Journal of Economic Behavior and Organization 189 (2021) 234–256

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Experimental methods: Eliciting beliefs

Gary Charness[a,*], Uri Gneezy[b], Vlastimil Rasocha[c]

[a] Department of Economics, Rady School of Management, University of California, Santa Barbara, United States
[b] San Diego, and CREED, University of Amsterdam, Netherlands
[c] Department of Economics, Stanford University, United States

ARTICLE INFO

Article history:
Received 17 January 2021
Revised 14 June 2021
Accepted 15 June 2021
Available online 15 July 2021

JEL classifications:
B49
C90
C91
C92

Keywords:
Experimental methodology
Belief elicitation
Comprehension
Scoring rules
Incentives

ABSTRACT

Expectations are a critical factor in determining actions in a great variety of economic interactions. Hence, being able to measure *beliefs* is important in many economic environments. In this article, we review the approaches that have been used to measure beliefs and make comparisons of their effectiveness. We also discuss belief elicitation when the truth is not verifiable. We find that introspection (non-incentivized responses) seems to do as well as rather complex incentivized methods. We conjecture that simple and easy-to-comprehend incentivized methods are superior to introspection, in fact there are very few studies making such comparisons; this is an avenue for future research. We also discuss ways in which relatively complex methods could be made easier for usage in experimental work.

1. Introduction

Beliefs are central to understanding economic behavior. They determine behavior in uncertain investments and more generally in choices between sets of risky options. They may also affect behavior in interactions with other people in which we form beliefs about actions and reactions. In such cases, economic models use beliefs in order to prescribe and predict behavior. A problem in the empirical literature is that we cannot directly observe these beliefs, which must instead either be inferred or elicited. In addition, in some cases even the individuals themselves may not be aware of their own beliefs. The reply could also depend on the way we elicit the belief and the degree to which this is understood. It is also possible that by incentivizing (or even simply asking) people to state their beliefs, we may change their behavior; we discuss this in some detail later.

In this article we survey the leading methods for measuring beliefs, discussing in turn the issues involved with each. We discuss standard methods as well as approaches to eliciting beliefs in environments where outcomes of interest are not verifiable-an important setting scarcely discussed in the experimental literature. We then consider the research that makes comparisons across different approaches. There are some inevitable issues with comparing elicited beliefs empirically. For example, just because one method is best in a lab setting with complete information does not make it best outside of

* Corresponding author.
E-mail addresses: charness@econ.ucsb.edu (G. Charness), ugneezy@ucsd.edu (U. Gneezy), rasocha@stanford.edu (V. Rasocha).

https://doi.org/10.1016/j.jebo.2021.06.032
0167-2681/

the lab - with many different kinds of incomplete information. So, discussing accuracy or levels in general terms may be problematic.

There are some excellent surveys on belief elicitation, most notably Schotter and Trevino (2014) and Schlag et al. (2015). Our focus is different. A central finding in our article is that theoretically robust, but complex, methods do not systematically outperform introspection (asking for beliefs without providing incentives). In addition, there is little comparison of the effectiveness of simpler incentivized mechanisms (our usual preference) to either more complex methods or introspection. This is an important avenue for future research.

Our main organizing principle throughout this article is the degree of complexity. Complexity of the elicitation method stands in participants' way of making utility-maximizing belief reports. Participants may struggle with the computational demands of long payoff tables, fail to imagine anything concrete under mathematical formulae, and fall short of understanding abstract decisions in complicated environments. If participants are confused, their reports may be unreliable. This makes complexity a first-order concern in belief elicitation.

Incentivizing choices inherently comes with some complexity. Hence, non-incentivized introspection is the simplest way to elicit beliefs. It is also fast and cheap, which makes it particularly attractive in applications. Asking experimental participants to report their beliefs without any reward for honesty, however, has major drawbacks. If answering a belief-elicitation question is cognitively taxing, individuals may not put sufficient effort into their answer, resorting to defaults or focal reports.[1] Moreover, individuals may choose to distort their expressed beliefs to maintain a positive (self)image and/or justify their actions. As our discussion of empirical comparisons of elicitation methods show, these concerns are justified. But utilizing complex methods may also lead to systemic biases in elicited beliefs, and these appear to be as serious as those arising from introspection.

The most popular methods for incentivizing honesty in belief elicitation are complex. Their appeal lies in their attractive theoretical properties-the widely used *proper scoring rules* make truthful reporting the expected payoff-maximizing strategy; *probability matching* and *the binarized scoring rule* are additionally robust to deviations from risk-neutrality. If participants whose beliefs are being elicited attempt to derive their optimal action under these mechanisms, they need to sift through long payoff tables, be adept at working with mathematical formulae, and wrap their mind around abstract environments. The sophistication required to intuit that truth-telling is optimal is even greater when using methods such as *the Bayesian Truth Serum*, which are applicable in settings where the outcome of the random variable of interest is not verifiable. We suspect that it is precisely this complexity that leads to the inability of popular incentivized methods to empirically outperform introspection with all its flaws.

Our conjecture is that reliable elicitation is possible and can be achieved using simple incentivized methods that are easy to understand. Several such methods, such as the *frequency method* and the *interval method*, are available and sometimes used in applications. Under the frequency (interval) method a participant is paid a fixed reward if her guess exactly (approximately) matches the actual realization. For example, if a participant is asked to predict how many out of 10 people make a particular choice, the frequency method rewards her if the guess is exactly correct. The interval method rewards her if her guess is, say, within one person of being correct. These methods maintain some nice theoretical properties of more complex methods without large complexity costs. There is little research comparing the performance of simple incentivized methods to complex methods or introspection. This is an important avenue for future research, as is the development of new simple methods better suited to particular elicitation tasks.

What should researchers who want to incentivize truth-telling, but worry about complexity, do in practice? First, when possible, they could use a simple incentivized method. These are not always directly applicable - we discuss when and how the elicitation task may be transformed to enable simple incentivization, noting the tradeoffs involved therein. Second, they could choose a simpler implementation of a more complex method (this could be particularly beneficial when using probability matching). Finally, they may try to partially alleviate the complexity issue by using simpler, non-mathematical descriptions of the methods, graphs and figures, or extensive understanding checks.

The remainder of this article is structured as follows. In Section 2 we describe existing methods, listing these by type and describing them in detail. Section 3 presents comparisons across methods from existing literature. Section 4 provides a practical guide for reducing the complexity of the elicitation task. We conclude in Section 5.

2. Belief-elicitation mechanisms

In this section, we first provide some background. We then discuss belief-elicitation methods divided into three categories according to their complexity. Most of these methods elicit beliefs when the outcome of the random variable of interest is verifiable (the typical case). The other methods described can be applied when the researcher cannot verify the outcome. A more detailed literature review can be found in our Online Appendix.

[1] A reviewer makes a deeper point. Is the experimenter interested in getting the most accurate report of the truth (the original motivation behind the QSR) or the "true" belief? The latter may be more relevant if one is interested in examining how beliefs may have influenced choices that have already been elicited.

2.1. Background

Suppose we are interested in eliciting beliefs about a binary random variable V with possible outcomes E and E^c. In particular, we want to know what subjective probability a participant assigns to outcome E. We denote the participant's true subjective probability of E with p and the elicited probability with r. The goal of elicitation methods is to induce the participant to report $r = p$, or at least make a report from which the true probability can be extracted, using methods that estimate utility functions to back out the truth.

In the discussion that follows, we describe the key properties of popular elicitation methods in this simple setting. We stick to the basics – detailed discussions of elicitation methods' theoretical properties and their extensions to more general settings can be found in Schlag et al. (2015) and Schotter and Trevino (2014).[2]

A central issue in our discussion is the degree of complexity of the mechanism. There are (at least) two dimensions to this: the complexity of the method *per se* and the complexity of the execution. We conjecture that a method is more likely to yield meaningful responses if the responder understands the principles underlying the mechanism. For example, a researcher using the quadratic scoring rule can tell the subjects that, under some assumptions, they are best off reporting their true beliefs; one can even tell the subjects that this is easily proven and insert the proof at the end of the instructions. But will this induce the same responses as when the mechanism is understood by participants?

The issue of complexity can be summarized as follows: Assume that people have subjective beliefs and are aware of them. Assume further that the agent wishes to maximize some utility function, and that the elicitation mechanism in question is such that reporting the agent's true belief would fulfill her desires and maximize her utility. The complexity of the mechanism prevents the agent from understanding this, and hence creates a divide between desires and behavior.

We follow the notion that complexity is related to the cost of understanding and implementation. Herbert Simon proposed the idea of bounded rationality, where one's cognitive limitations stem from too high a cost for fully-rational behavior (this cost could be infinite in the case of a method that cannot be understood by participants). But to some degree, participants can compensate for limitations by what Simon calls "structures of the environment." Thus, complexity is a function of the cognitive cost of the elicitation mechanism, which can potentially be minimized by effective presentation. Building on Simon, Oprea (2020) considers complexity through the automata needed to characterize a system. The basic idea is rather intuitive: if the cost of comprehension seems too high, one will not put forth the effort needed to understand it, and behavior will be less reliable and meaningful.

In this light, here is a non-exhaustive list of complexity types in this realm:

- Computational complexity-long payoff tables (like those often used with scoring rules) can be difficult to absorb. Checking on whether it is optimal to respond in a certain way would involve making many calculations.
- Mathematical complexity-participants may not be familiar with the mathematical proofs or do not have the ability to understand them (even ones that are seemingly simple to researchers).
- Abstraction-participants may struggle to comprehend abstract experimental settings, aggregate the consequences of their decisions over hypothetical scenarios, and link abstract decisions to their beliefs.

Based on the above approach, we rank various belief-elicitation mechanisms (described in more detail below) from relatively simplest to relatively most complex.[3]

Simpler

- Introspection-simply report belief without incentives.
- Frequency-guess how many instances out of N draws result in outcome E. Win a prize if exactly correct.
- Interval-guess the proportion of instances out of N draws that result in outcome E. Win a prize if approximately correct.
- Belief-coordination games-guess beliefs of others. Win a prize if correct.

More complex

- Outcome matching-choose between a certain payment and lottery in each row of a table.
- Probability matching-choose between an "outcome lottery" that pays a prize if the outcome is E and a "constructed lottery" that pays a prize with probability r, for r increasing from 0 to 100. The complexity of this mechanism vitally depends on the implementation. From simplest to most complicated:
  - Multiple price list setting-choose between the outcome lottery and the constructed lottery in each row of a table.
  - Sequential draws setting-a belief is reported, and a uniform random variable is drawn. Depending on the draw, the participant faces either the outcome or the constructed lottery with r equal to the draw.
  - Auction setting-bid against a computer in an English or second price sealed bid auction.

[2] It is also possible to use intervals for beliefs, as in Rustichini and Villeval (2014). Schlag and van der Weele (2013) suggest an elicitation method for most likely confidence intervals. However, this approach introduces another layer of complexity and is not easy for subjects to understand.

[3] Of course, these are not black-and-white distinctions; for example, some readers may feel that the QSR is simple and others may feel differently.

Very complex

- Scoring Rules-receive a score for the belief report that depends on the realization of the random variable V. Higher scores correspond to a smaller "distance" of the belief from the realization. The Quadratic Scoring Rule (QSR, described below) is the most commonly used proper scoring rule.
- Binarized scoring rule-receive a score as in QSR. This score corresponds to the probability of winning a prize. Just as abstract as the most abstract versions of QSR and more due to an additional layer of randomization.
- Incentivized elicitation methods when the outcome of V is not verifiable-most of these methods are extraordinarily difficult to fathom.

In addition to complexity, there are some other relevant issues concerning belief elicitation. Since many of these have been discussed in detail in Schlag et al. (2015), we touch on these only briefly here.

2.1.1. Beliefs and actions

A number of studies have considered the correspondence between beliefs and actions. This concern goes beyond just making sure that elicitation is incentive-compatible, and subjects understand this. Even if a belief about the realization of a random variable V is elicited precisely, this belief may not be used by agents to choose their action.[4]

Nyarko and Schotter (2002) suggest that participants use their incentivized stated beliefs as the basis of their choices. They find that the stated-belief model substantially outperforms the empirical belief models they tested (fictitious play beliefs or Cournot beliefs). Danz et al. (2012) extends these results to 3×3 variable-sum games, finding the stated-belief model best in terms of goodness of fit, with similar best-response rates (between 60% and 80%).

Charness and Dufwenberg (2006) find first-mover play strongly depends on first-order beliefs. Rey-Biel (2009) finds that participants best respond to their stated beliefs in 69.4% (64.9%) of the time in the constant-sum (variable-sum) games played. Blanco et al. (2010) find that first movers almost always best-respond to their beliefs about second movers. Hyndman et al. (2013) find that 62% of participants best respond to the beliefs they state, with more extreme beliefs receiving the most consistent best response.

People who are higher on the cognitive hierarchy may consider second-order beliefs before choosing an action in a game, so these may also be important. Dufwenberg and Gneezy (2000) and Charness and Dufwenberg (2006) study guilt aversion, which predicts that the responder's actions will be correlated with their beliefs about first-mover beliefs. Both studies find a strong correlation between actions and second-order beliefs, consistent with the guilt-aversion hypothesis. Manski & Neri (2013) consider the coherence of first- and second-order beliefs and propose a method for eliciting second-order beliefs using distributional assessments (ranges) rather than point predictions. First- and second-order beliefs are largely coherent.

