survey because the aim is not one of compensating for time spent, but instead to provide incentives for accuracy of the provided information. Economists have focused on mechanisms that present participants with incentives that make it in their interest to report their beliefs truthfully.[2] An incentive-compatible belief elicitation is structured to reward consideration, to increase accuracy, and to reduce noise in the response.

To see how the incentives selected for a mechanism can achieve this goal, consider the case of the "quadratic-scoring rule" (Brier 1950), one of the earliest deployed elicitation mechanisms (initially developed to assess the accuracy of weather forecasts). Suppose we want to elicit an individual's probabilistic belief $q \in [0,1]$ over a binary event E (say, being in the top performance quarter). Under the quadratic-scoring rule, the individual's monetary reward $\pi(q) \in [\$0, \$X]$ depends

[2] Incentive compatible rules have been shown to outperform incompatible ones (Nelson and Bessler 1989; Palfrey and Wang 2009; Schotter and Trevino 2014) and these in turn dominate unincentivized elicitation (Gächter and Renner 2010; Wang 2011; Trautmann and van de Kuilen 2015).

on their stated belief q, and on the (squared) prediction error based on the realized event E:

$$\pi(q) \;=\; \begin{cases} \$X \cdot \left[1 - (1 - q)^2\right], & \text{if event } E \text{ occurs,} \\ \$X \cdot \left[1 - q^2\right], & \text{otherwise.} \end{cases}$$

As a numerical example, suppose that someone believes they have an 80 percent chance of scoring in the top quarter on a test. If they report a belief of $q = 0.8$ their payoff will be:

$$\pi(q) \;=\; \begin{cases} \$10 \cdot \left[1 - (1 - 0.8)^2\right], & \text{if event } E \text{ occurs,} \\ \$10 \cdot \left[1 - 0.8^2\right], & \text{otherwise.} \end{cases}$$

That is, this person receives \$9.60 if they actually are in the top quarter, but only \$3.60 if they are not. Given the true belief that there is an 80 percent chance of being in the top quarter, the person expects an 80 percent chance of the high payment and a 20 percent chance of the low payment, yielding an expected payoff of reporting $q = 0.8$ of $0.8(\$9.60) + 0.2(\$3.60) = \$7.68 + \$0.72 = \$8.40$.

Central to the quadratic-scoring rule is that participants who maximize expected payoffs have an incentive to report their prediction accurately. For example, suppose that instead of reporting their true belief of $\theta = 0.8$, they report $q = 0.6$ on being in the top quarter. Given the incentives under the quadratic-scoring rule, a reported belief of 0.6 leads to a payoff of \$8.40 if they are in the top quarter and \$6.40 if they are not. While the participant may report any q they wish, their actual belief of $\theta = 0.8$ that they are in the top quarter is fixed, and so their expected payoff of making this incorrect prediction is $0.8(\$8.40) + 0.2(\$6.40) = \$6.72 + \$1.28 = \$8.00$. As a result, their expected payoff is lower under a report of 0.6 than if they had reported their true belief of 0.8. Reporting a higher belief of say $q = 1.00$ is also disadvantageous. Here the payoff would be \$10 when in the top quarter and \$0 when not in the top quarter, and so the expected payoff is $0.8(\$10) + 0.2(\$0) = \$8.00$, again lower than reporting the actual belief.

As this example illustrates, individuals who want to maximize their expected earnings will prefer to report their true belief θ, because any other report lowers their expected earnings.[3] To put it another way, participants of type θ prefer the incentives meant for them, over those intended for other types.

While the quadratic-scoring rule is theoretically incentive compatible for agents aiming to maximize their *expected* earnings, truthful revelation depends on

[3] More generally, given an actual belief of θ that E occurs, the participant's expected payoff when reporting q is given by:

$$E_\theta \pi(q) \;=\; \$X \cdot \left[\theta \cdot \left[1 - (1 - q)^2\right] + (1 - \theta) \cdot \left(1 - q^2\right)\right].$$

By deriving a first- and second-order condition over the reported value q, we confirm that the unique maximizer is to truthfully report $q^\star(\theta) = \theta$.

how individuals respond to the presented incentives. They may make mistakes when attempting to calculate their expected earnings or apply behavioral rules-of-thumb when faced with such problems, or they may not make choices to maximize their expected earnings. For example, individuals who are risk averse over the stakes will be drawn to report a more conservative belief, closer to the center $\left(q = 1/2\right)$, to get payoffs that vary less with the realized event. Indeed, concerns that the quadratic-scoring rule is not incentive compatible for risk-averse individuals (Winkler and Murphy 1970), and experimental evidence that it may not be behaviorally incentive compatible (for example, Offerman et al. 2009), has led to the development of belief elicitations that are incentive compatible for arbitrary risk preferences (for example, Hossain and Okui 2013; Mobius et al. 2022). Next, we discuss the experimental techniques that have been used to assess whether a mechanism is behaviorally incentive compatible.

Indirect Assessments of Behavioral Incentive Compatibility

We begin by reviewing the experimental tests that assess truthful reporting under the mechanism. That is, we provide participants with information on the likelihood of an event to induce the participants' belief θ that the event occurs and assess if, when presented with the incentives under a mechanism, reports on their type, q, correspond to the induced type, θ.

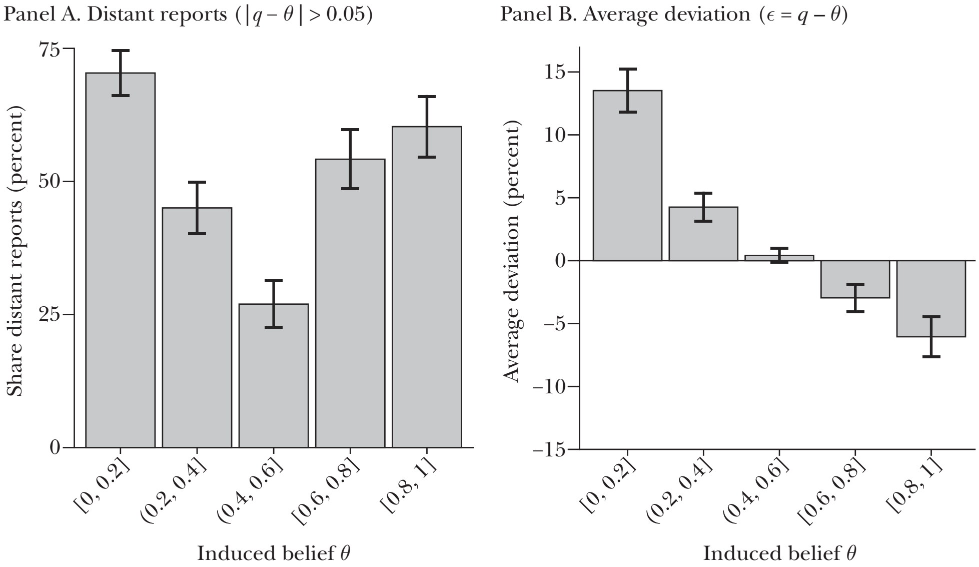
While informative on truthful revelation under the mechanism, these tests do not isolate the effect of incentives from a particular mechanism or directly evaluate preferences over the incentives within that mechanism. Hence, we refer to these tests as *indirect* assessments of behavioral incentive compatibility. They include performance evaluations within a particular mechanism and across mechanisms to determine which comes closer to truthful revelation, as well as assessments of what might cause deviations.

