BELIEF ELICITATION:
A USER'S GUIDE[†]

PAUL J. HEALY[*] & GREG LEO[**]

ABSTRACT. In this chapter we review the various incentive compatible methods for eliciting many different statistics of one's beliefs. We discuss theoretical and behavioral properties of each mechanism and review the relevant literature. Where appropriate, we provide our own recommendations and guidance on best practices for conducting a belief elicitation.

Keywords: Belief elicitation; Uncertainty; Experimental methodology

Draft: January 5, 2025

[†]We thank Ned Augenblick, Glenn Harrison, Muriel Niederle, Kirby Nielsen, Ryan Oprea, Lise Vesterlund, and Alistair Wilson for valuable comments and suggestions.
[*]Dept. of Economics, The Ohio State University; healy.52@osu.edu.
[**]Dept. of Economics, Loyola Marymount University; greg.leo@lmu.edu.

1. How To Use This Chapter

Incentivized belief elicitation is a surprisingly broad topic. There are many possible statistics that can be elicited, and a variety of mechanisms that can be used. Each is incentive compatible under different assumptions, and each comes with its own implementation issues. As a result, this chapter is quite large, and we do not expect the typical user to read every subsection. Instead, we have tried to design this chapter for researchers seeking advice on a particular elicitation problem they face in their own work. To that end, we designed the chapter to be a bit like a "choose your own adventure" novel, with references from each section to the next relevant section based on the particular research goal. We recommend using the "Guide to this Chapter" box below to pick your starting location according to your goal.

To illustrate the choices you might face, imagine that you want to ask experiment participants their beliefs about the scores of a group of people on an exam that is scored out of 100 points, and a test-taker is said to "pass" if their score is above, say, 75 points. (We will use this illustrative example throughout the chapter.) One common elicitation is to have the participant guess the group's average score (out of 100) and pay them \$10 if their guess is within ± 5 points. But what exactly this elicits is a bit complicated: The participant should consider their entire belief distribution over the possible values of the average score, identify the 10-point-wide interval that has the highest overall probability and then report the midpoint of that interval.

Perhaps you want a statistic that is easier to interpret. You could elicit the mean of that distribution over average scores. Section 5.8 provides several mechanisms that you can use, such as having them guess the average score and penalizing them according to the quadratic distance from the true average. A linear penalty would elicit the median (Section 5.6). Or maybe you would be interested in their beliefs about the fraction of people who pass the exam. Again, they would have a distribution over that quantity ranging from 0% to 100% and you could elicit the mean or median or modal interval of their belief distribution. But perhaps a simpler method is to ask them the probability that a randomly selected person from the group passes the exam. Now their belief is a single number, the probability, and there are many incentive compatible mechanisms for eliciting a single probability (Section 5.1).[1]

[1]A mechanism is said to be incentive compatible for some statistic (such as a single probability or a median or a modal interval) if it incentivizes participants to reveal that statistic truthfully under some specified preference assumptions. As we discuss, there are often several mechanisms that are incentive compatible for the same statistic, but under different preference assumptions.

We believe that it is important to be thoughtful about which statistic you elicit in cases like this. For example, if you elicit someone's beliefs about the average score, then you will not be able to back out their beliefs about the percentage of people who will pass the exam and vice versa. As another example, if you were interested in how beliefs update after learning some information, the formula you would use to calculate the Bayesian benchmark would differ depending on whether you have elicited the mean, median, or probability of a single person passing. Canen and Chakraborty (2023) provide an example where mean beliefs and median beliefs update in opposite directions in response to new information and emphasize that this can lead to incorrect interpretations. Section 4 provides a starting point for making the choice about what to elicit.

But even before we choose what to elicit, there is still an active debate about whether or not we should be using pecuniary incentives. Would we not be better off simply asking people their beliefs and eschewing these complicated issues of incentives? Wouldn't participants want to reveal truthfully even without being paid? Section 3 provides a discussion of this question and reviews existing evidence in the domain of belief elicitation.

Section 5 provides an encyclopedic overview of the various mechanisms for each statistic, with a quick discussion of the incentives and incentive compatibility of each. We also include brief discussions on how to elicit a coarse belief, reducing the possible reports to a finite grid, and whether this may simplify each mechanism. For an in-depth background on the theory of incentive compatible mechanisms, the interested reader should consult Section 2.

We have tried to write this chapter in a fairly objective and comprehensive manner. However, we include our own recommendations in separate pop-out boxes for the interested reader. Of course, we caution that these are often recommendations supported by limited available data or experience, and we know of many researchers who would disagree with some or all of our recommendations.

Ours is certainly not the first survey on belief elicitation. Schotter and Trevino (2014) provide an excellent review that focuses more on elicitation in strategic settings. For example, they discuss accuracy of first- and second-order beliefs, hedging issues, whether elicitation changes subsequent behavior (evidence seems mixed, but it is a valid concern), and the belief formation process itself.[2] The survey of Charness et al. (2021) provides many updated references. They focus more on the complexity of various mechanisms (which they admit is subjective)—including thoughts on how one might reduce

[2]The repeated games chapter of this handbook also discusses experimental design considerations for minimizing the effect of belief elicitation on behavior.

complexity—and provide a very nice review of experiments that compare different mechanisms. We attempt to replicate that review at relevant points throughout our chapter. The closest review to ours in spirit is probably Schlag et al. (2015), who focus more on the conditions for incentive compatibility of each mechanism. Like us, they also survey mechanisms for eliciting statistics other than a probability, such as the mean and median. Their review of experiments that compare mechanisms and those that test for evidence of hedging are excellent; we certainly stand on their shoulders here.

GUIDE TO THIS CHAPTER

Where should I start?

- **I want to learn more about the technical details of elicitation.**
  Go to Section 2: *The Theory of Incentive Compatible Belief Elicitation*
- **I need to decide whether to incentivize the elicitation of beliefs.**
  Go to Section 3: *Should I Use (and Emphasize) Pecuniary Incentives?*
- **I need to choose which statistic to elicit.**
  Go to Section 4: *Choosing What to Elicit*
- **I need to choose how to elicit a particular statistic.**
  Go to the menu in Section 4 and use that to navigate to the relevant statistic in Section 5: *The Mechanisms*.

2. The Theory of Incentive Compatible Belief Elicitation

This section gives a technical overview of how we model beliefs in a choice framework, describes the two main families of belief elicitation methodologies, and discusses the incentives involved in each, as well as what assumptions are required for incentive compatibility. We advise readers who are not immediately interested in these technical details to skip this section. The content is not required to understand the direction and advice given throughout the rest of the chapter. However, throughout the chapter, we have also provided references to the relevant portions of this section at locations where interested readers *may* want to refer back to specific technical details.