2.1.2. Elicitation and behavior

Simply eliciting beliefs might influence behavior, by either leading to bias or to a different understanding of the task at hand. As a result, a potential concern is that merely eliciting a belief might affect the subsequent behavior of the party whose beliefs were elicited. One standard approach is to elicit beliefs only after all game choices have been made. A problem with doing this is that beliefs might be shaped by actions. In addition, a researcher might wish to elicit beliefs after every period of a multi-period experiment.

Does elicitation affect behavior? Overall, the results are mixed, although Schotter and Trevino (2014, p. 103) write "beliefs elicited in the laboratory are meaningful (i.e., they are generally used as the basis for behavior), and the process of eliciting beliefs seems not to be too intrusive."

Nyarko and Schotter (2002) argue that elicitation had no effect on the likelihood of choosing one of the two available strategies in their game (although some of their tests suggest that it may have). Results in Costa-Gomes and Weizsacker (2008) and Ivanov (2011) also support the view that the act of belief elicitation does not alter behavior. On the other hand, Croson (2000) presents evidence that eliciting beliefs in an incentivized manner leads to participants decreasing their contributions in public-good games. Gächter and Renner (2010) find the opposite effect, while Wilcox and Feltovich (2000) find no difference. Rutström and Wilcox (2009) do find that belief elicitation can alter paths of play in a two-player repeated asymmetric matching-pennies game, although this effect occurs only during early periods and only for players with strongly asymmetric payoffs.

In our view this is not a settled issue. We recommend proceeding cautiously in this regard, asking first about the data that seems the most critical for the purpose of the study. There is a need for future research to systematically test this question.

2.1.3. Hedging

If participants in an experiment are receiving income from both the actions they take and their guesses about their opponent's actions, participants could try to hedge and coordinate their actions and guesses so as to provide an expected

[4] To illustrate the issue, suppose that a researcher is interested in the behavior of an agent who bets on the outcome of a biased die. If the participant is aware of the bias, her truthfully-elicited beliefs will reflect it. But she may still bet on the outcome of the die "as if" it were unbiased-perhaps because it is easier to calculate the expected returns of different betting strategies this way.

payoff with less variance. Does a hedging possibility with belief elicitation affect behavior? If so, are there approaches that can help?

Blanco et al. (2010) provide the first direct test of hedging in an experiment with belief elicitation, concluding (p. 434) that "hedging-like behavior may be difficult to completely eradicate." Both Nyarko and Schotter (2002) and Costa-Gomes and Weizsäcker (2008) use indirect evidence from their experiments to conclude there is no apparent hedging bias. Haruvy et al. (2007) compare incentivized beliefs about the value of assets by traders in an asset market with stated beliefs by "observers" who do not trade and are only paid for belief accuracy. They find no significant differences between the accuracy of predictions of traders and observers.

On the other hand, Armentier and Treich (2013) generalize the analysis of proper scoring rules to richer environments and study issues regarding potential bias. They specifically consider the issue of the agent hedging her prediction with an additional action in their experiment. They show that the prediction and the additional action are, in general, different from what each decision would be if made separately, so that hedging creates an additional source of bias in the reported probabilities. The positive correlation between the amount bet on an event and the bias in the reported probability for that event provides evidence of hedging.

Overall, it seems that no ideal solution has been found regarding hedging when beliefs are elicited. Even not paying for beliefs might still induce some hedging, as pointed out by Blanco et al. (2010).[5] Researchers should be cautious about hedging with belief elicitation. The standard method of only eliciting beliefs when the game is over is reasonable, but one must still be concerned about the actions in the game affecting (stated) beliefs. To that end, it makes sense to not divulge the results until after the belief elicitation. Paying randomly for either a belief or an action (as suggested in Blanco et al., 2010) may not be perfect, but it may well serve to substantially decrease the severity of the problem.

2.2. Simpler mechanisms

2.2.1. Introspection

One approach to eliciting beliefs is to simply ask for them. Participants report the probability of E without any reward for the accuracy of their report. This method is widespread in surveys, where incentivization is typically difficult (Manski, 2004; Hurd, 2009), and is sometimes used in experiments (e.g. Bosman and van Winden, 2002).[6]

Introspection is simple, fast, and inexpensive. However, there is no guarantee that people will state the true subjective probability accurately. Since there are no effort-boosting extrinsic incentives (if one needs to exert effort to engage cognitively), we may worry that participants report defaults, salient probabilities such as 0, 0.5, or 1, or answer randomly. They could also deliberately bias their reports to, say, justify their prior behavior or boost their image in experimenter's eyes (see Zizzo (2010) for a discussion of experimenter demand effects).

One might also expect that biases (such as ex-post rationalization) would be ameliorated by incentives, but there appears to be little evidence that this is the case. One could argue that there are already incentives present, if the marginal utility gained from demonstrating (to the experimenter or even to oneself) that one made the right decision is strong relative to the monetary benefit from answering truthfully, the incentivization would make no difference. In addition, there is also the difficult issue of whether the act of (incentivized) elicitation affects beliefs itself. Bauer and Wolff (2018) provide a nice discussion of different biases in reported beliefs with different frames of incentivized elicitation.[7] They consider three different frames (opponent, random-other, and population) and four psychological processes (ex-post rationalization, wishful thinking, hindsight bias, and consensus bias). The results indicate that different processes are triggered by different frames, so that there may not be an "innocent" or "neutral" way in which to elicit beliefs. They describe a de-biasing process by which they eliminate the treatment effects, but caution that this is a rather thorny issue. This is certainly an issue to consider for future experimental designs.

Experimental economists tend to be skeptical about introspection. A case in point is the considerable early criticism of unincentivized experiments (see e.g. Smith and Walker (1993)). Without pecuniary incentives, carelessness and unobserved motives may bias results. But this is ultimately an empirical question and may depend on the situation.

2.2.2. Frequency method

When beliefs are elicited using the frequency method, participants are rewarded with a fixed prize F if and only if their estimate exactly matches the empirical frequency. This means if participants report that six out of 10 draws will result in E, they are rewarded only when this turns out to be the case empirically.[8] We feel that this approach is quite simple, since it only involves natural inumbers and there is no calculation involved.

[5] From p. 435: "In social-dilemma experiments, however, stated beliefs may well be used as a justification of selfish behavior-making it difficult to distinguish whether true beliefs are correlated with subjects' own behavior or whether the non-incentivized stated beliefs were biased in order to justify subjects' own behavior. Moreover, subjects might play a hedge-like strategy simply because they want to be right at least some of the time. Not paying for beliefs will not eliminate such psychological hedging." It should be noted, however, that we are not aware of any empirical evidence that demonstrates this.

[6] Another common method is to pay some randomly-drawn participants a substantial flat fee.

[7] We thank an anonymous reviewer for this reference.

[8] A careful reader will notice that there is a formal difference between the belief elicited by the frequency and interval methods, and those elicited by other methods discussed in this section. In particular, other methods described in this section elicit the subjective probability that a draw of a random

Besides its simplicity, the main advantage of this method is its theoretical robustness (see Schlag and Tremewan, 2021 for detailed discussions, and Hurley and Shogren, 2005 for an application). When beliefs about a random variable V are elicited, it is optimal for participants to report truthfully, that is report $r = p$.[9] In more general settings, it is incentive-compatible to report the mode of the subjective probability distribution.[10] This is the case regardless of risk-preferences and probability weighing. The peaked nature of the incentives may offer a strong motivation to think about the elicitation problem and report carefully.

On the other hand, if N is large, rewarding correct guesses would only result in a very low probability of winning regardless of the report. This may lead participants to care about the elicitation task less. Choosing a suitably-sized N may, therefore, prove important in applications.[11]

2.2.3. Interval method

Participants are asked to report the percentage of N draws of the random variable of interest that result in outcome E (see e.g. Dufwenberg and Gneezy (2000) and Charness and Dufwenberg (2006) for applications).[12] They are then rewarded if their guess is within some number of percentage points of the truth. In other words, this method is the same as the frequency method, only with a confidence band around the point estimate of p. This method is still fairly simple, but it does require just a slight knowledge of probability. Still, the approach is very easy to understand and few people seem to have problems in practice.

Creating a confidence interval around $\hat{p}$ tackles the low probability of winning under the frequency method if N is large. This comes at the cost of some theoretical robustness. If this method is used to elicit beliefs from a skewed probability distribution, for example, it may be optimal to over-report or under-report to maximize the probability of winning.

The previous mechanisms in this section can be used when the truth is verifiable, the most common case. However, there is also at least one relatively simple mechanism for determining beliefs in some cases even when the truth is unverifiable.

2.2.4. Belief-coordination games

In belief-coordination games, people are rewarded if their report matches the average or modal report of the (other) players. As far as we know, this method was first suggested in Xiao and Houser (2005). It was used in Houser and Xiao (2011) and later popularized by Krupka and Weber (2013).

Xiao and Houser (2005) ask people to assign a message to one of three categories. Three messages (out of 75) were randomly-drawn for payoff purposes at the end; if a subject's evaluation matched the modal evaluation, she earned additional money. Houser and Xiao (2011) used a very similar approach with classifying the messages from Charness and Dufwenberg (2006). Krupka and Weber (2013) also give subjects incentives to match the modal response provided by others rating the same choice environment.

Agents are incentivized to report truthfully if they believe that their true belief p corresponds to their expectation of the average or modal report of other agents. In most experimental settings, this is an unrealistic assumption and subjects may severely bias their reports.[13] In some settings, however, this assumption is more justified. In particular, the method seems relatively well-suited to assessing beliefs about social norms.[14]

Of course, another concern is that there are multiple equilibria in coordination games. Still, if the observed reports are highly concentrated in just one region, this would suggest that a multiplicity is not present. So, this theoretical concern may not be a problem in applications, particularly since the truthful equilibrium would seem to be rather focal here.

2.3. More complex mechanisms

2.3.1. Outcome matching

Participants are offered a lottery that pays a fixed reward F if and only if the outcome of the random variable of interest V is E (Kadane and Winkler, 1988; Heinemann et al., 2009). They are shown a table, where each row represents a choice between this lottery, and a certain payment $C(r)$. $C(r)$ starts at 0 and increases in each row until it reaches F. The participant

variable V results in outcome E. By contrast, the frequency and interval methods ask participants to guess the frequency of outcome E if V is drawn repeatedly with replacement. More discussion of the difference between the two settings to come in Section 4.

[9] To be more precise, it may not be possible to make reports that correspond to p because the reports can only be made in whole numbers. But it is always optimal to make a report that is close to the true belief. For example, if the true p is 0.67 and N is equal to 10, an expected-utility maximizer would report that 7 out of the 10 draws will result in E.

[10] By a general setting, we mean a situation where a single belief parameter p does not describe the whole subjective distribution of frequencies. Example: An agent believes that there is a 50 percent chance that exactly 2 out of 10 coin tosses result in heads, and a 50 percent chance that exactly 4 out of 10 coin tosses result in heads.

[11] As noted by an anonymous referee, choosing N = 100 (or, we would add, N = 10) can make this mechanism particularly simple to understand.

[12] See footnote 12.

[13] For instance, suppose we ask subjects to report their subjective probability that god exists and reward them if their report matches the average report. Both atheists and highly religious individuals would have a tendency to bias their report towards 0.5 because they probably expect to hold more extreme beliefs than average.

[14] For example, suppose that the experimenter is interested in subjects' beliefs about the average population subjective probability that god exists. In this setting, it makes sense to reward participants' guesses according to their accuracy relative to the average guess.

is asked to choose between $C(r)$ and the lottery in each row knowing that one of these will be randomly chosen for payment. The switching point identifies the participant's certainty equivalent of the offered lottery.

The switching point, however, is not enough to pin down beliefs exactly, because the participant's choices also depend on her risk preferences (the measurement of which is subject to error). For instance, if we assume that the agent is risk-neutral, then the true belief corresponds to the value of p that equates the expected payoff of the lottery with the switching point $C(r^{switch})$. This is clearly a major issue, because there is a great deal of evidence showing that people are not risk-neutral in experiments (Rabin, 2000). Risk preferences can be estimated via additional measurements (more details on this in the proper-scoring-rules section below), but this makes the experiment longer, more expensive, and therefore distracts from the task of interest.

2.3.2. Probability matching

Probability matching (DuCharme and Donnell, 1973; Karni, 2009; Holt, 2006; Grether, 1981; Allen, 1987) is known under several names (most notably the revealed probabilities method, the stochastic Becker-DeGroot-Marshak mechanism and the Karni mechanism) and is implemented in a variety of ways. It uses randomizations to ensure robustness to violations of risk neutrality.

Probability matching consists of participants choosing between an outcome lottery A, which pays a fixed price F if and only if event E takes place, and a constructed lottery $B(r)$, which pays the fixed price F with probability r. This choice is repeated for all r's of interest (e.g. $r = 0$, $r = 0.01$, …). Expected utility maximizers will choose lottery $B(r)$ whenever $r > p$, and lottery A otherwise.[15] This ensures that she always chooses the option that gives her the largest probability of winning. Hence the switching point enables the researcher to identify participants' beliefs.

There are several ways in which the probability matching procedure has been implemented in the literature. The *multiple price list (MPL) implementation* consists of the method described above, in which participants choose either lottery $B(r)$ or A for each considered r. One of their choices is then randomly selected for payment (Baillon and Bleichrodt, 2015). The *sequential draws implementation* asks participants to report r directly (Holt and Smith, 2009). After participants report r, a uniformly distributed random variable is drawn from the interval from zero to one. If the draw is below r, the participant is paid according to the outcome lottery-she receives the fixed price F if event E takes place. If the draw is above r, the participant is paid according to the constructed lottery $B(r)$-she receives the fixed price F with probability equal to the draw. This implementation is often accompanied with figures used to explain the procedure better (see Fig. 2).