For the purposes of this paper, we will focus on the elicitation of simple induced beliefs, where probabilities are straightforward to see and can be understood with virtually no computational effort, like probabilities based on rolling a die or drawing from an urn. There are of course many studies that compare belief elicitations when induced beliefs require greater computation (for a review, see Benjamin 2019), or when subjective beliefs are elicited (over the behavior of other players in a strategic game, as in Nyarko and Schotter 2002). Our focus on simpler settings rules out confusion that could arise from determining the induced belief and makes it easy to determine whether the reported beliefs differ from the induced ones. If we find that a mechanism fails in a simple setting, we should not expect it to fare better when eliciting more complex beliefs.

Assessing Truthful Revelation Within and Across Mechanisms

Behavioral incentive compatibility is often assessed within a mechanism by simply checking how often reports under the mechanism correspond to the

Figure 1
Reporting Behavior in the Quadratic Scoring Rule



Source: Figures based on the published data from elicitations using the quadratic scoring rule in Offerman et al. (2009); Hossain and Okui (2013); Erkal, Gangadharan, and Koh (2020); and Danz, Vesterlund, and Wilson (2022). Total sample size is 426 participants and 3,213 total decisions.
Note: The figure shows the fraction of distant reports (panel A) and the direction of deviations (panel B) by induced belief (binned into intervals).

induced belief. To demonstrate, we report on studies examining reports under the quadratic-scoring rule, pooling more than 3,000 decisions from Offerman et al. (2009), Hossain and Okui (2013), Erkal, Gangadharan, and Koh (2020), and Danz, Vesterlund, and Wilson (2022).

In panel A of Figure 1, we show by ranges of the induced belief, θ, the fraction of reports, q, that were more than 5 percentage points from θ. We refer to these as "distant reports." For example, the first bar shows that when the induced belief is a number in the range of 0 to 0.2, a full 70 percent of reports deviated by more than 5 percentage points from the induced belief. Across all induced beliefs, 49 percent of reports deviated by more than 5 percentage points and only 43 percent of reports were exactly equal to the induced belief. Furthermore, we see a systematic decrease in the frequency of distant reports when the induced belief is closer to the center, with it being smallest in the center range from 0.4 to 0.6. For noncentered induced beliefs (outside of the 0.4 to 0.6 range) the majority of distant reports pull toward the center and 10 percent claim an exactly centered belief of $q = 0.5$. Evidence of center-biased reporting is also seen in panel B of Figure 1 where the average deviation from the induced belief tends to be positive when the induced belief is less than one-half and negative when the induced belief is more than one-half. In addition, the average size of the deviation is largest when the induced belief is large or small, because centered reports are farther from these values.

As an assessment of the performance of the quadratic-scoring rule, panels A and B of Figure 1 demonstrate that participants within the mechanism largely fail to report the induced belief. That is, the mechanism does not appear to be behaviorally incentive compatible. Particularly concerning is that deviations from the induced beliefs are large and systematic. Econometrically, center-biased reporting will bias the underlying estimates if we use the reported beliefs q in place of the true beliefs θ, as either an explained or explanatory variable in a regression.

Another popular experimental technique for assessing behavioral incentive compatibility is to compare the performance of different mechanisms to determine which comes closer to truthful revelation. For example, this "horserace" methodology has been used by Hossain and Okui (2013), Erkal, Gangadharan, and Koh (2020), and Danz, Vesterlund, and Wilson (2022) to compare the classic quadratic-scoring rule and the binarized version of the quadratic-scoring rule. The latter was designed to be incentive compatible for individuals irrespective of their risk preferences, and thus address the concern that risk aversion may cause center-biased reporting (Hossain and Okui 2013).

In both the classic and the binarized quadratic-scoring rule, the participant's payment depends on their (squared) prediction error. While payment is decreasing in the prediction error for the classic quadratic-scoring rule, payment for the binarized-scoring rule is a percentage chance of winning a fixed monetary prize, say \$10, and a larger prediction error instead decreases the chance that the participant wins the prize. Specifically, participants under the binarized-scoring rule are incentivized by a state-contingent lottery pair, where a reported belief of q on a binary event E is compensated with a $1 - (1 - q)^2$ chance of winning \$10 if the event occurs; and a $1 - q^2$ chance of winning \$10 if the event does not occur. Thus, if a participant believes and reports that there is an 80 percent chance of an event happening, then the chance of winning \$10 is 96 percent if the event occurs and 36 percent if the event does not occur, where the chance of winning the prize is maximized when the true belief is reported. While the classic quadratic-scoring rule is theoretically incentive compatible for risk-neutral individuals, the binarized version of the scoring rule is theoretically incentive compatible for arbitrary risk preferences.

A common measure of performance success used in horseraces between mechanisms is the square root of the sum of the squares of the deviations between the reported belief and the belief induced by the researcher (specifically, the root-mean-squared-deviation is RMSD $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(q_i - \theta_i)^2}$). Pooling results from Hossain and Okui (2013), Erkal, Gangadharan, and Koh (2020), and Danz, Vesterlund, and Wilson (2022), we can compare the root-mean-squared-deviation under the classic and binarized quadratic-scoring rule.

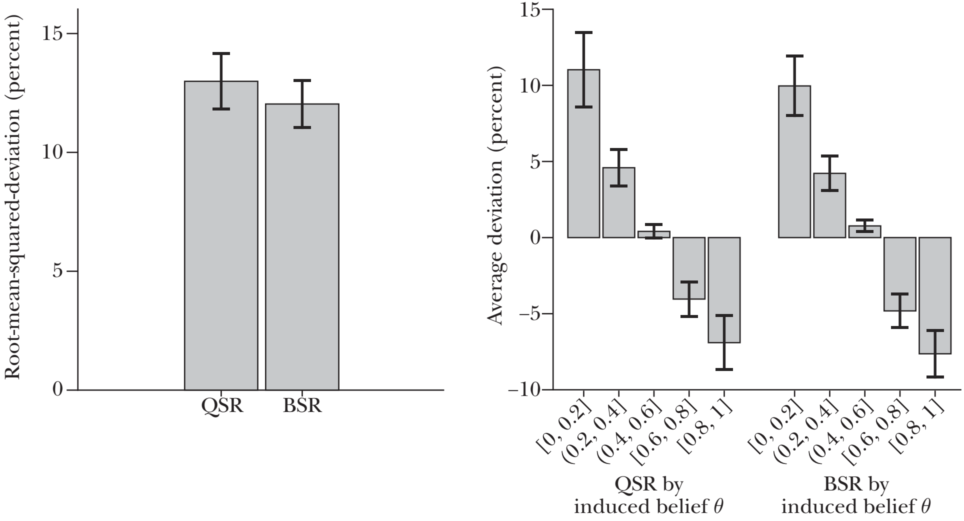
The results for the pooled data are shown in panel A of Figure 2. Revealing that while there is substantial deviation from the induced belief under both elicitations, the average root-mean-squared-deviation is smaller in the binarized (BSR) than in

Figure 2

A Comparison of the Quadratic Scoring Rule and the Binary Scoring Rule

Panel A. Average root-mean-squared-deviation (RMSD)

Panel B. Average deviation ($\epsilon = q - \theta$)



Source: Figures based on the published data from binary scoring rule (BSR) and quadratic scoring rule (QSR) elicitations in Hossain and Okui (2013); Erkal, Gangadharan, and Koh (2020); and Danz, Vesterlund, and Wilson (2022).