2.1. What is a Belief?

Suppose a participant would rather have a bet that pays \$10 if a randomly chosen student scores an "A" on an exam than a bet that pays \$10 if that student scores a "B". We can reasonably conclude that they must think that the grade "A" is more likely. Thus, information about beliefs can easily be inferred from choices. We begin this section by formalizing this idea into a definition of beliefs in a pure choice framework.

To formalize this, suppose that there is a state space (finite or countable) Ω and only two outcomes (or prizes): a high prize $\overline{x}$ and a low prize $\underline{x}$. This is the case, for example, when all payments are binarized (see Section 2.4 for a discussion of binarization in the context of scoring rules). Following the Anscombe-Aumann [AA] setup (Anscombe and Aumann, 1963), we assume that there exist subjective bets on events in $\mathscr{E}$ as well as objective lotteries that pay with a known probability.

Formally, for any $p \in [0,1]$, let

$$L^p = (\overline{x}, p; \underline{x}, (1-p))$$

represent the *objective lottery* that pays the high prize with probability p and the low prize otherwise. Then, for any finite partition $\mathscr{E} = \{E_1, \ldots, E_n\}$ of Ω, let

$$f = [L^{p_1}, E_1; L^{p_2}, E_2; \ldots; L^{p_n}, E_n]$$

be the AA *act* that pays lottery L^{p_i} in each event E_i. Acts that pay the high prize with certainty (L^1) if an event E occurs and the low prize with certainty (L^0) if E^c occurs are called "pure acts" or "pure bets" and are of the form

$$f^E = [L^1, E; L^0, E^c].$$

For clarity this may also be written as $f^E = [\overline{x}, E; \underline{x}, E^c]$. The space of all possible AA acts is given by $\mathscr{A}$ and we assume our decision maker has a preference ordering $\succsim$ over $\mathscr{A}$. To make the problem non-trivial we assume $\overline{x} \succ \underline{x}$.

2.1.1. Defining Beliefs

With this notation, we can formally define what it means to have an *ordinal belief* (or qualitative probability) in pure choice terms.

Definition 1. (*Ordinal Belief*) For any events $E, F \subseteq \Omega$, a participant believes E is more likely than F if $f^E \succsim f^F$.

Many beliefs we discuss in this chapter are ordinal. For example, Definition 1 is sufficient for defining and eliciting the most likely event from a set (see Section 5.3), the relative likelihood of events (see Section 5.4), or even the mode/ modal interval of a random variable (see Section 5.5).

To go further and provide a cardinality to beliefs, we need a measuring stick. The measuring stick we use are the objective lotteries. Suppose that instead of comparing a bet on "Grade of A" to a bet on "Grade of of B", we instead compare it to $L^{0.5}$, the lottery that pays $\overline{x}$ with probability 0.5. Let E be "Grade of A". If $f^E \succsim L^{0.5}$ it is reasonable to say that the participant believes that "Grade of A" is more likely than 50%.

More generally, suppose that we compare f^E against the set of all pure lotteries $\{L^p\}_{p\in[0,1]}$. If we find some L^p such that they are indifferent between betting on "Grade of A" and getting paid L^p then we can say that a person believes the probability of an "A" grade *is* p. This definition follows Ramsey (1931), De Finetti (1937, 1962), Savage (1954), Anscombe and Aumann (1963), and Machina and Schmeidler (1992, 1995).

Definition 2. (*Probabilistic Belief*) A participant believes the probability of E is $p \in [0,1]$ if $f^E \sim L^p$.

This indifference point in probability space has many different names in the literature, including "matching probability", "indifference probability", "probability equivalent", and "uncertainty equivalent".

2.1.2. *Existence and Uniqueness*

What assumptions on $\succsim$ guarantee that a belief exists for every $E \subseteq \Omega$? In other words, how do we ensure that the bet f^E is indifferent to some objective pure lottery L^p? First, we need that preferences are complete and transitive.[3]

Axiom 1 (Preference)**.** Preferences $\succsim$ over $\mathscr{A}$ are complete and transitive.

Next, we need the space of objective pure lotteries to be ranked so that the decision maker prefers higher probabilities of winning the prize. Since there are only two outcomes, this just means that $\succsim$ should respect stochastic dominance over pure lotteries. We refer to this as monotonicity over pure lotteries.

Axiom 2 (Monotonicity over Pure Lotteries)**.** $p \geq q$ if and only if $L^p \succsim L^q$.

Axiom 2 is enough to ensure the existence of a continuous, strictly increasing utility functional V on $[0,1]$ such that $V(p) \geq V(q)$ if and only if $L^p \succsim L^q$.[4] Indeed, the identity function for which $V(p) = p$ for all $p \in [0,1]$ would represent these preferences.

We can also extend the domain of V to represent preferences over pure bets. In other words, we can write $V(E) \geq V(F)$ if and only if $f^E \succsim f^F$. Having a belief p for event E thus requires that $V(E) = V(p)$. To ensure this indifference exists we next we need to rule out the possibility that $V(E) > V(1)$ (which would imply the belief of E is over 100%) or $V(E) < V(0)$ (the belief of E is negative). This is done by assuming preferences respect statewise dominance for subjective events. Since there are only two possible outcomes, bet f^E statewise dominates bet f^F if and only if $E \supseteq F$. This gives the following axiom.

Axiom 3 (Statewise Monotonicity over Pure Bets)**.** If $E \supseteq F$ then $f^E \succsim f^F$, and if $E \supset F$ then $f^E \succ f^F$.

Statewise monotonicity ensures that V ranks events by set inclusion. We also assume (without stating it as an axiom) that $L^1 \sim f^\Omega$ and $L^0 \sim f^\emptyset$ since these represent logically identical objects. Statewise monotonicity over pure bets then implies that that $V(E) \leq V(\Omega) = V(1)$ and $V(E) \geq V(\emptyset) = V(0)$. By the fact that V is continuous and strictly increasing on $[0,1]$, there must exist a unique p such that $V(E) = V(p)$, meaning $f^E \sim L^p$.

Proposition 1. (*Existence and Uniqueness of p*) If $\succsim$ satisfies Axioms 1, 2, and 3 then for every event $E \subseteq \Omega$ the decision maker has a probabilistic belief $\mu(E) \in [0,1]$ according to Definition 2.

[3] Our notion of completeness includes the case of $f \succsim f$ and thus implies reflexivity.

[4] The rationals on $[0,1]$ provide the needed countable $\succsim$-dense subset.

2.1.3. Well-Behaved Beliefs

This notion of a belief above does not guarantee that the beliefs are well behaved in the sense of Kolmogorov (1950). In particular, they may not be finitely additive, meaning that if E and F are disjoint it may not be the case that $\mu(E \cup F) = \mu(E) + \mu(F)$. For example, it is possible that $\mu(E) + \mu(E^c) \neq 1$. In addition, it does not guarantee that more complex AA acts are reduced to equivalent simple lotteries. The axioms of Machina and Schmeidler (1995), which imply the three axioms above, gives not only the existence and uniqueness of beliefs, but finite additivity and reduction. We present them here for our special setting with only two outcomes.