Finally, some have implemented this procedure in the form of a *second price,* or an *English auction* where participants "bid" for the outcome lottery A (Karni, 2009; Hao and Houser, 2012). Participants are asked to submit a bid r on the interval from 0 to 1. Their "opponent" is a computer that submits a bid drawn from the uniform distribution on this interval. If participants understand auctions, this mechanism is equivalent to the procedures described above. The disadvantage of using an English auction rather than a second-price auction is that the procedure sometimes terminates before participant indicates r (i.e., the computer drops out of the auction before the participant does). However, an English auction is relatively easy to understand.[16]

The main advantage of probability-matching elicitation procedures is their theoretical robustness. As already mentioned, it is robust to deviations from expected payoff maximization, as well as some deviations from expected utility maximization. For example, non-linear weighting of subjective probabilities still enables the identification of the subjective belief. However, the mechanism is not robust to violations of probabilistic sophistication. Suppose, for example, that a participant is ambiguity-averse and prefers known unknowns to unknown unknowns. This participant prefers lottery B(r) to A even for some range of $r < p$, because the participant only estimates that the probability of event E is p but does not know this with certainty. This is a common problem for elicitation mechanisms that rely on randomization.

A second issue lies in the complexity of the mechanism. If participants do not understand the randomization that underlies the mechanism well, it is unlikely that their response will reliably correspond to their true belief. Here, the fine print of the implementation matters a great deal. Complexity is perhaps the least worrying for the multiple-price-list implementation. The participant makes a series of relatively straightforward choices between lotteries. The randomization then simply consists of the experimenter picking one of these choices for payment. However, we conjecture that the other two implementations – sequential draws and auction settings - require a more sophisticated understanding of statistics. To derive her optimal report, a participant needs to (i) understand what draws from a uniform distribution mean, and (ii) engage in relatively sophisticated contingent reasoning. In principle, point (i) can be addressed by providing intuitive figures (for an example, see Massoni et al. (2014)) although we might still worry that participants will struggle. Point (ii) is known to be an issue in practice (see Martínez-Marquina et al., 2019). The auction-based implementation calls additionally for a careful explanation of auctions, not the task of interest.

[15] In fact, even the assumption of expected-utility maximization is too strong. This method can also recover beliefs when participants weigh subjective probabilities non-linearly. All that is needed for this method to work is probabilistic sophistication (i.e., agents treat known unknowns and unknown unknowns equivalently) and dominance.

[16] The English auction is simpler from a formal perspective as well-unlike the second price sealed bid auction, it is obviously strategy proof (Li, 2017).

2.4. Very complex mechanisms

2.4.1. Proper scoring rules

Scoring rules represent one of the most widely-used incentivized belief-elicitation methods. Under scoring rules, agents choose from a list of lotteries (indexed by r) that specify payment $S_E(r)$ if event E takes place and payment $S_{E^c}(r)$ if event E^c takes place. Each lottery corresponds to some reported probability r. Individual scoring rules differ in the way they specify $S_E(r)$ and $S_{E^c}(r)$. A scoring rule is *proper* if reporting $r = p$ is optimal under the assumption of expected payoff maximization. In principle, there are infinitely many functions of $S_E(r)$ and $S_{E^c}(r)$ that achieve this, but only three-the *quadratic scoring rule (QSR)* (Brier, 1950)*, the logarithmic scoring rule (LogSR)* (Good, 1952; Toda, 1963) and the *spherical scoring rule (SSR)* (Roby, 1965) – have received much attention in the literature.[17]

The *quadratic scoring rule* (QSR) is by far the most popular proper scoring rule. QSR specifies $S_E(r) = \alpha - \beta(1-r)^2$ and $S_{E^c}(r) = \alpha - \beta r^2$. Some researchers consider the QSR to be rather simple, since it involves what seem to be simple calculations. Indeed, people may follow the instructions without real comprehension. But understanding why the method works is not trivial for most subjects. Intuitively, reporting a high r pays off when E takes place. Conversely, reporting a low r pays off when E^c takes place. An intermediate value for r leads to similar payoffs after both outcomes. These motives are just balanced so that reporting $r = p$ is optimal, assuming expected payoff maximization. Specifically, under the assumption of expected payoff maximization, the agent maximizes $E[S(r)] = \alpha - \beta[p(1-r)^2 + (1-p)r^2]$. Taking the FOC with respect to r shows that it is optimal to choose $r = p$.[18]

Scoring rules are often implemented in the form of a table that lists all possible lotteries, with each row corresponding to a particular value of r. Fig. 1 provides an example from Sonnemans and Offerman (2001). A participant seeing this table is asked to select one row. The rows are often labeled with the corresponding probabilities or "chances-in-100″. Sometimes, they are simply presented as lotteries with no reference to the probabilities they represent. Other applications have given participants the formula corresponding to the scoring rule and asked them to state r directly.

Scoring rules are quite flexible and can be adapted to more general settings than the binary random variable set up discussed here. This may have contributed to their popularity in applications. Their prominence in experimental applications may also be more prosaic-the method first appeared in the work of Brier (1950) and benefited from early adoption in other experiments, so its prominence may simply reflect longevity.

Despite their popularity, (proper) scoring rules are subject to two major disadvantages. First, they are not robust to violations of the stringent assumption of expected payoff maximization (Winkler and Murphy, 1970; Schlag and van der Weele, 2013). Risk-averse agents, for example, have an incentive to report probabilities closer to 0.5 to reduce the variance of their payoff. Some researchers have noted that lower stakes may lead to approximately risk-neutral behavior. But this goes against the very purpose of incentivizing beliefs in the first place – the lower the stakes, the smaller the incentive to report truthfully.[19] Besides, empirical evidence suggests that participants are risk-averse even when stakes are small (Holt and Laury, 2002).

Another proposed solution consists of adding a set of measurements to estimate individual participants' deviations from risk-neutrality (Offerman et al., 2009; Andersen et al., 2013). This method is based on revealed preferences and does not require subjects to fully understand properties of the QSR[20] nor does it require expected utility maximization. The idea is to ask subjects to report a belief (choose a lottery in the table) for an uncertain event and to then see which belief the subject reports for objective probabilities.[21] Participants' reports can then be adjusted to reflect their utility curvature (although of course the measurement of risk preferences is also subject to error; see Charness et al. (2013)). The adjustment can even be non-parametric if the researchers collect enough additional data. This method is not widely used, since it requires time-consuming additional measurement, which may well distract participants from the task of interest.

The second issue regarding proper scoring rules is their complexity. To illustrate this, consider the QSR as presented in Fig. 1. Here, the participants choose from a list of 100 lotteries. Even if participants seek to maximize their expected payoffs, we see no reason to think that they can intuit that reporting their true belief is optimal. Remember that changing even one entry in this table by an arbitrary amount could make truth-telling sub-optimal for some subjective beliefs. It would, therefore, require a lot of effort to truly make sure that the particular payoff configuration in the table makes truth-telling incentive compatible.

[17] Gneiting and Raftery (2007) show that any positive affine transformation of a proper scoring rule is also a proper scoring rule, and so is any convex combination of two proper scoring rules.

[18] Similar proofs of the incentive-compatibility of truth-telling go through for LogSR and SSR. These proper scoring rules are often contrasted with the *linear scoring rule* (LinSR), which is *not* proper. LinSR specifies $S_E(r) = \beta(1-r)$ and $S_{E^c}(r) = \beta r$. Under the assumption of expected payoff maximization, it is optimal to make boundary reports (either 0 or 1). LinSR is incentive-compatible if people happen to have log-utility, but not typically otherwise (in general under LinSR, a particular misreporting strategy may strictly dominate truth-telling).

[19] Arrow (1971) showed that when stakes are sufficiently small, expected utility maximizers are arbitrarily close to risk-neutral when their utility functions are differentiable (p.100). Ramsey (1926) mentions that paying small stakes may alleviate the issue of risk aversion in the context of belief measurement, but also notes that "the measurement is spoiled by introducing the new factor of reluctance to bother about trifles".

[20] It does require enough understanding for participants to make consistent decisions in the belief-elicitation stage and in the additional measurements stage. If the participant is so confused that she chooses her responses arbitrarily, the method will clearly fail.

[21] Offerman and Palley (2016) present another method to calibrate beliefs without the need for sophisticated understanding by the subjects.

Payoff table

Reported probability BLUE	Payoff if outcome is BLUE	Payoff if outcome is YELLOW	Reported probability BLUE	Payoff if outcome is BLUE	Payoff if outcome is YELLOW
0%	0	10000	51%	7599	7399
1%	199	9999	52%	7696	7296
2%	396	9996	53%	7791	7191
3%	591	9991	54%	7884	7084
4%	784	9984	55%	7975	6975
5%	975	9975	56%	8064	6864
6%	1164	9964	57%	8151	6751
7%	1351	9951	58%	8236	6636
8%	1536	9936	59%	8319	6519
9%	1719	9919	60%	8400	6400
10%	1900	9900	61%	8479	6279
11%	2079	9879	62%	8556	6156
12%	2256	9856	63%	8631	6031
13%	2431	9831	64%	8704	5904
14%	2604	9804	65%	8775	5775
15%	2775	9775	66%	8844	5644
16%	2944	9744	67%	8911	5511
17%	3111	9711	68%	8976	5376
18%	3276	9676	69%	9039	5239
19%	3439	9639	70%	9100	5100
20%	3600	9600	71%	9159	4959
21%	3759	9559	72%	9216	4816
22%	3916	9516	73%	9271	4671
23%	4071	9471	74%	9324	4524
24%	4224	9424	75%	9375	4375
25%	4375	9375	76%	9424	4224
26%	4524	9324	77%	9471	4071
27%	4671	9271	78%	9516	3916
28%	4816	9216	79%	9559	3759
29%	4959	9159	80%	9600	3600
30%	5100	9100	81%	9639	3439
31%	5239	9039	82%	9676	3276
32%	5376	8976	83%	9711	3111
33%	5511	8911	84%	9744	2944
34%	5644	8844	85%	9775	2775
35%	5775	8775	86%	9804	2604
36%	5904	8704	87%	9831	2431
37%	6031	8631	88%	9856	2256
38%	6156	8556	89%	9879	2079
39%	6279	8479	90%	9900	1900
40%	6400	8400	91%	9919	1719
41%	6519	8319	92%	9936	1536
42%	6636	8236	93%	9951	1351
43%	6751	8151	94%	9964	1164
44%	6864	8064	95%	9975	975
45%	6975	7975	96%	9984	784
46%	7084	7884	97%	9991	591
47%	7191	7791	98%	9996	396
48%	7296	7696	99%	9999	199
49%	7399	7599	100%	10000	0
50%	7500	7500			

Fig. 1. (Sonnemans and Offerman, 2001; experiment 1).

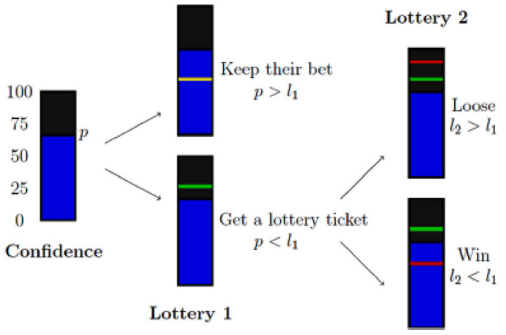

Fig. 2. (Massoni et al., 2014).

Researchers have addressed this complexity in several ways. Some experiments provide participants with the mathematical formula used to derive the rule. Many even explicitly state that truth-telling maximizes expected payoffs and offer the formal proof upon request. However, proving incentive-compatibility calls for comprehension of calculus, statistics, and the

concept of strategic dominance, which is not typical among the participant body.[22] Moreover, it distracts participants from the task of interest. We may hope that participants will simply believe the experimenter when she says that truth-telling maximizes expected payoffs; but this argument implies that the validity of scoring-rule-based elicitation rests on the credibility of the researcher rather than the theoretical properties of the incentivization method. In this case, it seems odd to claim that the form the scoring rule takes really matters for reliable belief elicitation.

2.4.2. Binarized scoring rule

As explained above, when beliefs are elicited using proper scoring rules, violations of risk-neutrality make truth-telling a sub-optimal strategy. The *binarized scoring rule* introduces an additional layer of randomization to scoring rules to tackle this issue (Schlag and van der Weele, 2013; Hossain and Okui, 2013).

BSR elicitation proceeds as follows: As with proper scoring rules, participants first choose from a list of lotteries (indexed by r) that specify "BSR payoff" $S_E(r)$ if event E takes place and $S_{E^c}(r)$ if event E^c takes place. After this choice is made, the participant observes the outcome of the random variable and is paid the corresponding BSR payoff.

This is where the fundamental difference between BSR and proper scoring rules lies. Whereas proper scoring rules pay the participants in dollars at this stage, BSR participants are paid in lottery tickets. Each of these tickets gives them a one-in-N chance of winning a fixed monetary prize-the more tickets they have, the higher the chance of winning. Any expected-utility maximizer, irrespective of her specific risk preferences, wants to maximize her chance of winning-therefore, any such individual will report a belief that will maximize her expected number of lottery tickets. If the BSR payoffs are generated by a proper scoring rule, then the expected-lottery-ticket-maximizing report is the truth, $r = p$. Hence, truth-telling is incentive-compatible even for risk-averse or risk-loving individuals.[23]

In practice, the BSR is usually implemented as follows. (1) Participants choose their most-preferred lottery. (2) The outcome of the random variable of interest is observed and participants earn a score, $S_E(r)$ or $S_{E^c}(r)$, in points on a scale from 0 to x. (3) A new random variable, distributed uniformly on the interval from 0 to x, is drawn. If the realization of this random variable is lower or equal to the participant's score, then the participant receives reward F. She receives no reward otherwise.