Note: All data use the Hossain and Okui definition of "betweenness" to exclude participants with reports far from the induced belief, in the opposite half of the probability space. Total sample size is 391 participants and 2,554 decisions. For panel A, a nonlinear test of the difference in root of the squared-deviation, using paper-fixed effects, is significantly different ($p = 0.046$).

the classic quadratic-scoring rule (QSR), suggesting a higher frequency of truthful revelation under the former. In panel B of Figure 2, we further explore the average difference between the reported and induced beliefs in the two mechanisms. Despite a lower spread in the reports around the induced belief in the binarized-scoring rule, the data surprisingly indicate comparable average deviations and similar deviation patterns under the two mechanisms. Both elicitations show evidence of pull-to-center reporting, with positive deviations when the induced belief θ is less than 0.5 and negative deviations when it exceeds it. Overall, reports under both elicitations differ from the induced beliefs and do so in a manner that is likely to affect econometric inference from the elicited beliefs. While risk aversion only should affect deviations under the quadratic-scoring rule, we see center-biased reporting under both mechanisms, suggesting that neither mechanism is behaviorally incentive compatible.

Why Do Individuals Fail to Reveal the Induced Belief? Explanations and Remedies

In efforts to design better mechanisms, it is critical that we understand why a mechanism fails. While our results make clear that something in the classic and binarized quadratic-scoring rule is malfunctioning, it is not clear what. Experimental techniques have been essential in exploring why individuals do not reveal their types. We offer a few examples to demonstrate the designs used for uncovering possible explanations. For more detail, see the excellent and comprehensive reviews by (Schlag, Tremewan, and van der Weele 2015; Schotter and Treviño 2014; Charness, Gneezy, and Rasocha 2021).

Initial assessments of what drives false reports were focused on understanding whether risk aversion affected deviations under the quadratic-scoring rule. Later investigations have moved to explore a broader set of causes and mechanisms. Three classic experimental-design techniques have been used to shed light on what drives deviations: (1) *design-by-correlation*, where an external measure of a potential driver is used to assess its correlation with the behavior of interest; (2) *design-by-manipulation*, where treatment variation is introduced that will attenuate/exacerbate the effect the driver has on the behavior of interest; or (3) *design-by-subtraction*, where a treatment removes the potential role for the driver of interest entirely, holding everything else constant.

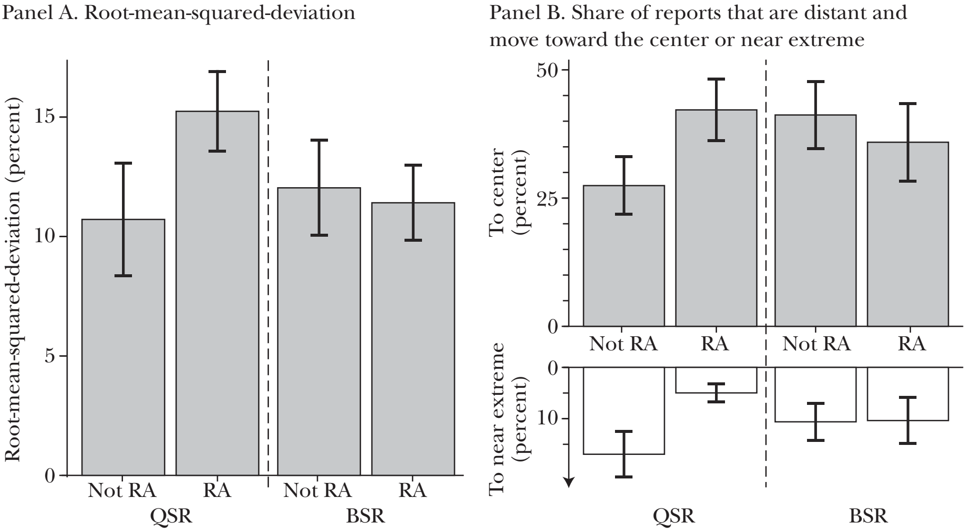
What would these design techniques look like in the context of evaluating whether risk aversion is causing reports to differ from the induced beliefs under the classic quadratic-scoring rule? Design-by-correlation would entail separately eliciting a measure of the participant's risk preference and determining whether it correlates with report deviations. In contrast, design-by-manipulation would explore treatment variations where risk aversion is predicted to further distort the deviations in particular ways, for example by comparing reports when we do and do not give participants an additional stake in the event (and a theoretical motive for a risk-averse individual to hedge). Finally, design-by-subtraction would introduce a treatment, where holding everything else constant the potential for risk aversion is removed.

Design-by-correlation hinges on securing an accurate external measure of the driver of interest (in this case, risk aversion) and a measure that is uncorrelated with other factors that may influence the behavior of interest (for example, confusion). Design-by-correlation is seen as the weaker of the three designs because it does not identify a causal relationship and because inference hinges on the quality of the external measure. Nonetheless it can offer insight. For example, we can elicit risk preferences by presenting participants with a lottery (say a 50 percent chance of winning $10) along with a list of certain payments ($1 to $10 in dollar increments) and ask that participants select the certain payments they prefer to the lottery. A participant indicating that they would prefer certain payments of $4 or more to the lottery would be categorized as risk averse, while a participant, who prefers the lottery unless the certain payment exceeds $6, would be categorized as risk-seeking. The correlation between risk aversion and misreporting can then serve as an indicator for whether center-biased reporting in the quadratic-scoring rule results from it not being incentive compatible for risk-averse individuals.

The data from Hossain and Okui (2013), Erkal, Gangadharan, and Koh (2020), and Danz, Vesterlund, and Wilson (2022) make possible a design-by-correlation

Figure 3

Distortion on Noncentered Induced Beliefs, by Elicitation and Risk Preference



Source: Figures based on the published data from binarized (BSR) and quadratic (QSR) scoring-rule elicitations in Hossain and Okui (2013); Erkal, Gangadharan, and Koh (2020); and Danz, Vesterlund, and Wilson (2022).