Axiom 4 (FOSD Dominance)**.** If $p_i \geq q_i$ then

$$[L^{q_1}, E_1; \ldots; L^{p_i}, E_i; \ldots; L^{q_n}, E_n] \succsim [L^{q_1}, E_1; \ldots; L^{q_i}, E_i; \ldots; L^{q_n}, E_n],$$

with strict preference whenever $p_i > q_i$.

The dominance axiom is a statewise monotonicity axiom for the case where binary objective lotteries are paid in each state. And it implies the previous axiom that preferences are monotonic over pure lotteries (Axiom 2) by letting $E_i = \Omega$.

Axiom 5 (Horse/Roulette Replacement)**.** For any partition $\{E_1, \ldots, E_n\}$, if for some $\alpha \in [0,1]$ and pair of events E_i and E_j:

$$f^{E_i} = \begin{bmatrix} \overline{x}, E_i; \\ \underline{x}, E_j; \\ \underline{x}, E_{k \neq i,j} \end{bmatrix} \sim \begin{bmatrix} L^{\alpha}, E_i; \\ L^{\alpha}, E_j; \\ \underline{x}, E_{k \neq i,j} \end{bmatrix}$$

then for any lotteries $\{L^{p_1}, \ldots, L^{p_n}\}$ and any $i, j \in \{1, \ldots, n\}$,

$$\begin{bmatrix} L^{p_i}, E_i; \\ L^{p_j}, E_j; \\ L^{p_k}, E_{k \neq i,j} \end{bmatrix} \sim \begin{bmatrix} L^{\alpha p_i + (1-\alpha) p_j}, E_i; \\ L^{\alpha p_i + (1-\alpha) p_j}, E_j; \\ L^{p_k}, E_{k \neq i,j} \end{bmatrix}.$$

The α in the replacement axiom can be thought of as the conditional probability of E_i given $E_i \cup E_j$. Intuitively, it is the rate at which the decision maker replaces subjective uncertainty (of E_i conditional on $E_i \cup E_j$) with objective randomness. The axiom says that this same rate is applied regardless of the payment lotteries received conditional on E_i and E_j, and regardless of what is paid in all other states. This is the key axiom behind both reduction and finite additivity.

Axiom 6 (Mixture Continuity)**.** For any partition $\{E_1,\ldots,E_n\}$, if

$$[L^{p_1},E_1;\ldots;L^{p_n},E_n] \succ [L^{p'_1},E_1;\ldots;L^{p'_n},E_n] \succ [L^{p''_1},E_1;\ldots;L^{p''_n},E_n]$$

then there exists a probability $\alpha \in (0,1)$ such that

$$[L^{p'_1},E_1;\ldots;L^{p'_n},E_n] \sim [L^{\alpha p_1+(1-\alpha)p''_1},E_1;\ldots;L^{\alpha p_n+(1-\alpha)p''_n},E_n].$$

This final axiom is simply a continuity axiom that guarantees all needed indifference points exist. It was not necessary for Proposition 1 because $V(p)$ over pure lotteries inherits continuity from the monotonicity assumption, but on this larger domain monotonicity over pure lotteries alone is no longer sufficient.

Theorem 1 (Machina and Schmeidler, 1995)**.** A preference relation $\succsim$ over $\mathscr{A}$ satisfies Axioms 1, 4, 5, and 6 if and only if there exists a finitely additive belief $\mu(\cdot)$ and a continuous and strictly increasing functional $V(\cdot)$ such that $[L^{p_1},E_1;\ldots,L^{p_n},E_n]$ is evaluated according to:

$$V\left(\sum_{i=1}^{n}\mu(E_i)p_i\right).$$

In other words, the decision maker has a finitely additive belief, uses that to reduce AA acts to a simple lottery that pays $\overline{x}$ with probability $\sum_i \mu(E_i)p_i$, and then ranks those simple lotteries according to their payoff probability. Since V is continuous and strictly increasing, $V(p) = p$ is an admissible representation of $\succeq$. Machina and Schmeidler (1995) refer to a decision maker as *probabilistically sophisticated* if their preferences satisfy the axioms of Theorem 1.

2.2. Eliciting "Beliefs" Under Ambiguity Aversion

There are some cases where probabilistic sophistication is known to be violated. For example, in the Ellsberg paradox players choose bets that reveal $\mu(\cdot)$ must sum to less than one. Our Definition 2 does not rule out this possibility. Thus, we can elicit "beliefs" even when those beliefs do not sum to one. However, binarized scoring rules require probabilistic sophistication for incentive compatibility (see below) and thus should not be used to elicit non-additive beliefs. The BDM and MPL mechanisms do not require probabilistic sophistication and therefore can still be used.

Testing for this and measuring the degree of ambiguity aversion (or ambiguity loving) is, in principle, simple: Measure the participant's probability equivalents for E and its complement — paying for one of the two randomly — and comparing their sum to

100%. The difference between 100% and the sum of the probability equivalents is a sensible measure of this participant's degree of ambiguity aversion for these two events. For a more general measure of ambiguity attitudes, see Baillon et al. (2018); we do not cover ambiguity elicitation in this chapter.

When probabilistic sophistication is violated (for example, due to ambiguity aversion) we arguably should not refer to the elicited quantities as "beliefs" or "probabilities" as they do not satisfy the standard properties of a probability distribution (Kolmogorov, 1950). Instead, we will refer to them as "probability equivalents," since they are simply the objective probabilities that the participant views as equivalent to the ambiguous bet.

Although we can still measure these probability equivalents, interpreting them is more nuanced. Under the Maxmin Expected Utility model (Gilboa and Schmeidler, 1989) the participant is assumed to (behave as if they) have a closed interval of probabilities, and their stated probability equivalent for each event will be the minimum probability in that interval. In that case, the elicited probability equivalents have a clear meaning. But for other models of ambiguity aversion (such as Klibanoff et al., 2005) the interpretation is less clear. In general, the elicited probability equivalent captures both the participant's "likelihood" of the event and their degree of ambiguity aversion, and disentangling these is impossible unless the question is viewed through the lens of a specific model.

2.3. Incentive Compatibility of the MPL and BDM Mechanisms

Mechanisms in the BDM-based family (which includes the MPL) are designed to elicit the indifference probability from Definition 2 by asking the participant to choose between a pure bet f^E and an objective lottery L^p for many possible values of p. If they prefer the subjective bet to the objective lottery that pays if p' their belief about E must be larger than p'. If they prefer the objective lottery that pays with p'' to the subjective bet, their belief about the probability of E must be smaller than p''. In this case, p must be in the interval $[p', p'']$. This is formalized in the following straightforward proposition which follows from monotonicity over pure lotteries (Axiom 2) and transitivity of preferences.