Clearly, the theoretical robustness of BSR comes with further complexity costs compared to proper scoring rules. Not only do participants have to intuit that reporting truthfully maximizes their expected number of lottery tickets by examining the list of available bets or the payoff formula, they also have to understand the additional layer of randomization this procedure involves. They also need to be able to reduce compound lotteries, which is known to be an issue (Halevy, 2015). Moreover, there is long-standing doubt over whether payments in lottery tickets actually manage to produce risk-neutrality (Selten et al., 1999). A study by Danz et al. (2020) reinforces this argument. The authors demonstrate that explaining the quantitative incentives involved in BSR precisely creates similar deficiencies in elicited beliefs to those elicited using QSR.

2.4.3. Elicitation when truth is not verifiable

The validity of the incentivized elicitation methods described above relies on the assumption that the outcome of the random variable V is verifiable-the experimenter can observe the outcome of a coin toss, record the strategic behavior of an experimental participant, or verify the result of an election. In some applications, however, the experimenter may be interested in eliciting beliefs about random variables that are non-verifiable by their very nature. For instance, the probability p of outcome E may correspond to the probability that an asteroid impact will end life on Earth within the next 50 years,[24] the probability that god exists, or the probability that we live in a computer simulation. Prelec (2004) demonstrated that incentivizing truthful responses is possible for these beliefs as well, at least in principle.

There is an expansive, mostly theoretical literature spurred by Prelec's original paper that spans economics and computer science. Here, our goal is to highlight the key methods that could be of use to experimentalists. We aim to provide an informal and intuitive account of these methods. Technical details can be found in the cited papers.

As we will discuss, methods that have been considered in this literature are very complex, so much so that applications typically make no attempt to explain how payoffs are actually determined to participants. We see this as a problem; and an opportunity-an easy-to-understand incentivized elicitation mechanism in settings where outcomes are not verifiable would be valuable for researchers. Choice matching (Cvitanić et al., 2019) seems like a step in the right direction. We look forward to future research in this area.

2.4.3.1. Bayesian truth serum

When the outcome of the random variable of interest V is unknown to the experimenter, she loses a vital source of information that is useful in designing elicitation mechanisms. The key insight of Prelec's (2004) *Bayesian Truth Serum (BTS)* is that one can be compensated for this by appropriately collecting additional data from the study's participants.

[22] Technically speaking, it is incentive-compatible for participants to learn calculus, game theory and statistics when confronted with the belief-elicitation task. We suspect that the typical incentives are not strong enough to ensure participants do this, however.

[23] Note that truth-telling is also incentive-compatible for individuals who weight their subjective probabilities non-linearly. The key necessary assumptions are probabilistic sophistication (i.e., equivalent treatment of known and unknown probabilities) and dominance (i.e., the assumption that a lottery that wins a fixed prize F with probability p_1 is preferred to a lottery that wins F with probability p_2 if and only if $p_1 \geq p_2$) (Hossain and Okui, 2013).

[24] Whilst this event is verifiable in a very real sense, we would suggest that participants will generally not find the promise of payment in the event of the Earth's destruction to be credible.

In particular, each subject makes two reports under BTS: (1) a subjective probability or "opinion" report r and (2) an estimate of the opinion distribution among other participants. The latter estimate (2) consists of one prediction y_r^i for every admissible value of r, where y_r^i corresponds to i's reported expectation of the proportion of all participants in the study that claim to hold opinion r. For example, i may report that her opinion is $r = 0.5$ and claim that she expects that 20% of participants report $r = 0.1$ ($y_{0.1}^i = 0.2$) and the rest report $r = 0.9$ ($y_{0.9}^i = 0.8$). The actual proportion of all participants who report opinion r is equal to x_r. The participants' collective average prediction of this proportion is $\bar{y}_r$.[25]

After all reports are collected, each participant receives a payment equivalent to the sum of two scores. The *prediction score* rewards each participant i in proportion to the accuracy of her prediction reports y_r^i compared to the actual proportion x_r. This is done using a proper scoring rule. As a result, the agent maximizes her expected prediction score by reporting the expected opinion distribution truthfully. Additionally, if participant i reports that her opinion is r, she receives an *information score* equal to:

$$\log\frac{x_r}{\bar{y}_r}$$

The numerator corresponds to the empirical frequency of the participant's own report r in the population, and the denominator to the collective prediction of this frequency. It follows that the participant who wants to maximize her information score seeks to report the opinion r whose true popularity is collectively under-estimated by others. If certain assumptions hold, the participant thinks that this *most surprisingly common report* corresponds to her true belief, leading her to truthfully report $r = p$.

The first three necessary assumptions are (i) expected payoff maximization,[26] (ii) a large enough sample size,[27] and (iii) the assumption that all agents believe that others are reporting truthfully.[28] Additionally, Prelec (2004) makes two assumptions that relate to the belief-formation process itself. Some version of these assumptions is central not only to BTS, but also to most other elicitation mechanisms described in this section, so it is worthwhile to discuss them in some detail.

Informally, it is key that every agent treats her opinion as a signal about the opinion distribution in the population – if agent i holds an opinion, this signals to her that others do too. Suppose, for example, that we ask agents A and B to report their subjective probability p that god exists. Suppose further that person A is atheist-leaning and thinks that $p = 0.1$ whilst person B is religious and thinks that $p = 0.9$. Roughly speaking, if both think of their personal opinions as signals that others share their view, then person A's estimate of the proportion of people with opinion $p = 0.1$ is higher compared to that of person B. From A's perspective, person B is underestimating the share of people with opinion $p = 0.1$. It follows that A thinks that the population on average underestimates the popularity of opinion $p = 0.1$-she expects this to be the most surprisingly common opinion. This makes $r{=}p$ the payoff-maximizing report for A.

Formally, the argument in the preceding paragraph goes through if two formal assumptions hold: (1) agents' views on the opinion distribution in the population are consistent with updating from a common prior, and (2) agents A and B agree on the population opinion distribution if and only if they have the same opinion p. Combined, these assumptions imply that personal opinions provide information about the population-opinion distribution, but that the informativeness of each opinion is not individual-specific – i.e. if agents A and B agree on p, their conclusion on the population-opinion distribution is equivalent, because p provides the same information to both A and B.

These assumptions are stringent and may be unrealistic in many applications.[29] Most fundamentally, do agents really treat their opinion as a signal about the population-opinion distribution? It is true that people generally believe that their own opinion is more common than it actually is (see the literature on the false consensus effect Marks and Miller (1987), Ross et al. (1977), see Dawes (1990) for the Bayesian explanation of this phenomenon). However, this alone is not sufficient to make truth-telling incentive compatible under BTS. Individuals also must be aware of others' tendency to overestimate the popularity of their opinions. Only then would agent i conclude that the collective prediction of the population-opinion distribution will underestimate the frequency of i's opinion.

BTS is very complex. Even if the participants behaved in line with the assumptions outlined above, we doubt that they can intuit that truth-telling is optimal just by looking at the score functions alone. Moreover, the proof of the incentive-compatibility of truth-telling under BTS requires a sophisticated understanding of statistics and game theory. This complexity is typically even greater in extensions of the standard mechanism that are robust to some deviations from the baseline assumptions.

Prelec (2004) is highly aware of this complexity. He suggests telling the participants that it is in their best interest to report the truth, without explaining the precise reasons for this. This is what is typically done in applications. For example,

[25] The average prediction $\bar{y}_r$ is actually a geometric average $\log \bar{y}_r = \sum_{i=1}^{n} \log y_r^i$. This is a technical requirement.

[26] In theory, this could be remedied by paying participants in probability rather than in money. In practice, we would expect this remedy to be as ineffective as it is in other applications (Selten et al., 1999; Danz et al. 2020).

[27] Variants of BTS that work with small sample sizes have since been proposed by computer scientists (Witkowski and Parkes, 2012; Radanovic and Faltings, 2013, 2014).

[28] Other equilibria also exist.

[29] For instance, it is unreasonable to assume that a census official and a bus driver have the same beliefs about the population-opinion distribution about the existence of god, even if they agree that the probability that god exists is p = 0.9. Several variants of BTS that are robust to some deviations from Prelec's original assumptions have since been proposed (Radanovic and Faltings, 2014). These mechanisms still assume that people with the same opinion p give similar estimates of the population opinion distribution, even though they do not have to be exactly equal.

Weaver and Prelec (2013) tell participants in their experiment that "Truth Scoring [was] recently invented by an MIT professor and published in the academic journal Science" and that participants "are most likely to maximize [their] earnings if [they] answer every item truthfully".

We view this approach as problematic for two reasons. First, the statement above is true only if all of the stringent assumptions above hold-if not, it may be viewed as deceptive. Second, the reliability of belief elicitation now depends on the credibility of the experimenter, not the properties of the mechanism.

2.4.3.2. Peer-Prediction method

Like BTS, the peer-prediction method assigns scores to participants' estimates of the population-opinion distribution (Miller et al., 2005). Unlike BTS, however, the peer prediction method does not elicit reports y_r^i about this distribution, but rather posits that the experimenter already knows this distribution for each admissible true belief p.[30] The experimenter then simply asks participants to make a report r about their subjective probability or opinion p. The experimenter subsequently computes i's prediction score for the implied opinion distribution conditional on i's reported opinion r. If the prediction score is calculated using a proper scoring rule and the participants maximize their expected payoff, there exists a truth-telling equilibrium.

The clear issue with this mechanism is that the experimenter generally does not know the actual opinion distribution of the agents. This is why BTS elicited two separate reports from each participant in the first place. Some research has described how these opinion distributions could be estimated through distributional assumptions or machine-learning methods (Radanovic and Faltings, 2015; Radanovic et al., 2016; Shnayder et al., 2016; Agarwal et al., 2020; Liu and Chen, 2017). Additional reports may also be collected in a sequential setting to inform the estimation of subjective opinion distributions (Witkowski and Parkes, 2012). In many applications, however, reasonable estimates of the opinion distributions are not available, and the method is of little use.

Besides this major issue, the peer-prediction method suffers from virtually all of the theoretical flaws of BTS.[31] Importantly, even though participants typically only report their opinion p under the peer-prediction method, this elicitation mechanism hardly tackles BTS's complexity and its lack of transparency - a full understanding of the mechanism would require an appreciation of how the experimenters go about estimating an agent's beliefs about the opinion distribution. It is unrealistic to expect this from most typical experimental subjects.

2.4.3.3. Choice matching

The choice-matching mechanism (Cvitanić et al., 2019) requires each participant to make two reports: (1) a subjective probability or "opinion" report r and (2) a report that can be scored and rewarded. In principle, the latter report (2) may take many forms-in the baseline version described by Cvitanić et al. (2019), subjects report their estimate of the distribution of opinions among other participants. More specifically, each agent i reports her expectation y_r^i of the proportion of all participants in the study who claim to hold opinion r for each considered r. Since this is also the report used in the BTS, we stick with this baseline version of choice matching for now and later explain extensions.

As is the case under BTS, each choice-matching agent receives a sum of two scores. The *individual score* is calculated using each agent's report (2). In the baseline version of choice matching, the individual score rewards the accuracy of the agent's reported population-opinion distribution-i.e. the accuracy of her reported frequencies of each opinion $y_r^i \ \forall \ r$.[32] If the agent seeks to maximize her expected payoff, she reports her true estimate of the population-opinion distribution.

The agent's second score corresponds to the average individual score of all other agents that reported the same opinion r as agent i. Call this the *group score*. If the agent believes that like-minded people are also more likely to receive high individual scores, she should report truthfully that $r = p$. In addition to expected payoff maximization and the belief that others are also truthful, this simply requires that each agent thinks that like-minded individuals are more likely to be "right" than others.

As was already insinuated, it is not necessary that the agents' individual and group scores are calculated using estimates of population-opinion distributions. In fact, these scores may be completely unrelated to the elicitation task. For instance, all participants can take an intelligence test. The individual score of agent i then corresponds to i's performance on this test. The participant i's group score corresponds to the average performance of all agents who reported the same opinion r as agent i. Each participant is then incentivized to report truthfully if she believes that like-minded individuals are smarter than others. The "scores" may also correspond to different kinds of gifts (Cvitanić et al., 2019): Suppose that participants are asked to assess the probability of purchasing a new product offered by a firm. In addition to making the required report r, each agent chooses one of the company's current products. She then receives this selected product (=individual score) or the product selected by a random person who reported the same opinion r (=group score). If each agent believes that like-minded individuals are likely to have similar preferences, truth-telling is incentive compatible.

[30] More precisely, Miller et al. (2005) assume that the experimenter knows the common prior about the population opinion distribution. Under the assumption of Bayesian updating with opinions as informative signals, the experimenter can specify the posterior belief about the opinion distribution conditional on each admissible opinion.

[31] As with BTS, there also exist variants that are robust to deviations from some baseline assumptions (e.g. see Witkowski and Parkes, 2012).

[32] The authors use a proper scoring rule.

Of course, it is not always reasonable to assume that like-minded agents are more likely to be right, score higher on IQ tests or have similar tastes. For instance, if agents are asked to report their profession and then take a statistics test, all agents have an incentive to report that they are statisticians because this increases the agents' expected group score. Finding an appropriate basis for individual and group scores is, therefore, vital in experimental applications.