Note: All data use the Hossain and Okui definition of "betweenness" to exclude participants with reports far from the induced belief, in the opposite half of the probability space, and include only noncentered beliefs (outside of the 0.4 to 0.6 range). The sample includes 389 participants and 1,851 decisions. Inferentially, in panel A the differences in the RMSD between risk-averse (RA) and not risk-averse (not RA, so either risk-neutral/loving) participants is significant in the QSR (p = 0.046) but not in the BSR (p = 0.625). Similarly, in panel B there are significant differences in both movement directions across risk-preference for the QSR (p = 0.001 both center and near extreme) but not for the BSR (p = 0.314 and p = 0.936).

evaluation of the role played by risk aversion in reports under the standard and binarized quadratic-scoring rules. Focusing on noncentered induced beliefs where risk aversion is predicted to cause a distortion (induced beliefs θ outside of the central 0.4 to 0.6 range), we can assess if deviations in reports for risk-averse (RA) respondents are different from those who are not risk-averse (not RA, so risk-loving or risk-neutral).

The first two bars of Figure 3, panel A, show that under the quadratic-scoring rule the root-mean-squared-deviation is greater for the risk-averse participants, revealing a positive correlation between risk aversion and the size of the deviations. The next two bars show under the binarized-scoring rule no correlation between risk preferences and deviations. Taken in combination, the results are consistent with risk-aversion driving deviations under the quadratic-scoring rule.

We can use the same techniques to examine the interaction between risk attitudes and the direction of the deviations. Focusing on noncentered induced beliefs, panel B of Figure 3 shows the direction of the deviation by the belief elicitation and the participants' risk preferences. The figure shows the direction of the distant reports, moving either towards the center (gray bars) or towards the near extreme (white bars).

The first two sets of bars show for the quadratic-scoring rule the predicted correlation with risk aversion: for risk-averse participants, 42 percent of reports are distant and move toward the center, while for not-risk-averse participants only 27 percent of reports are distant and distorted towards the center (and consistent with risk-seeking preferences, a significantly larger proportion make distant reports toward the near extreme). The next set of two bars show for the binarized-scoring rule that the participants' risk preferences do not correlate with the share of distant reports, neither toward the center nor the near extreme. Instead, independent of risk aversion we find that approximately 40 percent of reports are distant and towards the center and 10 percent are distant and towards the near extreme. In short, assessing the correlation between reported beliefs and participants' risk preferences suggests that risk preferences contribute to the rate of false reports under the quadratic-scoring rule.

Design-by-correlation has also been used to understand the effects of bounded rationality on distortions in belief reports. Burfurd and Wilkening (2022) use a measure of probabilistic sophistication and show that this measure of bounded rationality correlates with larger deviations. Enke and Graeber (2023) examine behavior in a belief-updating task with a shifting prior probability using a binarized-scoring rule. Using a measure of cognitive uncertainty, they assess the impact of bounded rationality on reporting and show that much of the non-Bayesian updating behavior is driven by cognitively-uncertain participants.

For an example of using design-by-manipulation to explain the deviations from induced beliefs, Armantier and Treich (2013) introduce experimental variation over: (1) the size of the incentives used in the quadratic-scoring rule (the maximal prize amount $\$X$), (2) the extent to which the participant has a financial stake in the event being elicited (a separate bonus payment if the elicited event happens), and (3) whether the participant could make a bet on the event being elicited, separate from the elicitation incentives. Relative to a control, these treatment manipulations are predicted to affect reports by risk-averse participants, but to have no effect on reports by those who are risk-neutral. For example, an increase in the size of the incentives should have no impact on reports by risk-neutral participants, while it should make centered reports relatively more attractive for risk-averse participants. Paying a bonus if the event E occurs makes it more attractive for risk-averse participants to report a lower belief, as the bonus decreases the ratio of marginal utilities for the payoff when the event occurs relative to the payoff when the event does not occur. Consistent with risk aversion impacting deviations, they find increased distortions in the reports for all three treatments, leading to the conclusion that risk-aversion contributes to the deviations seen under the quadratic-scoring rule.

Design-by-manipulation has also been used to explore other drivers of deviations. For example, Offerman and Palley (2016) use a manipulation of the classic quadratic-scoring rule. Specifically, they modify the payments to reduce the distortions from loss aversion, where the core treatment variation increases payoffs in the unlikely state where relative losses occur. Consistent with loss aversion affecting deviations, they show that treatment variation reduces false reports (measured by the root-mean-squared-deviation and by the fraction of centered reports).

In an example of design-by-subtraction to explore drivers of reported deviations, Benoît, Dubra, and Romagnoli (2022) assess the role of participants' preference for events they control.[4] They use an elicitation over the respondent's confidence that they are above the median for performance on a task. However, the mechanism used (a mechanism called the probabilistic BDM, which we discuss further below) makes use of two payment arms: one with an exogenous lottery, and one with a lottery based on their performance. A posited channel for false reports is that participants prefer incentives based on realizations under their control, and so distort their beliefs upward. In a clever design-by-subtraction, Benoît, Dubra, and Romagnoli (2022) remove this feature by replacing the exogenous lottery arm with an equivalent incentive that is based on the respondent's performance. As such, the treatment holds constant the incentives, but removes the control motive. The comparison provides evidence that a preference for control is driving false reports, as reports exhibiting self-confidence decrease substantially in the treatment without the control motive.

To summarize, a range of experimental tests and designs have been used to explore why participants under a theoretically incentive-compatible mechanism fail to report their induced type truthfully. In these studies, much of the experimental focus has been the distortive effects of risk aversion under the quadratic-scoring rule. The literature has responded to these findings in one of two ways. One approach involves patching up the misfunctioning mechanism, by collecting additional behavioral measures and applying a correction to the reports. For example, Offerman et al. (2009) gather additional data on preferences and construct corrections to the reports for both risk preferences and ambiguity.[5] The other approach involves updating the mechanism to remove the distortions, as in developing elicitations that are incentive compatible for risk-averse individuals (for example, Hossain and Okui 2013; Benoît, Dubra, and Romagnoli 2022; Mobius et al. 2022).

[4] While it is tempting to see a comparison of the classic and binarized versions of the quadratic-scoring rule as design-by-subtraction, here we are not holding everything constant except risk aversion, as the entire incentive structure is also changing. While design-by-subtraction is seen as the gold standard for experimental design, it is also one of the more challenging design methods, when the driver we wish to identify is more abstract.

[5] While adding supplemental type information such as risk preferences has proved useful for belief elicitations, in more-general mechanisms the designer's goal will typically depend on these type features too. As such, we cannot use supplemental individual assessments of, say, risk and loss aversion to correct the reported types in auctions or other mechanism, as bids will depend on these features as well as the valuations. Uncovering the "true" type will impact the designer's action or interact with the bids of others. So these additional elements of type must be directly accounted for within the mechanism. Revelation and implementation require that the mechanism is incentive compatible for any outcome-relevant type (for example, information on risk preferences).

Direct Tests of Behavioral Incentive Compatibility

Indirect tests of behavioral incentive compatibility indicate when a mechanism malfunctions, but they do not tell us whether failure results from the mechanisms' incentives. To assess whether a mechanism is behaviorally incentive compatible, recent assessments instead look directly at how participants respond to the incentives of a mechanism and ask whether participants perceive them as intended (Danz, Vesterlund, and Wilson 2022).