Proposition 2. (*Belief Interval*) Let p be a participants belief about event E according to Definition 2. If $f^E \succsim L^{p'}$ and $L^{p''} \succsim f^E$ then $p \in [p', p'']$.

By having a participant make many such comparisons, the BDM "traps" the participant's belief in the smallest such interval. This can be done literally by presenting many binary choices as a large list (the "multiple price list" method) or by asking the participant to report belief q and then randomly drawing the value of p from $[0,1]$ and paying them based on what their choice would be in for the resulting pair given their reported belief.

This approach is based on a mechanism for eliciting willingness-to-pay, originally by Becker (1968). Thus, BDM-based mechanisms are a class of mechanisms developed specifically within the experimental economics literature. This is in contrast to scoring rules (discussed below), which were developed originally in the statistics and decision-theory literature, originating from the seminal work of Brier (1950).

To our knowledge, the idea of using a BDM-like mechanism to elicit probabilities was originated by Savage (1971) and was first used in an actual experiment by Grether (1981). The idea has been independently rediscovered multiple times, including in the work of Karni (2009), Holt and Smith (2009, 2016), and Möbius et al. (2022). We believe that Holt and Smith (2016) were the first to represent it as a multiple price list (MPL).

The strength of the BDM and MPL approach is the simplicity of the incentives. This is because it is based on simple comparisons between objective lotteries and pure bets. If you ask someone to make a choice between a single objective lottery L^p and a pure bet f^E, then there are no further assumptions you need to make to infer their preference between these options from their choice. However, the BDM/MPL requires participants to make *many* such choices. The mechanism then randomly picks one for payment.

To formalize this, imagine a second state space $\mathscr{T} \subseteq [0,1]$ representing which row of the MPL is chosen for payment, or which probability is randomly chosen in a BDM. The realized $\tau \in \mathscr{T}$ is drawn according to some absolutely continuous probability measure. For example, if the MPL has a finite number of rows, then each is drawn with positive probability. In practice, this is usually the uniform distribution.

For a participant that announces q, they are paid f^E (a bet on E) if $\tau \le q$ and are paid L^τ (a lottery that pays with probability τ) if $\tau > q$. This is visualized in Figure 1.

If their true belief is p then they prefer f^E to L^τ if and only if $\tau < p$. Now suppose that they switch from truth-telling (shown in Figure 1) to announcing $p+0.01$. In branch $\tau = p+0.01$, they would switch from being paid $L^{p+0.01}$ to being paid f^E. But they strictly prefer $L^{p+0.01}$, so they are now being paid something they like less on that branch. For all other branches, their payment is unchanged. Therefore, if we think of τ as the "state", this misreport leads to a lottery dominated state-by-state.

In the theory developed above, preferences $\succsim$ are defined over two-stage acts, where the subjective stage resolves first, followed by an objective stage. The second and third

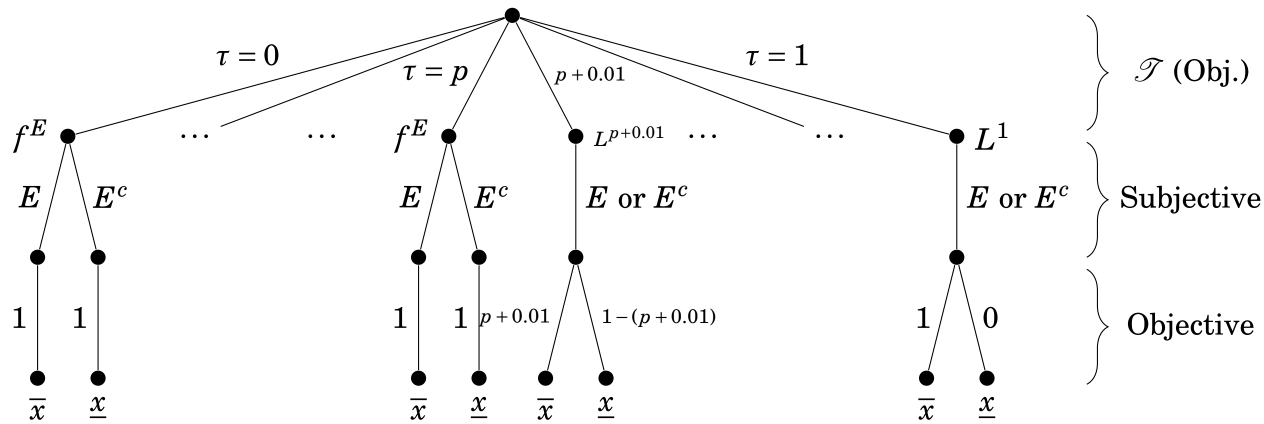


FIGURE 1. The three-stage act generated by the MPL when the participant announces their true belief p. Thus, $f^E \succ L^\tau$ if and only if $\tau < p$.

stages in Figure 1 have this structure. But the BDM and MPL generate three-stage acts, so it becomes necessary to extend our theory to this domain.

Let $\phi : \mathscr{T} \to \mathscr{A}$ represent a three-stage act, and $\phi(\tau) \in \mathscr{A}$ be the two-stage act paid at each $\tau \in \mathscr{T}$. The space of all such three-stage acts is given by Φ. We need to extend the preferences defined in the previous section to Φ. Refer to this extension as $\succsim^*$. We assume consistency between $\succsim^*$ and $\succsim$, meaning if $\phi(\tau) = f$ and $\phi'(\tau) = f'$ for every τ then $\phi \succsim^* \phi'$ if and only if $f \succsim f'$.[5]

Axiom 7 ($\mathscr{T}$ (Statewise) Monotonicity)**.** Given $\succsim$ over $\mathscr{A}$, a consistent extension $\succsim^*$ over Φ satisfies statewise monotonicity if $\phi(\tau) \succsim \phi'(\tau)$ for every τ implies $\phi \succsim^* \phi'$, with strict preference if $\phi(\tau') \succ \phi(\tau')$ for some $\tau' \in \mathscr{T}$.

The following is then a simple application of Azrieli et al. (2018).

Proposition 3. A BDM or MPL is incentive compatible for $\succsim$ whenever the extension $\succsim^*$ satisfies $\mathscr{T}$ Statewise Monotonicity.

Given this, incentive compatibility is relatively easy to explain to participants. If they lie about their belief, they change their answer on some rows of the list to something they prefer less. Thus, the only way to they get their most-preferred lottery/bet on every row is to tell the truth.