The precise form of the individual and group scores also determines the complexity of the mechanism itself. If participants are rewarded using complex payoff formulas, as is the case with proper scoring rules, choice matching can be complicated and difficult to understand. If participants are scored using "simple incentives" as described in section 2.2, marked using intelligence tests or receive gifts, the method becomes much simpler. If they think that like-minded individuals are alike in other respects, they may be able to intuit that truth-telling is optimal.

3. Comparisons across elicitation mechanisms

The purpose of belief elicitation is pragmatic – to provide an accurate measure of subjective beliefs. As a result, picking the best elicitation method is fundamentally an empirical issue. A critical consideration is the quality of elicited beliefs. Schlag et al. (2015) describe four metrics for assessing the quality of beliefs: Inducing beliefs and testing for accurate reporting of the induced beliefs, identifying deviations from the empirical distribution, measuring consistency with behavior, and checking for additivity (to 100%) of beliefs. Additionally, we might be interested in studying the empirical frequency of focal or default reports, or the time participants spend assessing the data to construct and update their beliefs. All of these have some merit, but all have some concerns as well. We refer the interested reader to the excellent discussion in Section 1 of their paper.

This section summarizes empirical literature comparing the performance of elicitation methods in applications and outlines the general lessons that this line of research has taught us.[33] First, we demonstrate that there is little empirical support for the belief that complex incentives do better than no incentives at all – both complex incentives and introspection generate biases in elicited beliefs. Second, while we believe that simple incentivization is effective and provides a promising road forward, more research needs to be done before drawing definitive conclusions. Despite this uncertainty, we think that it is evident that (1) the search for reliable elicitation methods is far from over, and (2) that the tendency of researchers studying belief elicitation to design ever-more theoretically-robust methods with little consideration of complexity has not led to systematic improvements in empirical belief measurements.

3.1. Introspection versus complex incentives

Economists' skepticism about unincentivized experiments has motivated the use of complex incentivized methods in belief elicitation. It is therefore unsettling to find that the literature has not demonstrated that these methods do any better than introspection empirically.

To our knowledge, there are only two studies with credible statistical analysis in which at least one complex method dominates introspection according to measures of belief quality. Wang (2011) concludes that QSR outperforms introspection when participants are asked to predict behavior in an asymmetric-matching-pennies game. In this experiment, participants first observed five periods of play of a matched pair recorded in an earlier experiment. They then indicate the likelihood that their counterpart chooses a given action for another five rounds.

The author uses three measures to argue that QSR performs better than introspection. First, she records the absolute distance of the elicited probability from 50 percent to create a measure of "extremeness" and shows that QSR's extremeness first-order stochastically dominates introspection's extremeness. In other words, beliefs elicited via introspection are closer to 0.5 than those elicited via QSR. This, the author states, provides an indication that QSR beliefs are less random. Second, the author shows that, though economically small, the correlation between beliefs about actions and actual behavior is statistically significant and positive only for QSR beliefs. Third, the author runs a random-effects regression of actions on stated beliefs, showing that only QSR beliefs are better calibrated than random guesses. This deviation from randomness was statistically significant, though not economically sizeable.

We think that this evidence is not quite a ringing endorsement for QSR over introspection. The fundamental problem is that the tendency of introspection beliefs to appear "more random" may simply reflect the honest uncertainty of participants about players' behavior rather than the deficiencies of the elicitation method. Moreover, even if we were to accept that "less random" beliefs signal a better elicitation method, QSR does not do particularly well either. For example, the correlation coefficient between actions and QSR beliefs is only 0.077.

In a signal-detection task, Massoni et al. (2014) elicit confidence to compare probability matching (sequential draws implementation), introspection, and QSR. Participants observe two circles with a similar number of dots and indicate which circle has more dots. They then report their confidence in this guess in the form of their subjective probability of being correct. The difficulty of the task is calibrated; the participant is extensively trained and gains further experience as the experiment progresses through its 100 trials of varying difficulty, and she receives feedback on success rate and earnings. The

[33] A reviewer notes the difficulty in comparing across studies with different mechanisms due to the high variance in belief reports and the typically low power of the studies involved. Accordingly, we do not attempt rigorous statistical comparisons across these studies.

authors assess the performance of the methods by constructing a "signal detection (SD) agent"-a model Bayesian agent that receives noisy Gaussian signals about the true number of dots in each circle and always picks the circle that is more likely *a posteriori*. Careful pre-testing makes it possible to calibrate the task for individual participants to permit constructions of this SD agent. Given the sequence of trials, it is possible to exactly specify the implied confidence rating distribution and the receiver operating characteristic (ROC) curve for the SD agent.[34] These measures are then compared to their empirical counterparts for each elicitation method.

Both the empirical confidence rating distribution and the ROC curve of probability-matching beliefs are closer to the SD predictions when compared to introspection. The introspective beliefs also reveal that a non-negligible number of participants were lax with their responses, sticking to the default confidence rating of 75%. It should be noted that the lengthy training and experimental session along with extensive feedback may have helped participants gain a good grasp of the complex sequential-draws implementation of probability matching. For introspection, the length of the task is a drawback, perhaps leading to a loss of interest and motivation. Moreover, as the authors acknowledge, the results are only as robust as the SD model they use. Still, this result should not be overlooked.

Although probability matching appears to have elicited beliefs well in this experiment, QSR did not. Contrary to Wang (2011), QSR performed worse than introspection along almost every measure used. The poor performance stemmed from large peaks in the confidence level distribution at 50% and 100%. The former peak can be explained by risk-aversion. The latter is hard to square with anything but confusion. This confusion may have been strengthened by the fact that individual QSR bets were not indexed by corresponding confidence levels, but rather by letters, which is usually not the case in applications. Nevertheless, this would suggest that the incentive-compatibility of truth-telling under QSR is so obscure to experimental participants that even long experiments and training sessions do not help them internalize it.

The preceding two papers can be contrasted with the preponderance of research on this topic, which paints an even more ambivalent picture of the benefits of complex incentives. For instance, Sonnemans and Offerman (2001) ask people to guess the proportion of a wheel of fortune that is blue. The composition of this wheel is unobserved and changes in each period. Before making their guess, participants may spin the wheel up to 20 times and observe the outcome, so there is a direct measure of both effort and belief quality. Beliefs are elicited using QSR and introspection. Average reported beliefs are virtually identical in the two treatments. Furthermore, if participants in the introspection treatment were rewarded for their reports using QSR, their earnings would have been very close to the actual earnings in the QSR treatment.

Moreover, while participants ask for slightly more draws in the initial periods with introspection, this insignificant difference quickly disappears as the experiment progresses. So, QSR participants do not seem to put more effort into the task than the introspection participants.

Trautman and van de Kuilen (2014) conduct an extensive examination of elicitation mechanisms, covering QSR, introspection, and outcome matching. They also estimate risk-aversion and probability-weighing corrections to QSR and outcome matching with the help of additional measurements. Participants initially play an ultimatum game with six possible discrete allocations. Before this game is resolved, proposers report their beliefs about the acceptance probability for each possible proposal. Responders state their beliefs about the chance that each allocation was proposed. To check additivity, proposers (responders) also report their belief that one of the proposals (allocations) is not accepted (proposed).

In terms of additivity, introspection fares better than incentivized elicitation methods. That is, the sum of the elicited probabilities for complementary events was closer to 100% for introspection, and the share of perfectly and well-calibrated participants was also larger for this elicitation method. On the other hand, beliefs elicited via incentivized methods were more in line with proposer's optimal actions judged through the lens of expected payoff maximization, expected utility maximization with CRRA utility, and expected utility maximization with Fehr and Schmidt (1999) social preferences. Even incentivized methods only predict around 30% of the choices, however, which could mean that these elicited beliefs are not of a particularly high quality, or that the behavioral assumptions used to deduce optimal actions are a poor guide to actual behavior in this experiment. We note that the average belief distributions elicited using all mechanisms are quantitatively very similar, without significant differences in accuracy.

Burfurd and Wilkening (2021) compare the performance of probability matching (sequential-draws implementation) and introspection in a belief-updating bucket task. The participants face one of two buckets with different compositions of black and white balls. They know the color composition of the two buckets and are told that the prior probabilities of facing bucket A and bucket B are equal. They subsequently make draws from the bucket and report their posterior belief of facing bucket A. The results show no significant overall differences in the accuracy of beliefs elicited using the two methods. Interestingly, the authors further divide the participants into "consistent probabilistic reasoners" and "inconsistent probabilistic reasoners" in relation to grasping a relatively subtle statistical intuition in a bucket task based on Charness and

[34] To construct the ROC curve in this experiment, the authors first select some confidence level x. They then conduct the following thought experiment. Suppose that we predict that all of the agent's left or right circle guesses are incorrect if her reported confidence level is below x, and that all of her guesses are correct if her reported confidence level is above x. This classification has a certain true positive rate (the share of actual correct guesses predicted to be correct) and a certain false positive rate (the share of actual incorrect guesses predicted to be correct). This thought experiment is repeated for all $x \in [0, 1]$ resulting in a recording of the true and false positive rates for each x. The result is graphed with the true positive rate on the vertical axis and the false positive rate on the horizontal axis. This graph corresponds to the ROC curve. The ROC curve thus gives us a measure of how good a predictor the confidence level is of the participant making mistakes in the signal-detection task.

Levin (2005).[35] The sequential-draws implementation of probability matching outperforms introspection in terms of the deviations from Bayesian posteriors for individuals of high cognitive or statistical ability, but it performs no better for low-ability individuals.

Burfurnd and Wilkening (2021) further demonstrate that introspection dominates probability matching when agents receive uninformative signals, particularly for low-ability individuals. This suggests that whilst the complex sequential draws implementation of probability matching appears to outperform introspection in a difficult task (Bayesian updating), introspection dominates in a simple task (reporting a prior). Thus, task selection may well be a critical factor in choosing an elicitation mechanism.[36] Different mechanisms may very well be suited to different tasks, and this is a topic for future research.

Overall, our view is that there is no reliable evidence that complex incentivized elicitation systematically outperforms introspection in experimental applications. Since introspection is also much cheaper, simpler, and faster, it seems preferable in common experimental applications. This stands in stark contrast with the status quo in the literature. Yet, without new evidence to the contrary, we struggle to see how we could hold a different view.

This should not be understood as a ringing endorsement of introspection. Eliciting beliefs without incentives comes precisely with the kinds of problems we would expect. Burfurd and Wilkening (2021) show that high-ability participants in the introspection treatment were significantly more likely to make focal reports of 0, 50 or 100 compared to participants in the probability matching treatment. Massoni et al. (2014) find that unincentivized subjects had a higher propensity to stick with a default report. However, complex incentives come with biases of their own-biases that are no less serious than those of introspection.

3.2. More complex incentives compared

The literature on more complex incentivization points to an overall disadvantage of QSR relative to other methods, especially those that are robust to violations of expected payoff maximization. Recent work has cast some shade on the performance of the BSR, which some researchers seem to feel is the state-of-the-art belief-elicitation mechanism. Probability matching appears to be the tentative winner, but even this mechanism may prove difficult to use when training and checking participants' understanding is not feasible. Its success may also depend on the particular implementation - as already noted, our view is that the complexity of probability matching varies widely with implementation.

3.2.1. Scoring rules

Under the assumption of expected payoff maximization, all proper scoring rules should perform just as well, and outperform scoring rules that are not proper, such as the linear scoring rule. Palfrey and Wang (2009) test this by comparing the performance of the quadratic, logarithmic, and linear scoring rules (QSR, LogSR and LinSR respectively). Participants predict the actions of players from a Nyarko and Schotter (2002) asymmetric-matching-pennies experiment after observing their behavior for five periods.

A calibration regression shows that QSR and LogSR beliefs perform significantly better than random in predicting matched players' behavior, while LinSR beliefs do not. Moreover, QSR and LogSR beliefs have a positive, though small, raw correlation coefficient with choices; LinSR beliefs have a small negative correlation. This appears to reaffirm the theoretical superiority of proper scoring rules relative to improper ones. However, some caution in interpreting the result is called for because the observed negative correlation of LinSR and actions is actually relatively difficult to square with the theory. Under the assumption of expected payoff maximization, LinSR should lead to extreme (0 or 100%) predictions depending on which action the participant deems more likely. A negative correlation could arise if the participants only have a poor idea of how the Nyarko and Shotter (2002) participants behave, leading them to make mistakes very frequently. If this is the case, judging the performance of scoring rules by their correlation with actual choices is not a good indicator of belief quality.

LinSR also resulted in a more extreme distribution of guesses than QSR and LogSR, judging by the distance from the 50/50 report, and resulted in a larger proportion of participants making boundary reports (about 33%). This is in line with the theory, although it should be noted that the proportion of QSR participants making boundary reports was also relatively large (about 20%, compared to 2.5% for LogSR). Interestingly, the distributions of QSR and LogSR beliefs were different, with QSR first-order stochastically dominating LogSR.

On the other hand, Andersen et al. (2014) do not find significant differences between the proper QSR and the improper LinSR. Participants in this experiment reported their beliefs about three questions related to the 2008 US presidential election, which took place one week after the session, and about the performance of men relative to women in a psychological test for empathy known as the Eyes Test (Baron-Cohen, 2003) that they completed at the start of the session. The LinSR and QSR belief distributions were statistically indistinguishable for three out of the four questions. Moreover, the LinSR beliefs were not more likely to take extreme (0 or 100%) values. This stands in contrast to the theoretical predictions, leading the

[35] In a follow-up experiment, Burfurd and Wilkening (2021) additionally classify participants into groups of high and low fluid intelligence according to the Raven's Advanced Progressive Matrices test, and into high effort and low effort groups according to performance on the Cognitive Reflection Test.