We discuss two direct tests of behavioral incentive compatibility. The first, a powerful *incentives-only test*, presents participants with a pure choice over the incentives available under the mechanism and evaluates whether most participants select the presumed maximizer. The second, an *info/no-info test*, uses design-by-subtraction to evaluate whether participants are more likely to reveal their induced type truthfully when provided with clear quantitative information on the incentives.

Incentives-Only Test

The *incentives-only test* strips the mechanism of its belief-elicitation framing and presents participants with a choice over the available incentives, asking them to choose their preferred event-contingent payoffs. For example, participants are informed that their earnings depend on whether a red ball is drawn from an urn with red and blue balls where the share of red balls corresponds to an induced belief of θ. The test presents the incentives under the mechanism as pairs of event-contingent payoffs—a payoff if the ball is red, a payoff if the ball is blue—where each pair corresponds to the incentives from a report of q in the mechanism being tested.

Table 1 provides an example of an incentives-only test of the binarized-scoring rule. The eleven options (A through K) correspond to the event-contingent payoffs from each implied report q on the chance of a red ball being drawn, ranging from 0 to 100 percent in 10 percent increments. For example, suppose that participants are informed that the chance of drawing a red ball is $\theta = 0.2$ and are asked to select their preferred event-contingent payoff pair. For participants selecting choice A, the chance of winning \$8 is 0 percent if the selected ball is red and 100 percent if the ball is blue, so a 20 percent chance of \$0 and an 80 percent chance of \$8. Selecting choice B, the chance of winning \$8 is 19 percent if the selected ball is red, and 99 percent if the ball is blue, and so on. For the objective probability of $\theta = 0.2$ on red, participants will maximize their chance of winning \$8 if they select option C, where, as seen in the right-most column (not visible to participants), selecting C corresponds to reporting a belief of $q = 0.2$.

The incentives-only test shows whether participants see the intended (truthfully revealing) choice as maximizing—that is, whether they make a choice corresponding to $q = \theta$. While truthful revelation is predicted for a rational expected-utility-maximizing agent, deviations may result because of cognitive limitations or nonstandard preferences, and because deviations from the intended choice are relatively inexpensive. To see this, consider again the case where there is a $\theta = 0.2$ chance of drawing a red ball. With the theorized maximizing C choice,

Table 1
Incentives-Only Test: Payoffs Available under the Binarized-Scoring Rule

	Binarized scoring rule (BSR)		
	Chance of $8 prize by event		
	Red ball	*Blue ball*	
Lottery option	*(Prob. θ)*	*(Prob. 1 – θ)*	*Implied report q*
A	0%	100%	0.0
B	19%	99%	0.1
C	36%	96%	0.2
D	51%	91%	0.3
E	64%	84%	0.4
F	75%	75%	0.5
G	84%	64%	0.6
H	91%	51%	0.7
I	96%	36%	0.8
J	99%	19%	0.9
K	100%	0%	1.0

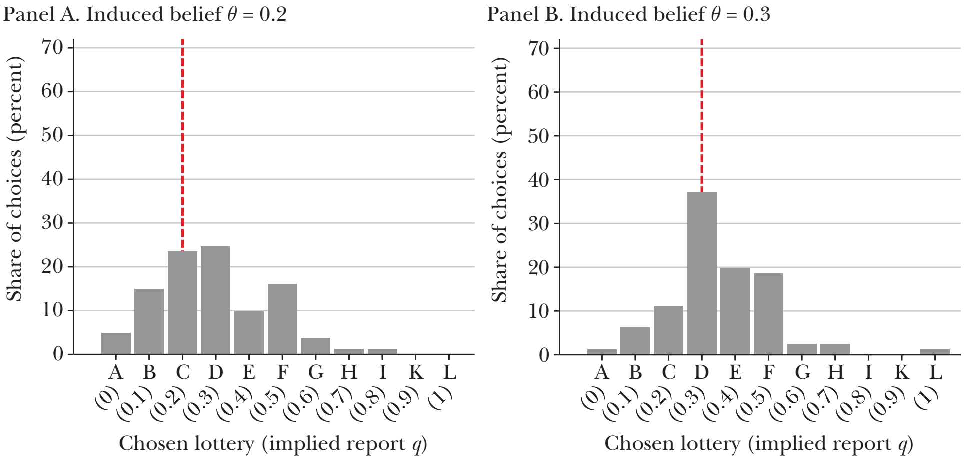
Source: Authors' creation.
Note: Participants are shown the menu of options under the binarized-scoring rule (BSR) and are asked to select their preferred option of event-contingent payoffs conditional on a θ chance that the ball is red. With the theorized maximizer under each elicitation being the option corresponding to $q = \theta$. The implied report q column (which is not shown to participants) indicates the report in the BSR to which this lottery incentive is matched.

the chance of winning $8 is 36 percent when the ball is red and 96 percent when blue. This compound lottery yields an 84 percent chance of winning $8, the largest total chance over the available options. However, a choice such as *D* (corresponding to a more-conservative report of $q = 0.3$) increases the chance of winning by 15 percentage points on red (from 36 percent to 51 percent) while decreasing the chance of winning by only 5 percentage points on blue (96 percent to 91 percent). By design, moving from choice *C* to *D* decreases the overall chance of winning, but note that the decrease is a mere one percentage point. The inexpensive deviation to *D* may therefore tempt individuals who prefer smaller differences in the chance of winning across the binary event outcome.

Figure 4 illustrates the results from an incentives-only test of the binarized-scoring rule for induced probabilities on a red ball of $\theta = 0.2$ or 0.3, respectively. Most participants choose event-contingent payoff options that differ from the assumed maximizer under the mechanism (shown by the vertical dashed line), showing directly that the incentives from the binarized-scoring rule are not behaviorally incentive compatible. Further, the test demonstrates the expected direction of deviations under the mechanism, in this case showing preferences for lottery pairs toward the center choice of F, consistent with the center-biased reporting seen in Figure 2, panel B, and Figure 3, panel B.

Figure 4

Chosen Options in the Incentives-Only Test of the Binarized-Scoring Rule



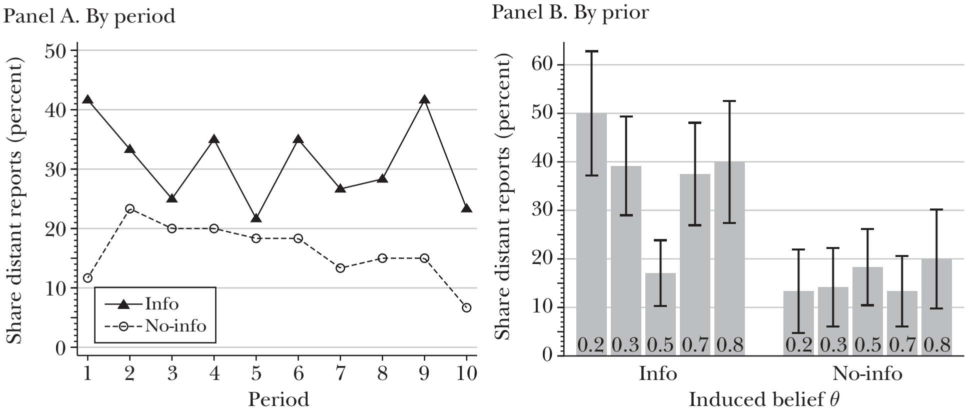
Source: Figure based on the published data from Danz, Vesterlund, and Wilson (2022, Figure 9).
Note: Figure shows distribution of participants' chosen lottery for induced beliefs of $\theta = 0.2$ (panel A) and $\theta = 0.3$ (panel B). The x-axis shows the lottery options (A–K) with corresponding implied belief reports (not shown to participants; 0–1).