2.4. Scoring Rules

A scoring rule is a function $s_x(q)$ that maps the announcement of a participant (about the probability or statistic being elicited) and the true realization of the event or random

[5]See Azrieli et al. (2018) for a more precise definition of preference extensions.

variable into some payment. In *dollar-denominated* scoring rules, these payments are amounts of money. In *binarized* scoring rules, these payments are probabilities of a fixed amount of money.

When eliciting the probability of an event, a scoring rule is a pair of functions: $s_1(q)$ is the payment received if the participant reports a probability q and the event occurs, and $s_0(q)$ is their payment if the event does not occur. If their true belief is p then their expected payment is

$$p\,s_1(q)+(1-p)\,s_0(q). \tag{1}$$

Scoring rules date back to Brier, 1950), who introduced the *quadratic scoring rule* [QSR]. This rule is still the most famous and most widely used in experimental economics. For eliciting a probabilistic belief, it has the form:

$$\begin{aligned} s_1(q) &= 1-(1-q)^2 \text{ and} \\ s_0(q) &= 1-(0-q)^2. \end{aligned} \tag{2}$$

A scoring rule is said to be strictly *proper* if the expected payment is uniquely maximized when $q = p$. For instance, it is easy to verify that the QSR is strictly proper for eliciting probability p. However, both the definition of properness and this rule itself can be generalized to settings with more outcomes. If X is a real-valued random variable (say, the number of students who will pass an exam) and the participant is asked to announce their belief about the mean number q, then the scoring rule would provide different payments for each realization of the number of students who pass. This resulting rule would be strictly proper for eliciting the mean if announcing $q = \pi$ (their true belief about the mean) truthfully maximizes their expected payoff over all possible outcomes of X.

The QSR, when extended to real-valued X is the function $S_x(q) = 1-(x-q)^2$ and it is strictly proper for eliciting the mean.[6] However, other scoring rules are proper for other statistics. For example, if the payment is $1-|x-q|$ then the optimal report is the median, and if it is $1-\mathbf{1}_{(x\neq q)}$ (meaning \$1 if $x = q$ and \$0 otherwise) then the optimal report is the mode. Lambert et al. (2008) provide a complete characterization of the various statistics that can be elicited in this way. Gneiting and Raftery (2007) survey the general theory

[6] The probabilistic version is really a special case of the real-valued version of the QSR since E can be represented by a Bernoulli random variable where $X = 1$ if E occurs and $X = 0$ otherwise. In this case, the expected value for the participant of X is simply their probability of E.

of scoring rules, which was first developed by Savage (1971) and later generalized by Schervish (1989).

2.4.1. Accounting for Risk Preferences

These scoring rules described above are all said to be (strictly) proper because truthful reporting (uniquely) maximizes expected earnings. That means they are incentive compatible — participants will report truthfully —*if the participants are risk-neutral expected utility maximizers*. But risk neutrality is well known to be a tenuous assumption. So, what can be done to accommodate risk preferences?

The problem with these dollar-denominates scoring rules is that they may not be incentive compatible if the participants are not risk-neutral expected utility maximizers. There are three solutions that have been discussed for this problem. The first, offered by Savage (1971), is simply to shrink the size of the payments. The logic is that people should be approximately risk neutral over very small stakes. However, unfortunately, there is ample evidence that this is not the case in practice.

A second solution, developed primarily by Viscusi and Evans (1998, 2006), Offerman et al. (2009), and Andersen et al. (2014) is to use the original dollar-denominated scoring rule, but to try to estimate the degree to which a given participant will misreport. Specifically, we can estimate a function that maps their true belief into their actual report. Using that, we can debias their report by inverting that function to uncover their true belief. Offerman et al. (2009) do this directly by first measuring how much a participant misreports when there is a correct, objective probability. Andersen et al. (2014) instead assume a structural model of preferences and use an array of lottery choices to estimate the participant's Bernoulli utility index and probability weighting functions. From this they can predict the degree of misreporting on any given scoring rule and use that prediction to debias actual reports. The downsides to these procedures are that they require both extra data and extra assumptions. However, it does appear that correcting the QSR for risk (or loss) aversion improves performance (Offerman et al., 2009; Harrison et al., 2014; Offerman and Palley, 2016).

The third solution is to pay in probabilities instead of dollars. The idea that paying in probabilities induces risk neutrality is a direct application of the Von Neumann and Morgenstern (1944) observation that probabilities can be used as a measure of cardinal utility (see also Raiffa, 1968). Savage (1971) first suggested its use for scoring rules, crediting Smith (1961) with the idea (who in turn says it was adapted from Savage (1954)). For the special case of the quadratic scoring rule the idea was rediscovered independently by both Allen (1987) and McKelvey and Page (1990). The latter appears to be the first to implement the procedure. Other early applications (outside of the belief elicitation context) are Roth and Malouf (1979) and Berg et al. (1986).

The use of binary lotteries to linearize preferences is often called the *binary lottery procedure* [BLP]. In this sense, the procedure of paying the QSR in probabilities could be called the BLP-QSR. However, Hossain and Okui (2013) have popularized the term "binarized" for scoring rules that pay in probabilities. Thus, the quadratic scoring rule (QSR) becomes the binarized quadratic scoring rule (BQSR). Given the prevalence of this term, we will use it throughout the chapter. Specifically, fix a single prize (say \$10) and let s_x represent the probabilities that the participant wins that prize conditional on x (whether the event occurs or the realization of X). For eliciting a belief with a binarized scoring rule, for example, equation 1 represents the overall probability of winning the prize when the true belief is p.

Binarized scoring rules are incentive compatible if we assume a particular version of the reduction of compound lotteries. Specifically, we require that the participant multiply their subjective probability of each event occurring times the objective probability of payment dictated by the scoring rule. But, since only two payments are used, this reduction only needs to occur for two-outcome lotteries. This theoretical requirement was described by Harrison et al. (2013, 2014) and formalized by Healy and Kagel (2023), who refer to it as "subjective-objective reduction," or S-O reduction. A formal discussion of this assumption is presented in the next Subsection 2.4.2.

Regarding the effectiveness of paying in probabilities, early experiments found that it does not induce risk neutrality (e.g. Walker et al., 1990; Cox and Oaxaca, 1995; Selten et al., 1999).[7] On the other hand, most tests find that the binarized quadratic scoring rule (BQSR) performs relatively better than the dollar-denominated QSR (Hossain and Okui, 2013; Harrison et al., 2014; Erkal et al., 2020; Healy and Kagel, 2023), especially for risk-averse subjects.

2.4.2. *Incentive-Compatibility of Binarized Rules*

For the BQSR, if the participant reports a belief q and E occurs, then the participant is paid the high prize $\overline{x}$ with probability $s_1(q) := 1-(1-q)^2$ (and $\underline{x}$ with probability $1-s_1(q)$). If E^c occurs, then the participant receives the high prize with probability $s_0(q) := 1-(0-q)^2$ (and the low prize with probability $1-s_0(q)$). Thus, the scoring rule generates a different two-stage compound lottery for each possible announcement.