[36] An insightful reviewer notes that, with a slight change of that design, a disingenuous researcher could write a paper where either method could outperform the other if the underlying decision task is ignored.

authors to conclude "this finding provides some support for those that would prefer to use the LinSR on the grounds that it is simpler to explain to participants than the QSR".

Andersen et al. (2014) conduct additional measurements to estimate the deviations from risk-neutrality and linear probability weighing. They show that estimating the curvature of the utility and probability-weighing function leads to statistically-significant differences in the implied subjective beliefs. The authors, however, do not examine empirical measures of belief quality, so it is unclear whether these adjustments actually improve raw elicited beliefs.

Offerman et al. (2009) analyze the effects of a similar adjustment to raw QSR beliefs. Participants in this study observe a time series of 32 stock prices and predict the probability that they fall into two small intervals and their union seven months after the evaluation date. These beliefs are elicited using QSR. Participants were paid either for all guesses (ALL treatment), or randomly for only one of their guesses (ONE treatment). Additionally, participants indicated their beliefs under QSR for about 20 events with objectively known probabilities. Theoretically, this makes it possible to establish an approximate mapping from QSR reports in the stock-price section of the experiment to true subjective probabilities.

The calibration results show that participants in the ONE treatment deviate from objective probabilities significantly in their raw reports, while the deviations for participants in the ALL treatment are much smaller. This dichotomy is also reflected in an additivity bias of the reports. There were virtually no differences in the additivity bias in the ALL treatment for adjusted beliefs relative to raw responses. On the other hand, the average additivity bias was reduced for 27 of 32 stocks in the ONE treatment when raw responses were adjusted.

Our sense is that the need for additional costly and time-consuming measurements makes the kinds of adjustments studied by Offerman et al. (2009) and Andersen et al. (2014) unpopular in applications.

3.2.2. The binarized scoring rule

The binarized scoring rule is a more widespread way to deal with scoring rules' lack of robustness to risk aversion. Hossain and Okui (2013) compare the empirical performance of the BSR and the QSR in two experiments – the "P-experiment" and the "M-experiment". Participants in the "P-experiment" report their beliefs about the probability that a ball of a certain color is or is not drawn from an urn with a known composition. The results from the P-experiment show that the squared difference between reports and the true probabilities is significantly larger under QSR than BSR. Regression analysis shows that the BSR's better performance reflects an improvement in elicited belief accuracy for risk-averse agents.[37]

The authors show that the distribution of QSR and BSR beliefs is significantly different for risk-averse agents, but not for risk-neutral or risk-loving agents. Moreover, when the true proportion of red balls was 0.25 or 0.75, the BSR belief distribution was closer to this true proportion than the QSR distribution. When the true proportion of red balls was 0.5, both mechanisms performed very well without statistically significant differences. When the true proportion of red balls was 0.1 or 0.9, both elicitation methods performed poorly with participants showing only a very limited tendency to report beliefs close to the truth. The belief distributions for this extreme composition were not significantly different under the QSR and the BSR. These measures point to an overall advantage of the BSR. But the poor performance for extreme urn compositions shows that even the BSR leaves much to be desired.

While differences between BSR and QSR are large and statistically significant in the P-experiment, this is not the case in the M-experiment. In this experiment, participants predict the realized earnings per share of stocks. For each considered stock, the participant first observes 10 independent forecast realizations drawn from a normal distribution centered at the true realized earnings per share. Subsequently, she indicates her prediction. This experiment revealed no statistically significant differences in the standard deviations from the truth for the BSR and the QSR. We note that the structure of the elicitation environment in the M-experiment actually makes it easier to intuit that truth-telling is profit-maximizing under both QSR and BSR.[38] We discuss how the elicitation environment interacts with the complexity of elicitation methods in Section 4.1.

Participants in Harrison et al. (2014) report their beliefs about the composition of ping pong balls in a bingo cage. They could observe the bingo cage for a duration of 10 spins. Their beliefs were elicited using QSR and BSR. A structural model based on expected utility theory and rank-dependent utility was also estimated to correct for risk-aversion and probability weighing in QSR. The display of the elicitation mechanism was sliders. BSR was found to perform better than QSR, although QSR adjusted for risk aversion and probability weighting was about as effective.

Danz et al. (2020) provides a striking account of the deficiencies of BSR. Participants in this experiment face either a Red urn or a Blue urn with a known prior probability. The participants then make three sequential guesses about the chance that the urn they face is Red. The first guess is made without any additional information; guesses two and three are made after observing one and two independent draws from the urn. The authors focus their analysis on the first guess because this makes it possible to study the quality of risk elicitation without the auxiliary hypothesis that participants are Bayesian.

[37] Erkal et al. (2020) reach a similar conclusion. Subjects in their experiment are told the composition of red and blue balls in an urn. They then report their belief that a red ball is drawn under QSR and BSR.

[38] In the M-experiment, the participant is penalized for the squared distance of her guess from the realized earnings per share in dollars under QSR. Under BSR, she is more likely to receives a prize if the squared distance from the realization is low. The participant may be able to intuit that it is in her interest to guess the realization as precisely as possible. This is somewhat more intuitive than the binary setting of the P-experiment – there, she is no longer guessing just the realization, but rather the probability of each possible realization.

All beliefs are elicited using the BSR. The interesting twist lies in the way BSR incentives are presented to the participants. The Information treatment presents BSR incentives precisely and carefully. The Reduction-of-Compound-Lotteries (RCL) treatment goes even further, providing participants with a calculator that allows them to explicitly calculate the total chance of winning the price for individual guesses, so that participants can verify that truth-telling maximizes the probability of winning. The No Information treatment takes the opposite approach: While participants are told that reporting the truth maximizes their chances of winning (as in the other treatments), they are given no quantitative description of BSR incentives.

The results show a large fraction of incorrect reports in the Information treatment (41.5%). This fraction was somewhat smaller, but still large, in the RCL treatment (32.5%). Interestingly, both of these treatments also display a significant dip in the fraction of false reports at the true probability of 0.5, when reporting the truth results in an equal chance of winning regardless of which urn is actually selected. The No-Information treatment performs significantly better than either of the previous two treatments. The fraction of incorrect reports is significantly smaller (20.6%) and roughly the same for all objective priors, including 0.5.

This finding means either that paying in lottery tickets fails to generate risk neutrality (as previously argued by Selten et al. (1999)) or that participants fail to internalize that truthful reporting maximizes their chances of winning, despite careful explanation. Since there is a reduction in the fraction of false information and RCL responses when the objective prior is 0.5 (and truthful reporting results in a constant probability of winning regardless of outcome), the authors argue that this lends support to the former explanation. Moreover, participants typically self-reported that they purposefully biased their responses toward 0.5 in the Information and RCL treatments to increase their chance of winning if they ended up facing the less likely urn.

3.2.3. Probability matching

Probability matching is perhaps the most empirically successful incentivized mechanism. As described in the previous subsection, Massoni et al. (2014) show that probability matching (sequential-draws implementation) outperforms both introspection and QSR in a signal-detection task. Raw confidence responses are compared to a theoretically constructed Bayesian "SD agent" with normally distributed internal signals in this paper. Holt and Smith (2016) compare QSR with probability matching: Participants face either a red cup or a blue cup with a known probability of 0.5. They observe draws with replacement from this cup and subsequently report their posterior belief that the cup they face is red (in some sessions, no balls are drawn, so participants report their prior belief).

The authors compare four treatments. Participants in the QSR-chance treatment pick a row in a standard QSR table with probability labels for each row. The QSR-numbers treatment removes the probability labels from the table. The BDM treatment uses the sequential-draws implementation of probability matching. The LC treatment is a small twist on the multiple price list implementation of probability matching.[39]

The average absolute distance from the true Bayesian posterior is smallest for the LC treatment and largest for the QSR-numbers treatment; those for BDM and QSR-chances lie between these, though only differences between the LC and QSR-numbers treatments are statistically significant. This pattern is repeated when attention is restricted to "simple" cases when the true posterior is 0.5, either because no draws are made or because the draws are balanced. Overall, this study points to the advantages of the multiple-price-list implementation of probability matching relative to QSR, but it also highlights that these advantages do not necessarily carry over to the abstract sequential-draws implementation.

Burfurd and Wilkening (2018) compare three different implementations of probability matching. As in the Holt and Smith (2016) study described above, participants face one of two buckets with different ball composition. Bucket A and bucket B are *a priori* equally likely to be selected. One ball is drawn from the selected bucket and participants subsequently report their posterior belief of facing bucket A. The first treatment consists of the sequential-draws implementation that explains the mechanism rigorously and precisely. The second treatment presents the mechanism in its multiple-price-list implementation with a small twist (this is equivalent to the LC treatments in Holt and Smith (2016)). The third analogy-based treatment, adapted from Hao and Houser (2012), presents the sequential draws implementation in simplified language using the following analogy: A number Y is then drawn from 1 to 100. If Y is greater than report *r*, the participant is paid if a black chip is drawn from a bag with Y black chips and 100 – Y white chips. If Y is lower or equal to *r*, the participant is paid if and only if the selected bucket is bucket A.

Consistent with Holt and Smith (2016)'s finding regarding its LC and BDM treatments, Burfurd and Wilkening (2018) find no significant differences between individual treatments in terms of accuracy (e.g. the absolute error of the participant's report relative to the Bayesian posterior) or precision (e.g. the standard deviation of absolute errors for each individual). However, the authors report a significant reduction in the length of the experiment in the analogy-based treatment compared to the other two treatments. Interestingly, a second experiment shows a significant reduction in the accuracy and

[39] Unlike the plain-vanilla MPL implementation of probability matching, this implementation progresses in two stages. In both stages, the participant chooses between the Red Cup Lottery that pays the prize if and only if the cup is red and each of the relevant Random Lotteries that pay the prize with probability r. In the first stage, the r's are presented in a table with r increasing in increments of 10 chances in 100 from 0 to 100. In the second stage, r's increase in increments of 1 chance in 100 for the 10-point range that corresponds to the participant's switching point in stage 1 (e.g., if the participant switched at r = 20 in the first stage, the second-stage table ranges from 10 to 20 in increments of 1).

precision of the rigorous descriptive implementation of probability matching when a pre-experimental quiz checking the participants' understanding of the mechanism is removed.

It is perhaps fitting to end this subsection by quelling some of the enthusiasm that the relatively good performance of probability matching compared to other complex methods may have generated by reminding the reader about the comprehensive study by Trautman and van de Kuilen (2014) described in the previous subsection. This study compares probability matching (multiple-price-list implementation) with introspection, outcome matching, QSR, as well as risk-aversion-robust adjustments to QSR and outcome matching. As already noted, this study reveals no systematic differences between the complex incentivized methods. Clearly, this could be due to the format of the experiment that asks the participants to predict something that is not objectively known, so similar performance could reflect true uncertainty rather than similar quality of elicitation methods. Nonetheless, this experiment serves as a reminder that belief elicitation is far from a settled methodological issue.

3.3. Comparisons with simple methods-a missing comparison

To date there is little evidence for empirical comparisons between either simple and complex incentivization or simple incentivization and introspection. We conjecture that simple incentivization will at minimum induce a smaller variance and likely result in more accurate reports relative to introspection by boosting effort and providing a counterweight to experimenter demand effects and image concerns. Unlike complex incentives, simple incentives could achieve this without confusing the participants and creating further biases in reported beliefs as a result. However, there is a surprising hole in the experimental literature on the performance of simple incentives relative to other methods.

Schlag and Tremewan (2021) provide the most explicit empirical comparison of simple and complex incentivization. This study elicits participants' beliefs using the frequency method and a complex sequential-draws implementation of probability matching in a Stag-Hunt game and an updating task with an urn.[40] In the Stag Hunt, participants in the frequency treatment guess how many out of 20 randomly chosen participants from the session play a certain action. In the probability-matching treatment, they guess the probability that a randomly-chosen participant plays this action. In the urn task, they face one of two urns with equal probability. After the urn is selected and one ball is drawn, participants in the frequency treatment guess how many out of 20 balls drawn with replacement from this urn will have the same color as the first. In the probability-matching treatment, they guess the probability that a ball randomly drawn from this urn would have the same color as the first. Additionally, participants also self-report their understanding of the task, since the instructions were not exactly user-friendly.

Most measures of the reported belief accuracy did not show statistically-significant differences between the two treatments. The exception to this finding is the proportion of participants who state the (approximately) true Bayesian posterior probability in the urn task, which is larger for the frequency method (with a p-value of 0.06). This experiment's failure to systematically observe statistically-significant differences is perhaps not particularly surprising, since participants can only guess other participants' behavior in the stag-hunt game, and arguably do not fully Bayesian-update their beliefs in the urn task.

If the two elicitation methods impact the quality of participants' responses, however, we might still expect to see some systematic differences in the elicited belief distributions, and this is the case in this experiment. In particular, in both the stag-hunt game and the urn task, the probability-matching belief distribution has a mode at 0.5. This is not the case for the frequency method. Moreover, the participants' propensity to report 0.5 in the probability-matching treatment is negatively correlated with their performance in a cognitive reflection test, indicating that those most likely to be confused by the complex method were most inclined to make this salient report. This pattern is not seen with the frequency method. Finally, self-reported task understanding is significantly better in the frequency method compared to probability matching.