Info/No-Info Test

With incentive-compatible belief elicitation, respondents should want to submit their most accurate belief after seeing the incentives. An *info/no-info test* can be used to assess how reports change when participants are given information on the incentives. Holding everything else constant, the test assesses as a minimal criterion for behavioral incentive compatibility whether knowing the offered incentives increases the likelihood that a respondent reveals their type.

The test uses two treatments: an *info* treatment with transparent quantitative information on the incentives, and a *no-info* treatment without the quantitative information on incentives. All other features are held constant. Participants in both treatments are given summary statements on the qualitative consequences of truthful reporting and the size of the stakes involved, $\$X$. The only difference is that participants in the info treatment also receive information on the precise quantitative incentives associated with any report under the mechanism. For example, participants in the no-info treatment for the binarized-scoring rule are only informed 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." Participants in the info treatment also received (1) a concise verbal description of how prize realizations were determined; (2) were shown the exact incentive for the provisionally selected belief report at the time of choice, and (3) were given feedback on the event outcomes and realized incentives at the end of each period.

Figure 5
Fraction Distant Reports in Info/No-Info Test of the Binarized Scoring Rule



Source: Figure based on the published data from Danz, Vesterlund, and Wilson (2022, Figure 4).
Note: Figure shows fraction of distant reports in the info and no-info treatments over time (panel A) and by induced belief (panel B). Distant reports are belief reports deviating by more than five percentage points from the induced belief.

Figure 5 illustrates the results from the info/no-info test of the binarized-scoring rule. The experiment was conducted over ten periods. At the start of each period, a simple belief was induced (based on probabilities of certain outcomes with a ten-sided die-roll), with the possibilities including 0.2, 0.3, 0.5, 0.7, and 0.8. Panel A of Figure 5 shows the rate of distant reports (those more than 5 percentage points from the induced belief) under the info and no-info treatments. Disturbingly, the rate of distant reports is substantially higher in the info than in the no-info treatment in every period of the experiment, revealing that participants are less likely to report the induced belief when they are presented with information on the quantitative incentives. Further, panel B of Figure 5 shows the rate of distant reports by treatment and for each induced belief. As evidence that incentives are distorting accurate reporting, we see that the rate of distant reports is independent of the induced belief in the no-info treatment (right-hand bars), but varies with the induced belief in the info treatment (left-hand bars), with distant reports being more likely for noncentered induced beliefs than for a centered belief of $\theta = 0.5$. Importantly, there is no evidence that risk aversion is the culprit for deviations under the info treatment, both because risk aversion theoretically should not play a role under the binarized-scoring rule, and because separately measured risk attitudes do not predict the likelihood of distant reports.

The no-info treatment demonstrates that participants have a reasonable understanding of the task at hand—as they report the induced beliefs at high rates in the absence of quantitative information on the incentives. Paradoxically, information on the incentives causes individuals to deviate from reporting their true type, demonstrating that the binarized-scoring rule is not behaviorally incentive compatible.

Other Applications of Direct Tests

Direct tests of incentive compatibility have been applied to several other belief elicitation mechanisms. For example, Danz, Vesterlund, and Wilson (2024) find that results for the quadratic-scoring rule are similar to those for the binarized-scoring rule. An incentives-only test of the quadratic-scoring rule shows that the majority of participants prefer payoffs that differ from the intended maximizer, and that many prefer the incentives consistent with center-biased reporting, where there are smaller differences in event-contingent payoffs. An info/no-info test of the quadratic scoring rule shows that information on the quantitative incentives *increases* distant reports, a difference that is maintained throughout the experiment. Further, mirroring the results from the binarized-scoring rule, distant reports under the classic quadratic-scoring rule are only sensitive to the induced belief in the info treatment, and are far more likely for noncentered induced beliefs. That is, direct tests of the incentives reveal that the classic quadratic-scoring rule is not behaviorally incentive compatible, and that the incentives directly contribute to the false reports seen under the mechanism.

Danz, Vesterlund, and Wilson (2024) also explore the behavioral incentive compatibility of the *probabilistic Becker-DeGroot-Marschak mechanism* (Becker, DeGroot, and Marschak 1964; Karni 2009; Mobius et al. 2022; see also Smith 1961; Grether 1980), an increasingly popular elicitation. Similar to the binarized-scoring rule, the incentives are designed to be incentive compatible for arbitrary risk preferences and ensure that truthful revelation maximizes the chance of winning a fixed prize. Under the probabilistic Becker-DeGroot-Marschak (p-BDM) mechanism, the participant reports a belief q for, say, the share of red balls in the urn out of a total of 100. The payment depends on the reported belief, the event realization, and a randomly drawn number $z \in [0,1]$. If z is higher than the reported number q, the participant receives \$$X$ with probability z. If the draw z is less than the estimated value q, then the participant receives \$$X$ if the event E occurs. That is, for a reported belief q of event E, the participant receives \$$X$ with probability $q + (1 - q^2)/2$ if the event occurs and with probability $(1 - q^2)/2$ if the event does not occur. While truthfully revealing the induced belief maximizes the chance of winning, note that the offered incentives differ markedly from those under the binarized-scoring rule. From Table 1, under the binarized-scoring rule an event-independent probability of winning (of 75 percent) can be ensured by a centered report of $q = 0.5$. In contrast, under the probabilistic Becker-DeGroot-Marschak mechanism, an event-independent probability of winning (50 percent) can be ensured by an extreme report of $q = 0.0$.

Danz, Vesterlund, and Wilson (2024) show in an incentives-only test of the probabilistic Becker-DeGroot-Marschak mechanism that the vast majority of participants prefer choices that differ from the intended maximizer, indeed 69 percent of participants opt for the event-independent choice corresponding to reporting $q = 0.0$. Results from the info/no-info test further confirm that the probabilistic

Becker-DeGroot-Marschak mechanism is not behaviorally incentive compatible. Distant reports are more likely when participants are informed of the incentives under the mechanism, and consistent with the incentives-only test, reports are pulled toward $q = 0.0$. For example, at an induced belief of $\theta = 0.2$ only 7 percent of reports are both distant and towards zero in the no-info treatment. In contrast, this figure jumps to 21 percent of reports in the info-treatment (with no differences in the fraction of distant reports in the other direction).