[7]See Harrison et al. (2013) for an excellent review of early tests of the binarization procedure.

Figure 2 illustrates the compound lottery generated by an announcement of q when the participant's true belief is p. Importantly, this compound lottery contains one stage that involves subjective uncertainty (whether or not E occurs) and one stage involving objective randomness (whether $\overline{x}$ is paid or not).

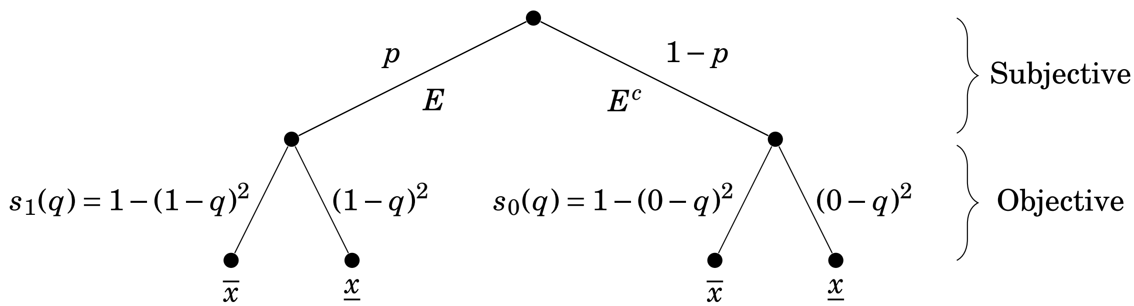


FIGURE 2. The compound lottery generated by the binarized quadratic scoring rule (BQSR) when the participant has true belief p and announces q. The two possible prizes are $\overline{x}$ and $\underline{x}$, with $\overline{x} > \underline{x}$

To analyze whether the BQSR is incentive compatible, we first need to understand how participants evaluate these compound lotteries. One hypothesis is that they only care about maximizing the overall probability of winning the high prize, meaning they choose q to maximize

$$U(q|p) = p\,s_1(q) + (1-p)s_0(q). \tag{3}$$

This hypothesis assumes that the participant seamlessly integrates their subjective beliefs p and $1-p$ with the objective payment probabilities $s_1(q)$ and $s_0(q)$, thus reducing the compound lottery to a simple one-stage lottery. We therefore refer to this as *subjective-objective reduction*.

Axiom 8. (*Subjective-Objective Reduction*) A participant satisfies **subjective-objective reduction (S-O reduction)** if they evaluate compound lotteries of the form shown in Figure 2 according to their overall reduced probability $U(q|p) = p\,s_1(q) + (1-p)s_0(q)$. In other words, they prefer to report q' over reporting q if and only if $U(q'|p) \geq U(q|p)$.

Fact 1. If the participant satisfies S-O reduction, then the BQSR is incentive compatible.

Mathematically, incentive compatibility under S-O reduction can be seen directly by taking the first-order condition of equation (3) with respect to q, and noting that (3) is strictly concave in q. Incentive compatibility can also be shown graphically, as in Figure

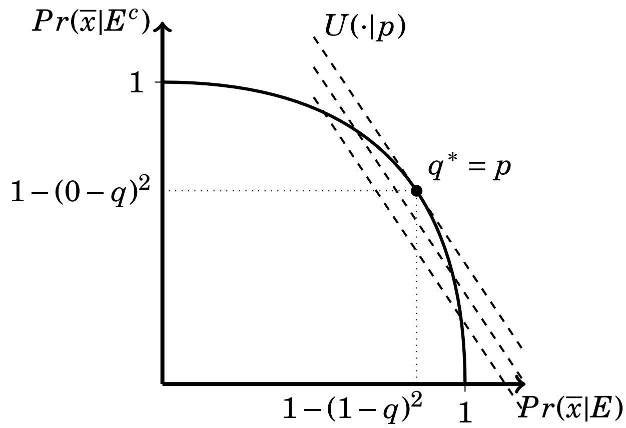


FIGURE 3. The state-contingent probabilities of winning the high prize for each possible announcement in the BQSR forms a strictly convex curve. Under S-O reduction, indifference curves are linear and the optimal announcement is truth-telling.

3. Each announcement q generates two different probabilities of winning the high prize: $s_1(q)$ if E occurs and $s_0(q)$ if E^c occurs. If we plot the point $(s_1(q), s_0(q))$ for every possible $q \in [0,1]$, we trace out a strictly concave curve spans from the point $(0,1)$ to the point $(1,0)$. Now, S-O reduction says that any point (s_1, s_0) in this space must be evaluated by

$$U(s_1, s_0|p) = p\,s_1 + (1-p)s_0.$$

In other words, the participant has linear indifference curves with slope $-p/(1-p)$. The BQSR is constructed so that the tangency between the strictly concave curve and this linear indifference curve always occurs at $q = p$, ensuring that truth-telling yields the participant's most preferred compound lottery from the curve.

From Figure 3 it is easy to see how other incentive-compatible scoring rules can be constructed: Simply draw a strictly concave, decreasing curve in (s_1, s_0)-space whose slope ranges from zero to negative infinity. Then, for each point on that curve, if the slope at that point is $-q/(1-q)$, let that point be the compound lottery paid to a participant who announces q. We refer to such scoring rules as *proper*. For instance, the *Binarized Spherical Scoring Rule* is given by

$$s_1(q) = \frac{q}{\left(q^2+(1-q)^2\right)^{\frac{1}{2}}} \qquad\qquad s_0(q) = \frac{1-q}{\left(q^2+(1-q)^2\right)^{\frac{1}{2}}},$$

which traces out a quarter-circle in (s_1, s_0)-space. Thus, it is also a proper scoring rule. The spherical rule is more concave than the BQSR for extreme reports (close to zero or one) but less concave for reports closer to 0.5. In other words, it gives stronger penalties for small deviations from truth-telling when beliefs are extreme, but weaker incentives when beliefs are intermediate. We discuss additional scoring rules and compare their incentives in the next section. Experimental tests of these "non-quadratic" binarized scoring rules are lacking; an interesting open question is whether altering the strength of incentives (by altering the concavity) over different ranges would have a significant impact on behavior. Without concrete evidence that some other rule is better suited to certain applications, our suggestion is to stick with the popular (and relatively simple) quadratic rule.

Figure 3 also highlights the necessity of S-O reduction: if a participant does not satisfy S-O reduction, then their indifference curves will not be linear. If this curvature in the indifference curve alters its tangency point, then their most preferred belief report will no longer be truthful. Now, it is possible that the curvature in their indifference curves occurs at different (s_1, s_0) points and therefore does not affect the tangency with the BQSR shown in Figure 3. But we can always construct a different binarized scoring rule by drawing a new strictly concave curve through the region where the indifference curves are nonlinear, and for this new scoring rule the participant will prefer to misreport. Thus, any instance of nonlinearity in preferences must lead to misreporting for *some* scoring rule.