Gächter and Renner (2010) show that a simplified non-incentive compatible scoring rule decisively outperforms introspection in their public-goods game. Participants play a 10-period repeated public-goods game, selecting contributions and estimating other group members' average contribution in each period. In the non-incentivized-beliefs treatment, there was no reward for the accuracy of their guess. Participants in the incentivized-beliefs treatment receive a fixed dollar reward if their estimate was sufficiently close to the actual realization, with non-linear reductions in the size of the reward as the guess diverges from the truth. The results show that elicited beliefs were significantly more accurate in the incentivized beliefs treatment. In particular, while first-period reports had roughly the same absolute estimation error, this error fell more quickly for participants in the incentivized-beliefs treatment in the subsequent periods.

The limited evidence outlined in this subsection is far from decisive. Nonetheless, the results reported here encourage us to think that developing and testing simple incentivized-elicitation methods is at the very least promising and should become an integral part of the belief-elicitation literature.

[40] A third game elicits quantiles using the frequency method. This is arguably not simple, so we do not discuss this part of the experiment here.

3.4. Elicitation when the truth is not verifiable

In many interesting and important economic settings, people form beliefs about random variables with non-verifiable outcomes. This would make reliable methods for incentivized belief elicitation in such settings extremely useful for researchers.

Weaver and Prelec (2013) incentivize participants using BTS incentives, explaining that participants would be wise to answer every item truthfully. Compared to introspection participants, BTS participants were less likely to claim that they recognized nonexistent foils (claiming to have seen an item was rewarded with a small bonus). A similar wording was used by Barrage and Lee (2010), Loughran et al. (2014) and Frank et al. (2017). We view these experiments as interesting, but not particularly informative about BTS. They test whether participants trusted the researcher's claims about incentive compatibility, but not whether the mechanism per se performs well.[41]

Staying truer to the formal structure of BTS, Shaw et al. (2011) tell participants that "there is no incentive to misreport" and that they "will have a higher probability of winning a lottery (bonus payment) if they submit answers that are more surprisingly common than collectively predicted". BTS, along with a method that punishes participants for disagreeing with the majority of others, showed a statistically significant improvement in performance relative to the unincentivized control.[42] Encouragingly, the study by Weaver and Prelec (2013) outlined above also included an experiment in which participants were not told that it was in their best interest to report truthfully under the BTS – instead, they were simply shown their BTS scores after each answer. Over time, performance improved, indicating that participants learned that truth-telling yielded higher payoffs.

Overall, we view these results as an indication that there is promise in exploring incentivized elicitation methods in settings where the truth is not verifiable. In particular, we encourage developing and testing simpler incentivized methods in these settings. At the same time, since most tests described above do not actually explain the incentive structure to participants, it would be premature to conclude BTS and its successors improve belief elicitation in practice.

4. Simple incentives-suggestions for experimenters

What should researchers who want to incentivize honesty but are concerned about complexity do in practice? One obvious option is to pick a simpler elicitation method. The frequency and interval methods, classified as simple in Section 2, are particularly attractive. However, it turns out that these methods are not directly applicable in every elicitation task. Fortunately, there is a simple way to transform many elicitation tasks to allow simpler incentivization. We describe how this transformation works in Section 4.1.[43]

If the researcher decides that transforming the elicitation task is not appropriate in her experiment, there is still plenty that she can do. For one, she can choose a simpler implementation of a given mechanism (e.g. the multiple price list rather than the sequential draws implementation of probability matching), or simplify the language of instruction. We discuss these topics in Section 4.2.

4.1. Changing the environment to simplify elicitation

When beliefs are elicited with the frequency method, participants are rewarded with a fixed prize if their guess matches the empirical frequency of outcome E. The interval method rewards participants if their guess is within some interval around the true empirical frequency. In both cases, the realized empirical frequency $\hat{q}$ is key – it provides an intuitive benchmark for belief elicitation and an estimator of the true underlying probability of E:

$$\hat{q} = \frac{\#\ of\ times\ (out\ of\ N)\ that\ E\ is\ drawn}{N}$$

Simple incentivized methods use the empirical frequency $\hat{q}$ to measure the accuracy of participants' guesses. They are then rewarded based on this accuracy. The story is simple and intuitive-participants get rewarded if their guesses turn out to be correct, or close to correct empirically.

[41] Frank et al. (2017) also includes a treatment in which participants observe dynamically calculated BSR scores next to each available answer (in addition to the claim that it was invented by an MIT researcher and rewards truth-telling). In one experiment, this treatment was statistically indistinguishable from the baseline BTS treatment. In another experiment, the distributions of answers in these two treatments were different. This is somewhat hard to interpret, because as the authors themselves write "it is difficult to validate these changes by comparison to some "honest ground-truth distribution" in the latter experiment.

[42] Several other financial incentives also yielded a quantitatively-similar average improvement in performance, but the effects were not statistically-significant. It should also be noted that authors attribute the good performance of BSR to participants' confusion about how they will be rewarded. In particular, they claim that this confusion led to more engagement with the individual questions. If they are correct, the good performance of BSR may not extend to belief elicitation settings of interest.

[43] We add here that psychologists have argued for decades that people are much better at frequencies as shown in pie charts and histograms than with probabilities. A classic reference for this is Gigerenzer and Hoffrage (1995), who discuss improving Bayesian reasoning by using frequency formats. In fact, these formats seem to be already becoming more common in experimental economics. Nevertheless, some papers in psychology (e.g., Griffin and Buehler, 1999) suggest that both frequency and probabilistic characterizations of problems carry their own behavioral issues. Please see the literature on Ecological Rationality for discussions.

Is using the empirical frequency $\hat{q}$ appropriate in every elicitation task? The answer is clearly no. For example, suppose we are interested in eliciting the subjective probability that an outcome of one coin toss is heads. Applying the frequency method to this setting would amount to asking the participant to guess the outcome of one coin toss and giving the reward if the guess turns out to be correct. This method would manage to elicit the subjectively more likely outcome, but it would not reveal anything else about the underlying subjective probability.

The issue in the example above is that the empirical frequency $\hat{q}$ may only be 0 or 1. Therefore, it does not provide a reasonable estimate of the true underlying probability of outcome E. This is the case whenever beliefs about a single draw of a random variable V are elicited. This type of elicitation task is prevalent in experimental applications.

Fortunately, there exists a straightforward transformation of such elicitation tasks that allows researchers to apply simple incentives directly. In particular, rather than drawing the random variable V just once and asking for the subjective probability of the outcome E, the variable is drawn N times with replacement, and the participants report how many of these N draws they expect to result in outcome E. For instance, in the example above the participant could be asked how many out of the next 100 coin tosses she expects to result in heads. For a large-enough N, the empirical frequency $\hat{q}$ is a reasonable estimator of the true probability of outcome E, and simple incentivization becomes possible.

The transformation of the elicitation task is feasible in most experimental settings that seek to elicit participants' beliefs. First, simple incentivization is feasible in experiments that study how people work with probability in stylized settings with randomization devices (bags or urns with chips, bingo cages, …). Rather than "what is the probability that I draw a red ball?" the experimenter simply asks: "how many out of the following 10 draws with replacement will be red balls?" Second, this method is applicable in experiments that elicit participants' beliefs about the behavior of others. Rather than "what is the probability that your counterpart plays 'defect'?", the experimenter asks, "how many participants in this room who play the role of the counterpart will play 'defect'?". Third, we can also often use this simple incentivization to elicit beliefs about real-world events. For instance, we can ask about "the expected percentage of Americans voting for Republicans in an election" or "the expected number of COVID-19 cases in two weeks."

Although this demonstrates that simple elicitation is entirely feasible in a wide range of experimental settings, this feasibility is not universal. In particular, it cannot be used to elicit beliefs about a binary random variable that cannot be drawn repeatedly by its very nature. For instance, we cannot implement this method to find the subjective probability that "the Republican candidate will win the 2020 presidential election" or "the likelihood that Italy will leave the Euro zone this month." This is a non-trivial drawback.

It should also be noted that from a theoretical perspective, beliefs are no longer described by a single subjective probability p in the transformed problem. Rather, they consist of a distribution over all possible empirical frequencies of outcome E. For instance, in the coin-toss example above, the participant's full beliefs specify the probabilities that 1 out of 100 tosses will be heads, 2 out of 100 tosses will be heads, etc. The frequency method elicits only the mode of this distribution (and the interval method approximately the mode). Since the mode of this distribution in the transformed problem actually (approximately) corresponds to the subjective probability p, we do not see this as a major problem. But transforming the problem does mean giving up on some theoretical properties of subjective beliefs, and this should not be forgotten.

Interestingly, the transformation of the belief-elicitation task allows for a significant simplification of elicitation using proper scoring rules. In particular, the *quadratic scoring rule* in this setting boils down to penalizing the participant in proportion to the squared distance of her report r from the actual realization $\hat{q}$, $(r-\hat{q})^2$. Unlike in the "complex setting", simple QSR does not require long payoff tables and asks for little in terms of the participant's prior mathematical knowledge apart from the understanding of distance.[44]

Some other methods, which essentially correspond to simplifications of scoring rules, are also sometimes used in experiments. For instance, it is common to reward correct guesses and let the payoff decline as some function of the distance from the actual realization $\hat{p}$ (see, e.g., Croson (2000) or Gächter and Renner (2010) for applications). The simplicity of these methods is traded-off against their theoretical robustness-truth-telling not being incentive-compatible in general.

4.2. Changing the implementation to simplify elicitation

If the researcher decides not to transform the environment and use simple methods (or if this is not feasible), there is still plenty that she can do to make complex methods easier to understand. For one, individual methods can typically be implemented in multiple ways. Some implementations are much easier to understand than others. The implementations of probability matching are the lead example here. The sequential draws implementation is abstract and its incentive compatibility is not particularly intuitive (at least for those unable or unwilling to do the math). By contrast, the multiple price list implementation consists of a series of relatively straightforward choices between (a) receiving a prize if the outcome E occurs and (b) receiving a prize with probability r.

There is not much evidence that compares individual implementations of elicitation methods. Holt and Smith (2016) find that the multiple-price-list implementation of probability matching outperforms QSR, whilst the sequential draws imple-

[44] As with "complex" proper scoring rules, truth-telling when beliefs are elicited using "simple" proper scoring rules is only incentive-compatible under the assumption of expected payoff maximization. Interestingly, as shown by Harrison et al. (2017), the theoretical bias for reasonable levels of risk aversion may be small in such settings if the full subjective probability distribution is elicited. Of course, asking subjects to provide a useful estimate of the full distribution might be rather optimistic.

mentation does not (although it should be noted that when compared with each other, the differences between the two methods were not statistically significant). Burfurd and Wilkening (2018) find a statistically-significant reduction in implementation time when using a simplified analogy-based version of the sequential-draws implementation. The performance of a more statistically-sophisticated version of the sequential-draws implementation significantly fell when a quiz checking participants' understanding was removed. More work comparing implementations of individual methods needs to be done.

One can also attempt to simplify the language in the instructions. Hao and Houser (2012) develop and Burfurd and Wilkening (2018) test an analogy-based description of the sequential-draws implementation of probability matching. Vespa and Wilson (2016) present a non-mathematical description of the binarized scoring rule. Graphs, figures, and sliders[45] can be used as visual aids for an improved understanding of the elicitation task. This is sometimes done with the sequential-draws implementation of probability matching (for example, see Massoni et al. (2014)). If a complex method is used, training and checking subjects' understanding is also crucial (Burfurd and Wilkening, 2021).

5. Conclusion

In this article, we have provided an extensive review of methods used to elicit beliefs. These methods range from simply asking people for their beliefs to extensive and complex methods that satisfy desirable theoretical properties. Our emphasis is pragmatic: Which approaches produce the highest accuracy regarding beliefs, and at what cost? The evidence does not show significant differences in this regard between introspection and the more complex methods. However, our conclusion that complex methods have not been so effective should not be understood as a triumph of introspection, or any form of endorsement of the idea that monetary incentives do not work in experiments more broadly. It should be understood as the failure of currently-used complex methods to outperform a flawed, but simple alternative.

Introspection may lead participants to pay less attention to the task or reduce their effort (Burfurd and Wilkening, 2021; Massoni et al., 2014). However, complex elicitation creates biases of its own, which are no less serious than those generated by introspection. These biases stem not only from the widely-debated issues of risk-aversion, non-linear probability weighing, or ambiguity aversion, but also arise from the complexity of the elicitation task. This complexity has sometimes been accentuated, rather than alleviated, through the development of new and ever-more theoretically-robust methods. Encouragingly, several simple alternatives and more intuitive implementations of standard methods are available, although more research is needed to determine how well they work in practice.

Beliefs are critical in a vast number of applications, not restricted to experimental work. We expect the need to consider beliefs to be an issue of increasing importance in social science. While effective belief elicitation may not yet be fully settled, we hope that our survey leads to considerably more research, particularly in the realm of simple methods.

Declaration of Competing Interest

We hereby declare that we have no declaration of interest relating to this submission (in fact, there is no funding).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2021.06.032.