To summarize, choices made under the incentives-only test for three commonly used belief elicitations reveal that the *majority* of participants *do not* prefer the theorized maximizing choice. Further, info/no-info tests show that providing participants with quantitative information on their incentives substantially increases the rate of false reports. That is, the incentives commonly used to encourage truthful revelation do not make it in the participant's "best interest to reveal their type," implying failures of behavioral incentive compatibility.

Conclusion

Economists have developed a range of mechanisms that are theoretically incentive compatible to provide participants with incentives to reveal their private type. Experimental economics has played a critical role in determining whether mechanisms are also behaviorally incentive compatible. The experimenter's ability to manipulate and induce an individual's type make it possible to determine whether the developed mechanism encourages truthful revelation. In reviewing the experimental techniques developed to assess behavioral incentive compatibility, we focus on the simple case of individual belief elicitation, showing both how indirect assessments can be performed within the mechanism, and how direct assessment can be done by directly evaluating the mechanism's incentives.

Applying the different experimental techniques to assess belief elicitations paints a dismal picture of the extent to which these encourage truthful revelation. Danz, Vesterlund, and Wilson (2024) show for the most-used belief elicitation mechanisms (the classic and binarized quadratic-scoring rule and the probabilistic Becker-DeGroot-Marschak rule) that participants largely prefer payoffs different from the intended maximizer under the mechanism, and that information on the incentives *increases* the rate of false reports.

The high rate of false reports has serious implications when using beliefs elicited under the mechanism. As an example, Danz, Vesterlund, and Wilson (2022) replicate the well-known Niederle and Vesterlund (2007) study on gender and competition. The original finding of Niederle and Vesterlund was that, conditional on performance, men enter competitions more than women, but that part of this difference was driven by men being more confident than women. Using an info/no-info comparison across the binarized scoring rule, Danz, Vesterlund, and Wilson (2022) elicit beliefs on relative performance for men and women. The no-info treatment replicates the prior finding that women are less confident about winning a competition than men, and that controlling for beliefs reduces the gender gap in preferences for competition. In contrast, for the info treatment, the results do not uncover a gender gap in confidence and controlling for beliefs does not help explain the gender gap in preferences for competition. Providing clear information on the quantitative incentives shifts reported beliefs and changes inference. Both the original study and the no-info treatment lead to a conclusion that differences in confidence between men and women are important, and contribute to the gender gap in competition. In contrast, for the info treatment, the gender gap in competition is solely explained by preferences. These results outline the large ramifications from using an elicitation mechanism that is not behaviorally incentive compatible. Inferences drawn from biased reports will attenuate estimated treatment responses when beliefs are used as a dependent (left-hand-side) variable and bias all estimates when used as an explanatory (right-hand-side) variable.

While we have focused on the case of belief elicitation, indirect assessments of behavioral incentive compatibility have been used to evaluate a broad set of mechanisms, including auctions, centralized clearing houses, and so on (for example, Kagel, Harstad, and Levin 1987; Coppinger, Smith, and Titus 1980; Kagel et al. 1989; Chen and Sonmez 2006; Roth 2017). However, direct assessments can also be extended to such settings, offering simple diagnostic tests directly targeted at the mechanism incentives. Info/no-info tests can be used to determine whether clear information on the incentives increases truthful revelation, while the incentives-only test can be used to convert the effective incentives into stark decision problems by holding constant the theorized behavior of other participants and directly evaluating whether individuals prefer the assumed maximizer.

For example, Danz, Vesterlund, and Wilson (2024) use the pure-incentives test to assess the "deferred acceptance" mechanism that Boston, New York, and other cities use to assign students to schools and that is used nationally to match newly graduated doctors to residency programs. Stripping away the mechanism and the strategic features, which typically require many participants to submit rankings of their potential options, they find the vast majority of participants prefer the outcome associated with truthfully revealing their ranking. That is, the incentives under deferred acceptance are behaviorally incentive compatible, and failures in truthfully revealing preference rankings must result from other aspects of the mechanism. This insight is particularly helpful in light of the evidence that individuals, when faced with the mechanism, fail to reveal their type (as in Echenique et al. 2016; Dreyfuss, Heffetz, and Rabin 2022; Rees-Jones 2018, and this symposium). Results from the pure-incentives test demonstrate that these failures are not driven by the incentives per se, but by other aspects of the algorithm.

Where static mechanisms might fail, behavioral research has opened up other design channels for improving mechanism performance. For example, dynamic framings in which types are revealed through a sequence of simpler, starker decisions, can make the dominant choice more obvious and increase truthful reporting (along the lines of Li 2017, and this symposium). For example, Hao and Houser (2017) demonstrate a substantial increase in truthful reporting when they reframe the probabilistic Becker-DeGroot-Marschak mechanism as a "clock auction"—that is, an auction with rounds of bidding where in each round participants reveal whether their belief is greater than the current clock value, rather than a declarative mechanism requiring a one-time report on q (see also Chapman et al. 2018; for impact of dynamic framing and more-careful instructions on deviations, see Healy 2017; Holt and Smith 2016).

Another approach—perhaps counterintuitive—is to provide less information on the mechanism's incentives. In the domain of belief elicitations, evidence of failed behavioral incentive compatibility has largely resulted in hiding the mechanism's incentives and instead providing participants with a summary statement of the incentives. For example, we may simply inform participants that truthful revelation maximizes the chance of winning an $8 prize (where statements on truthful revelation being in the participant's "interest" are deceptive given the pure incentives test). While this approach is tempting, we caution against it. If we are to incentivize truthful revelation, we recommend instead that the incentives provided be revised to encourage rather than discourage revelation. As part of this, it may be necessary to consider coarser mechanisms where simple and stark incentives are provided to secure truthful revelation. While this can reduce the precision of the provided reports, it may serve to reduce the hidden distortions in them, too. In developing and exploring new mechanisms, however, it is critical that attention be given to whether new candidates are behaviorally incentive compatible, and tests must be conducted to ensure that individuals see it as in their interest to truthfully reveal their type.

■ *We are grateful to Shengwu Li, Jonathan Parker, Nina Pavcnik, Alex Rees-Jones, Timothy Taylor, and Heidi Williams for invaluable comments and suggestions. We are also grateful to Garett Kohno, Priyoma Mustafi, and Brandon Williams for excellent research assistance with this paper.*

References

Armantier, Olivier, and Nicolas Treich. 2013. "Eliciting Beliefs: Proper Scoring Rules, Incentives, Stakes and Hedging." *European Economic Review* 62: 17–40.

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

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science* 9 (3): 226–32.

Benoıt, Jean-Pierre, Juan Dubra, and Giorgia Romagnoli. 2022. "Belief Elicitation When More than Money Matters: Controlling for 'Control.'" *American Economic Journal: Microeconomics* 14 (3): 837–88.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Burfurd, Ingrid, and Tom Wilkening. 2022. "Cognitive Heterogeneity and Complex Belief Elicitation." *Experimental Economics* 25 (2): 557–92.

Chapman, Jonathan, Erik Snowberg, Stephanie Wang, and Colin Camerer. 2018. "Loss Attitudes in the US Population: Evidence from Dynamically Optimized Sequential Experimentation (DOSE)." NBER Working Paper 25072.