Fact 2. If, for a given participant, we assume that every proper binarized scoring rule is incentive compatible, then that participant's preferences must satisfy S-O reduction.

See Healy and Kagel (2023) for a formal development of this argument.

2.4.3. Comparison of Scoring Rules

It is easy to prove that if the scoring rule for a probability is (strictly) proper, the expected payment at truth-telling is given by $G(p) = ps_1(p) + (1-p)s_0(p)$ must be (strictly) convex. The famous characterization of Savage (1971) and Schervish (1989) is that *any* strictly convex function $G(p)$ with domain $[0,1]$ can be used to construct a strictly proper scoring rule.[8] Thus, there are as many strictly proper scoring rules as there are strictly convex functions over $[0,1]$! Section 2.4.3 elaborates on this result.

A list of well-known proper scoring rules for eliciting a single probability appears in Table 1; see Gneiting and Raftery (2007) for more information. Interestingly, the MPL and BDM mechanisms can be "reduced" to a scoring rule, and we have included this

[8]If G is differentiable then this is done by setting $s_1(q) = G(q) + (1-q)G'(q)$ and $s_0(q) = G(q) - qG'(q)$.

"reduced MPL/BDM" it in Table 1 for comparison.[9] It is equivalent (assuming reduction of compound lotteries) to saying "We will flip a coin. If heads, we will ignore your report and you get the high prize if and only if E occurs. If tails, we will pay you according to the BQSR." Thus, the Reduced MPL/BDM has half the incentives of the BQSR. In addition, reducing the MPL or BDM in this way destroys its dominance property for incentive compatibility (see Section 2.3). For these two reasons we do not recommend reducing MPLs to a scoring rule.

Scoring Rule	$\mathbf{s_1(q)}$	$\mathbf{s_0(q)}$	**Strength of Truth-Telling Incentive**
Quadratic	$1-(1-q)^2$	$1-(0-q)^2$	2 (for all p)
Reduced BDM/MPL	$\frac{1}{2}+\frac{1}{2}\left(1-(1-q)^2\right)$	$0+\frac{1}{2}\left(1-(0-q)^2\right)$	1 (for all p)
Spherical	$\frac{q}{\sqrt{q^2+(1-q)^2}}$	$\frac{1-q}{\sqrt{q^2+(1-q)^2}}$	$\left(p^2+(1-p)^2\right)^{-3/2}$
Pseudospherical ($\alpha>1$)	$\frac{q^{\alpha-1}}{(q^\alpha+(1-q)^\alpha)^{(\alpha-1)/\alpha}}$	$\frac{(1-q)^{\alpha-1}}{(q^\alpha+(1-q)^\alpha)^{(\alpha-1)/\alpha}}$	$(\alpha-1)\frac{p^{\alpha-2}(1-p)^{\alpha-2}}{(p^\alpha+(1-p)^\alpha)^{(2\alpha-1)/\alpha}}$
Logarithmic*	$\ln(q)$	$\ln(1-q)$	$\frac{1}{p}+\frac{1}{1-p}$

TABLE 1. A list of well-known strictly proper scoring rules. The strength of truth-telling incentives is measured by the concavity of expected payoffs at the truth.
* The logarithmic scoring rule cannot be binarized.

How should we choose among the many options available for eliciting a probability with a scoring rule? One criteria is the strength of incentives they provide.

Lemma 1. (*Scoring Rules - Strength of Incentives*) For any differentiable proper scoring rule, the strength of incentives for truth-telling is given by $G''(p)$.

Proof in Appendix Section A.1.1

Lemma 1 shows that, for any proper scoring rule, the strength of incentives (in terms of the concavity of payoffs, or how fast the the payoffs drop as one moves away from

[9] In the MPL and BDM, there are two sources of objective randomness: which row is chosen for payment, and the lottery offered in Option B on each row. Thus, each possible report q generates a compound lottery. We can reduce this compound lottery into one with the same structure as a binarized scoring rule. The result is what we call the "Reduced MPL/BDM" and the probability-of-payoff functions are given by $s_1(q)=\frac{1}{2}+\frac{1}{2}\left(1-(1-q)^2\right)$ and $s_0(q)=\frac{1}{2}\left(1-(0-q)^2\right)$.

truth-telling)) for some belief p is measured by the convexity of the $G(p)$ function at the point p.[10]

Lemma 2. (*Average Strength of Incentives*) For any proper binarized scoring rule, the average strength of incentives must be weakly less than two.

Proof in Appendix Section A.1.2

Proposition 4. (*BQSR Maximizes Minimum Incentives*) The binarized quadratic scoring rule maximizes the minimum strength of incentives among all proper binarized scoring rules.[11]

Proof. The BQSR has $G(p) = 1 + p^2 - p$, so $G''(p) = 2$ for all p. Thus, the minimum strength of incentives is equal to the average, meaning it is maximized. $\square$

In contrast to the BQSR, the spherical scoring rule gives stronger marginal incentives for beliefs near one half. Thus, it might be useful in cases where the researcher thinks participants are likely to report such beliefs. Conversely, extreme beliefs are more likely for participants, then the pseudospherical scoring rule with α near one will provide the strongest incentives.

There have been several tests comparing different dollar-denominated scoring rules without any correction for risk aversion. A simple summary seems to be that the logarithmic rule gives (weakly) better calibrated beliefs and posteriors closer to Bayes's rule (Phillips and Edwards, 1966; Palfrey and Wang, 2009) compared to other rules, which could be due to its stronger incentives, but more data on the question would be useful before making a definitive recommendation.

An interesting open question is how the dollar-denominated logarithmic scoring rule—which has powerful incentives but cannot be binarized—compares to the BQSR. To our knowledge, there has not been an experiment that directly compares these two mechanisms, and a cross-study comparison is difficult since the domains in which the logarithmic rule has been studied and the criteria for its success differ substantially from existing tests of the BQSR. Another promising but unexplored direction would be to study risk-adjusted logarithmic scoring rules, perhaps compared against the BQSR.

[10]Formally, if $G(q|p) = p\,s_1(q)+(1-p)s_0(q)$ then Lemma 1 shows that for any incentive compatible scoring rule, $\partial^2 G(q|p)/(\partial q)^2$ at $q = p$ equals $-G''(p)$. The result is akin to the envelope theorem. Healy and Leo (2024) also provide a justification for using G'' in terms of the density used for each probability in the MPL-like mechanism that reduces to the desired scoring rule.

[11]This result is similar to Proposition 1 of Schlag et al. (2015). We note that while the logarithmic scoring rule gives strictly stronger incentives for every p, it cannot be binarized.