References

Agarwal, A., Mandal, D., Parkes, D.C., Shah, N., 2020. Peer prediction with heterogeneous users. ACM Tran. Econ. Comput. (TEAC) 8 (1), 1–34.
Allen, F., 1987. Discovering personal probabilities when utility functions are unknown. Manag. Sci. 33 (4), 542–544.
Andersen, S., Fountain, J., Harrison, G.W., Rutström, E.E., 2014. Estimating subjective probabilities. J. Risk Uncertain. 48 (3), 207–229.
Andersen, S., Harrison, G., Lau, M., Rutström, E., 2013. Discounting behavior and the magnitude effect. Economica 80, 670–697.
Armentier, O., Treich, N., 2013. Eliciting beliefs: proper scoring rules, incentives, stakes and hedging. Eur. Econ. Rev. 62, 17–40.
Arrow, K., 1971. Essays in the Theory of Risk-Bearing. Markham Publishing Company, Chicago, IL.
Baillon, A., Bleichrodt, H., 2015. Testing ambiguity models through the measurement of probabilities for gains and losses. Am. Econ. J. Microecon. 7, 77–100.
Baron-Cohen, S., 2003. The Essential difference: Men, Women and the Extreme Male Brain. Penguin, London, UK.
Barrage, L., Lee, M.S., 2010. A penny for your thoughts: inducing truth-telling in stated preference elicitation. Econ. Lett. 106 (2), 140–142.
Bauer, D., Wolff, I., 2018. Biases in beliefs: Experimental evidence," No. 109, TWI Research Paper Series, Thurgauer Wirtshaftsinstitut. Universität Konstanz.
Blanco, M., Engelmann, D., Koch, A., Normann, H., 2010. Belief elicitation in experiments: is there a hedging problem? Exp. Econ. 13, 412–438.
Bosman, R., van Winden, F., 2002. Emotional hazard in a power-to-take experiment. Econ. J. 112 (476), 147–169.
Brier, G., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3.
Burfurd, I., Wilkening, T., 2018. Experimental guidance for eliciting beliefs with the stochastic becker–degroot–marschak mechanism. J. Econ. Sci. Assoc. 4 (1), 15–28.
Burfurd, I., Wilkening, T., 2021. Cognitive heterogeneity and complex belief elicitation. Exp. Econ. 1–36.
Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74, 1579–1601.
Charness, G., Levin, D., 2005. When Optimal choices feel wrong: a laboratory study of bayesian updating, complexity, and affect. Am. Econ. Rev. 95 (4), 1300–1309.
Charness, G., Gneezy, U., Imas, A., 2013. Experimental methods: eliciting risk preferences. J. Econ. Behav. Organ. 87, 43–51.
Costa-Gomes, M., Weizsäcker, G., 2008. Stated beliefs and play in normal-form games. Rev. Econ. Stud. 75 (3), 729–762.
Croson, R., 2000. Thinking like a game theorist: factors affecting the frequency of equilibrium play. J. Econ. Behav. Organ. 41 (3), 214–299.

[45] See Andersen et al. (2014) for a use of sliders to simplify the presentation of their elicitation mechanism.

Cvitanić, J., Prelec, D., Riley, B., Tereick, B., 2019. Honesty via choice-matching. Am. Econ. Rev. Insights 1 (2), 179–192.
Danz, D., Fehr, D., Kübler, D., 2012. Information and beliefs in a repeated normal-form game. Exp. Econ. 15 (4), 622–640.
Danz, D., L. Vesterlund, and A. Wilson. (2020), "Belief elicitation: limiting truth telling with information on incentives," Unpublished paper.
Dawes, R.M., 1990. The potential nonfalsity of the false consensus effect. Insights Decis. Mak. A Trib. Hillel J. Einhorn 179–199.
DuCharme, W., Donnell, M., 1973. Intrasubject comparison of four response modes for 'subjective probability' assessment. Organ. Behav. Hum. Perform. 10 (1), 108–117.
Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games Econ. Behav. 30, 163–182.
Erkal, N., Gangadharan, L., Koh, B.H., 2020. Replication belief elicitation with quadratic and binarized scoring rules. J. Econ. Psychol. 81, 102315.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114, 817–868.
Frank, M.R., Cebrian, M., Pickard, G., Rahwan, I., 2017. Validating Bayesian truth serum in large-scale online human experiments. PLoS ONE 12 (5), e0177385.
Gächter, S., Renner, E., 2010. The effects of (incentivized) belief elicitation in public goods experiments. Exp. Econ. 13, 364–377.
Gigerenzer, G., Hoffrage, U., 1995. How to improve Bayesian reasoning without instruction: frequency formats. Psychol. Rev. 102 (4), 684–704.
Gneiting, T., Raftery, A., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102 (477), 359–378.
Good, I., 1952. Rational decisions. J. R. Stat. Soc. Ser. B (Methodol.) 14 (1), 107–114.
Grether, D., 1981. "Financial incentive effects and individual decision-making," California Institute of Technology, Working Paper 401.
Griffin, D., Buehler, R., 1999. Frequency, probability, and prediction: easy solutions to cognitive illusions? Cogn. Psychol. 38 (1), 48–78.
Halevy, Y., 2015. Time consistency: stationarity and time invariance. Econometrica 83 (1), 335–352.
Hao, L., Houser, D., 2012. Belief elicitation in the presence of naïve respondents: an experimental study. J. Risk Uncertain. 44 (2), 161–180.
Harrison, G., Martinez-Correa, J., Swarthout, T., 2014. Eliciting subjective probabilities with binary lotteries. J. Econ. Behav. Organ. 101, 128–140.
Harrison, G., Martinez-Correa, J., Swarthout, T., Ulm., E., 2017. Scoring rules for subjective probability distributions. J. Econ. Behav. Organ. 134, 430–448.
Haruvy, E., Lahav, Y., Noussair, C., 2007. Traders' expectations in asset markets: experimental evidence. Am. Econ. Rev. 97 (5), 1901–1920.
Heinemann, F., Nagel, R., Ockenfels, P., 2009. Measuring strategic uncertainty in coordination games. Rev. Econ. Stud. 76, 181–221.
Holt, C., Smith, A., 2016. Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. Am. Econ. J. Microecon. 8 (1), 110–139.
Holt, C., 2006. Markets, Games and Strategic Behavior. Pearson/Addison-Wesley, Boston.
Holt, C., Laury, S., 2002. Risk aversion and incentive effects. Am. Econ. Rev. 92, 1644–1655.
Holt, C., Smith, A., 2009. An update on Bayesian updating. J. Econ. Behav. Organ. 69, 125–134.
Hossain, T., Okui, R., 2013. The binarized scoring rule. Rev. Econ. Stud. 80, 984–1001.
Houser, D., Xiao, E., 2011. Punish in public. J. Public Econ. 95 (7), 1006–1017.
Hurd, M., 2009. Subjective probabilities in household surveys. Annu. Rev. Econ. 1, 543–562.
Hurley, T., Shogren, J., 2005. An experimental comparison of induced and elicited beliefs. J. Risk Uncertain. 30, 169–188.
Hyndman, K., Terracol, A., Vaksmann, J., 2013. "Beliefs and (in)stability in normal-form games," working paper, University of Texas, Dallas.
Ivanov, A., 2011. Attitudes to ambiguity in one-shot normal-form games: an experimental study. Games Econ. Behav. 71, 366–394.
Kadane, J., Winkler, R., 1988. Separating probability elicitation from utilities. J. Am. Stat. Assoc. 83, 357–363.
Karni, E., 2009. A mechanism for eliciting probabilities. Econometrica 77, 603–606.
Krupka, E., Weber, R., 2013. Identifying social norms using coordination games: why does dictator game sharing vary? J. Eur. Econ. Assoc. 11 (3), 495–524.
Li, S., 2017. Obviously strategy-proof mechanisms. Am. Econ. Rev. 107 (11), 3257–3287.
Liu, Y., Chen, Y., 2017. Machine-learning aided peer prediction. In: Proceedings of the 2017 ACM Conference on Economics and Computation, pp. 63–80.
Loughran, T.A., Paternoster, R., Thomas, K.J., 2014. Incentivizing responses to self-report questions in perceptual deterrence studies: an investigation of the validity of deterrence theory using Bayesian truth serum. J. Quant. Criminol. 30 (4), 677–707.
Manski, C., 2004. Measuring expectations. Econometrica 72, 1329–1376.
Manski, C., Neri, C., 2013. First- and second-order subjective expectations in strategic decision-making: experimental evidence. Games Econ. Behav. 81, 232–254.
Marks, G., Miller, N., 1987. Ten years of research on the false-consensus effect: an empirical and theoretical review. Psychol. Bull. 102 (1), 72.
Martínez-Marquina, A., Niederle, M., Vespa, E., 2019. Failures in contingent reasoning: the role of uncertainty. Am. Econ. Rev. 109 (10), 3437–3474.
Massoni, S., Gadjos, T., Vergnaud, J.-.C., 2014. Confidence measurement in the light of signal detection theory. Frontiers in Psychology 5, 1455.
Miller, N., Resnick, P., Zeckhauser, R., 2005. Eliciting informative feedback: the peer-prediction method. Manag. Sci. 51 (9), 1359–1373.
Nyarko, Y., Schotter, A., 2002. An experimental study of belief learning using elicited beliefs. Econometrica 70, 971–1005.
Offerman, T., Palley, A., 2016. Lossed in translation: an off-the-shelf method to recover probabilistic beliefs from loss-averse agents. Exp. Econ. 19 (1), 1–30.
Offerman, T., Sonnemans, J., van de Kuilen, G., Wakker, P., 2009. A truth serum for non-Bayesians: correcting proper scoring rules for risk attitudes. Rev. Econ. Stud. 76, 1461–1489.
Oprea, R., 2020. What makes a rule complex? Am. Econ. Rev. 110 (12), 3913–3951.
Palfrey, T., Wang, S., 2009. On eliciting beliefs in strategic games. J. Econ. Behav. Organ. 71, 98–109.
Prelec, D., 2004. A Bayesian truth serum for subjective data. Science 306 (5695), 462–466.
Rabin, M., 2000. Risk aversion and expected-utility theory: a calibration theorem. Econometrica 68 (5), 1281–1292.
Radanovic, G., Faltings, B., 2013. A robust bayesian truth serum for non-binary signals. In: Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 27, No. 1).
Radanovic, G., Faltings, B., 2014. Incentives for truthful information elicitation of continuous signals. In: Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 28, No. 1).
Radanovic, G., Faltings, B., 2015. Incentives for subjective evaluations with private beliefs. In: Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).
Radanovic, G., Faltings, B., Jurca, R., 2016. Incentives for effort in crowdsourcing using the peer truth serum. ACM Trans. Intell. Syst. Technol. (TIST) 7 (4), 1–28.
Ramsey, F., 1926. "Truth and probability," In antony eagle (Ed.), philosophy of probability: contemporary readings, routledge, 52–94.
Rey-Biel, P., 2009. Equilibrium play and best response to (stated) beliefs in normal form games. Games Econ. Behav. 65, 572–585.
Roby, T., 1965. Belief States: A Preliminary Empirical Study. Decision Science Laboratory, L.G. Hascom Field.
Ross, L., Greene, D., House, P., 1977. The "false consensus effect": an egocentric bias in social perception and attribution processes. J. Exp. Soc. Psychol. 13 (3), 279–301.
Rustichini, A., Villeval, M., 2014. Moral hypocrisy, power and social preferences. J. Econ. Behav. Organ. 107, 10–24.
Rutström, E., Wilcox, 2009. Stated beliefs versus inferred beliefs: a methodological inquiry and experimental test. Games Econ. Behav. 67, 616–632.
Schlag, K., Tremewan, J., 2021. Simple belief elicitation: An experimental evaluation. J Risk Uncertain doi:10.1007/s11166-021-09349-6.
Schlag, K., van der Weele, J., 2013. Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality. Theor. Econ. Lett. 3, 38–42.
Schlag, K., Tremewan, J., van der Weele, J., 2015. A penny for your thoughts: a survey of incentives and belief elicitation. Exp. Econ. 18 (3), 457–490.
Schotter, A., Trevino, I., 2014. Belief elicitation in the laboratory. Ann. Rev. Econ. 6, 103–128.
Selten, R., Sadrieh, A., Abbink, K., 1999. Money does not induce risk-neutral behavior but binary lotteries do even worse. Theory Decis. 46, 211–249.
Shaw, A.D., Horton, J.J., Chen, D.L., 2011. Designing incentives for inexpert human raters. In: Proceedings of the ACM 2011 conference on Computer supported cooperative work, pp. 275–284.
Shnayder, V., Agarwal, A., Frongillo, R., Parkes, D.C., 2016. Informed truthfulness in multi-task peer prediction. In: Proceedings of the 2016 ACM Conference on Economics and Computation, pp. 179–196.

Smith, L., Walker, J., 1993. Monetary rewards and decision cost in experimental economics. Econ. Inq. 31, 245–261.
Sonnemans, J., Offerman, T., 2001. "Is the quadratic scoring rule really incentive compatible?," Unpublished paper.
Toda, M., 1963. Measurement of Subjective Probability Distribution. Division of Mathematical Psychology, Institute for Research.
Trautman, S., van de Kuilen, G., 2014. Belief elicitation: a horse race among truth serums. Econ. J. 125, 2116–2135.
Vespa, E., Wilson, A., 2016. Communication with multiple senders: an experiment. Quant. Econ. 7 (1), 1–36.
Wang, S., 2011. Incentive effects: the case of belief elicitation from individuals in groups. Econ. Lett. 111, 30–33.
Weaver, R., Prelec, D., 2013. Creating truth-telling incentives with the Bayesian truth serum. J. Market. Res. 50 (3), 289–302.
Wilcox, N., Feltovich, N., 2000. "Thinking like a game theorist: comment," Working Paper, Department of Economics, University of Houston.
Winkler, R., Murphy, A., 1970. Nonlinear utility and the probability score. J. App. Meteorol. 9, 143–148.
Witkowski, J., Parkes, D., 2012. A robust Bayesian truth serum for small populations. In: Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 26, No. 1).
Xiao, E., Houser, D., 2005. Emotion expression in human punishment behavior. Proc. Natl. Acad. Sci. 102 (20), 7398–7401.
Zizzo, D., 2010. Experimenter demand effects in economics experiments. Exp. Econ. 13, 75–98.