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

Charness, Gary, and Dan Levin. 2005. "When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect." *American Economic Review* 95 (4): 1300–1309.

Chen, Yan, and Tayfun Sönmez. 2006. "School Choice: An Experimental Study." *Journal of Economic Theory* 127 (1): 202–31.

Coppinger, Vicki M., Vernon L. Smith, and Jon A. Titus. 1980. "Incentives and Behavior in English, Dutch and Sealed-Bid Auctions." *Economic Inquiry* 18 (1): 1–22.

Cox, James C., Bruce Roberson, and Vernon L. Smith. 1982. "Theory and Behavior of Single Object Auctions." *Research in Experimental Economics* 2: 1–43.

Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2022. "Belief Elicitation and Behavioral Incentive Compatibility." *American Economic Review* 112 (9): 2851–83.

Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2024. "The Incentives-Only Test: Assessing Behavioral Incentive Compatibility in Mechanisms." Unpublished.

Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2024. *Data and Code for: "Evaluating Behavioral Incentive Compatibility: Insights from Experiments."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E209063V1.

Dreyfuss, Bnaya, Ori Heffetz, and Matthew Rabin. 2022. "Expectations-Based Loss Aversion May Help Explain Seemingly Dominated Choices in Strategy-Proof Mechanisms." *American Economic Journal: Microeconomics* 14 (4): 515–55.

Echenique, Federico, Alistair J. Wilson, and Leeat Yariv. 2016. "Clearinghouses for Two-Sided Matching: An Experimental Study." *Quantitative Economics* 7 (2): 449–82.

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh. 2020. "Replication: Belief Elicitation with Quadratic and Binarized Scoring Rules." *Journal of Economic Psychology* 81: 102315.

Enke, Benjamin, and Thomas Graeber. 2023. "Cognitive Uncertainty." *Quarterly Journal of Economics* 138 (4): 2021–67.

Ewers, Mara, and Florian Zimmermann. 2015. "Image and Misreporting." *Journal of the European Economic Association* 13 (2): 363–80.

Gächter, Simon, and Elke Renner. 2010. "The Effects of (Incentivized) Belief Elicitation in Public Goods Experiments." *Experimental Economics* 13 (3): 364–77.

Grether, David M. 1980. "Bayes Rule as a Descriptive Model: The Representativeness Heuristic." *Quarterly Journal of Economics* 95 (3): 537–57.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing Information Provision Experiments." *Journal of Economic Literature* 61 (1): 3–40.

Hakimov, Rustamdjan, and Dorothea Kübler. 2021. "Experiments on Centralized School Choice and College Admissions: A Survey." *Experimental Economics* 24 (2): 434–88.

Hao, Li, and Daniel Houser. 2012. "Belief Elicitation in the Presence of Naïve Respondents: An Experimental Study." *Journal of Risk and Uncertainty* 44: 161–80.

Harrison, Glenn W., Jimmy Martínez-Correa, and J. Todd Swarthout. 2014. "Eliciting Subjective Probabilities with Binary Lotteries." *Journal of Economic Behavior and Organization* 101: 128–40.

Healy, Paul J. 2017. "Epistemic Experiments: Utilities, Beliefs, and Irrational Play." Unpublished.

Hollard, Guillaume, Sébastien Massoni, and Jean-Christophe Vergnaud. 2016. "In Search of Good Probability Assessors: An Experimental Comparison of Elicitation Rules for Confidence Judgments." *Theory and Decision* 80: 363–87.

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.

Hossain, Tanjim, and Ryo Okui. 2013. "The Binarized Scoring Rule." *Review of Economic Studies* 80 (3): 984–1001.

Hurwicz, Leonid. 1972. "On Informationally Decentralized Systems." In *Decision and Organization: A*

Volume in Honor of J. Marschak, edited by C. B. McGuire and Roy Radner, 297–334. Amsterdam: North-Holland.

Hurwicz, Leonid. 1973. "The Design of Mechanisms for Resource Allocation." *American Economic Review* 63 (2): 1–30.

Kagel, John H., Ronald M. Harstad, and Dan Levin. 1987. "Information Impact and Allocation Rules in Auctions with Affiliated Private Values: A Laboratory Study." *Econometrica* 55 (6): 1275–1304.

Kagel, John H., and Dan Levin. 1993. "Independent Private Value Auctions: Bidder Behaviour in First-, Second- and Third-Price Auctions with Varying Numbers of Bidders." *Economic Journal* 103 (419): 868–79.

Kagel, John H., Dan Levin, Raymond C. Battalio, and Donald J. Meyer. 1989. "First-Price Common Value Auctions: Bidder Behavior and the 'Winner's Curse.'" *Economic Inquiry* 27 (2): 241–58.

Kagel, John H., and Alvin E. Roth. 2000. "The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment." *Quarterly Journal of Economics* 115 (1): 201–35.

Karni, Edi. 2009. "A Mechanism for Eliciting Probabilities." *Econometrica* 77 (2): 603–06.

Kessler, Judd B., and Lise Vesterlund. 2015. "The External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effect." In *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter, 391–406. Oxford: Oxford University Press.

Köszegi, Botond, and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121 (4): 1133–65.

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.

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2022. "Managing Self-Confidence: Theory and Experimental Evidence." *Management Science* 68 (11): 7793–7817.

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?" *The Quarterly Journal of Economics* 122 (3): 1067–1101.

Nyarko, Yaw, and Andrew Schotter. 2002. "An Experimental Study of Belief Learning Using Elicited Beliefs." *Econometrica* 70 (3): 971–1005.

Offerman, Theo, and Asa B. Palley. 2016. "Lossed in Translation: An Off-the-Shelf Method to Recover Probabilistic Beliefs from Loss-Averse Agents." *Experimental Economics* 19: 1–30.

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.

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

Rees-Jones, Alex. 2018. "Suboptimal Behavior in Strategy-Proof Mechanisms: Evidence from the Residency Match." *Games and Economic Behavior* 108: 317–30.

Roth, Alvin E. 2017. "Experiments in Market Design." In *Handbook of Experimental Economics*, Vol. 2, edited by John H. Kagel and Alvin E. Roth, 290–346. Princeton: Princeton University Press.

Schlag, Karl H., James Tremewan, and Joël van der Weele. 2015. "A Penny for Your Thoughts: A Survey of Methods for Eliciting Beliefs." *Experimental Economics* 18: 457–90.

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

Smith, Cedric A. B. 1961. "Consistency in Statistical Inference and Decision." *Journal of the Royal Statistical Society: Series B (Methodological)* 23 (1): 1–25.

Thoma, Carmen. 2016. "Under- versus Overconfidence: An Experiment on How Others Perceive a Biased Self-Assessment." *Experimental Economics* 19 (1): 218–39.

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

Winkler, Robert L., and Allan H. Murphy. 1970. "Nonlinear Utility and the Probability Score." *Journal of Applied Meteorology and Climatology* 9 (1): 143–48.

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