3. Should I Use (and Emphasize) Pecuniary Incentives?

Although this chapter focuses entirely on incentivized elicitation, maybe it is sufficient (or even beneficial) to rely simply on unincentivized belief reports (often called "introspection") or hypothetical incentives?

The obvious argument in favor of incentives is that it induces participants to take the time to report truthfully, rather than to submit an easy or default answer. Similarly, it might improve the actual beliefs by encouraging them to deliberate consciously, leading to more precise reports. For example, unincentivized participants may not have thought about their precise belief and default to 50%, but adding incentives may help them refine their belief to 55% or 60%. As always, incentives can also help overcome other motives (Wilde, 1981; Smith, 1982), such as using stated beliefs to justify selfish behavior in other decisions (Blanco et al., 2010) or wanting to appear more confident than they are.

One argument against incentives is that experimental participants are generally averse to lying even when it would benefit them (Gneezy, 2005; Fischbacher and Föllmi-Heusi, 2013), so the pecuniary incentives to induce truth-telling are unnecessary. An even stronger argument is that the mechanism used may not be incentive compatible for actual participants, perhaps because the psychological effects of a complicated elicitation mechanism—especially one with flat payoffs at the maximum—can actually crowd out the extrinsic incentive to report truthfully. For example, Danz et al. (2022) show that providing on-screen calculators at the time of elicitation can lead to significant rates of misreporting, but this misreporting rate is greatly reduced when the calculator is removed and participants (who are still incentivized) are told only that truth-telling is in their best interest.

What does the data say? Several studies have shown that incentives improve beliefs according to various metrics. In the domain of updating, Phillips and Edwards (1966), Grether (1980), and Wright and Anderson (1989) find that biases in Bayesian posteriors are reduced when reports are incentivized.[12] One explanation is that incentives push some participants to think harder about their posterior belief. This was confirmed by Burfurd and Wilkening (2022), who find that people who have a basic grasp of Bayesian updating have lower errors when reports are incentivized. For people who lack Bayesian intuition, the results are more nuanced: their performance across mechanisms is similar

[12] Beach and Wise (1969) measure posterior beliefs either using unincentivized reports or choices between hypothetical bets. Accuracy is equal across methods. They also show that with sequential signals, step-by-step updating performs the same as all-at-once updating.

when they should be updating their belief, but they actually perform better without incentives when they receive an uninformative signal and, therefore, should simply report the prior.

For beliefs without updating, Wright and Aboul-Ezz (1988) find that beliefs about real-world variables, such as average GMAT scores, are more accurate with incentives. Incentives have also been shown to improve belief reports in games (Gächter and Renner, 2010; Wang, 2011). Harrison (2014) documents complex patterns of hypothetical bias in unincentivized belief elicitations, though he does find that adding a large flat payment to unincentivized reports can effectively eliminate the gap (see Haghani et al., 2021 for a survey of hypothetical biases). There is also evidence that incentives can improve the belief formation process (Rutström and Wilcox, 2009), consistent with the idea that belief formation is an effortful task. For example, without incentives, subjects are more likely to report default or focal values such as 50% or 100% (Massoni et al., 2014; Burfurd and Wilkening, 2022). Similarly, Grether (1992) found that participants without incentives were far more likely to say that event E is more likely than its complement but then report a belief of less than 50% for E.

Some studies show no effect of incentives, such as Sonnemans and Offerman (2001) and Trautmann and van de Kuilen (2015). Massoni et al. (2014) and Hollard et al. (2016) find that unincentivized beliefs in a classic signal detection task are more accurate than with a (dollar-denominated) quadratic scoring rule [QSR] —a result that could partly be due to risk aversion— but that a Becker–DeGroot–Marschak [BDM] mechanism performed as well as (Massoni et al., 2014) or even better than (Hollard et al., 2016) unincentivized reports. Armantier and Treich (2013) also show quite convincingly that scoring rule incentives led to worse reports in their domain, although again this could be driven by risk aversion. Interestingly, Trautmann and van de Kuilen (2015) find that unincentivized beliefs about complementary events are more likely to add to 100% than with either of their incentivized methods. Almost all studies agree that incentives reduce noise, at least weakly; see Camerer and Hogarth (1999), Gächter and Renner (2010), and Trautmann and van de Kuilen (2015) for example.

But what should we make of the Danz et al. (2022) results? Participants in their study are always incentivized via the same mechanism: the binarized quadratic scoring rule [BQSR]. The only difference between treatments is whether the incentives are shown on-screen or not. One compelling explanation they offer is that, by highlighting the incentives, participants become aware of and respond to an intrinsic asymmetry in payoffs: by decreasing their belief report from, say, 90% to 70% they reduce their payoff probability if the event occurs by only nine percentage points, but their payoff probability if the event does not occur improves by 32 percentage points. This asymmetry is necessary for incentive compatibility — a subject with a true belief of 90% should put far more weight on the payoffs conditional on the event occurring — but perhaps highlighting this trade-off alters the way subjects integrate these two payment probabilities. In particular, they may push towards a report of 50% to equalize payment probabilities across states. We discuss the incentives of the BQSR further in Section 2.4.

Danz et al. (2022) and Healy and Kagel (2023) offer a possible solution to this problem: remove the calculator from the decision screen and only explain the incentives in the opening instructions. Thus, subjects are still incentivized and those incentives are clearly explained, but the incentives do not distract subjects when choosing their report. An alternative method (which has been used but not tested formally) is to provide a button that subjects can click to view details about incentives. In our own limited experience, most subjects never click the button and simply trust a statement that says "you maximize your overall chance of payment if you report truthfully."

We end this section by noting that it can be difficult to test the effectiveness of belief elicitation methodology. In practice, the methods in this chapter are used to elicit beliefs that are unknown to the experimenter and may need to be refined through substantial reflection. However, this means that there is no source of truth to compare the elicited beliefs against to determine effectiveness. For this reason, many tests elicit beliefs about an objective event where the probability is well known or given to the participants directly. We think, especially the the later setup puts participants in a very unnatural scenario– being asked to do something trivial but given non-trivial and often complex incentives to do it. We think this dissonance could cause behavior that would make tests unreliable relative to how the same procedures would perform in a scenario where they are being used to elicit a belief that is unknown to the experimenter and requires non-trivial reflection.

OUR RECOMMENDATION #1

Incentives

(1) **Use incentives.** Although not unanimous, most studies show that incentives help increase accuracy (perhaps through improved belief formation) and reduce noise. As Holt and Smith (2016) write: “In the absence of a reliable set of criteria to determine when incentives matter and when they do not, it seems prudent to use incentives.”

(2) **Don’t put the incentives on the decision screen.** It appears sufficient to explain the incentives and/or provide a calculator only in the opening instructions (Danz et al., 2022; Healy and Kagel, 2023) or in an optional pop-up window accessed by clicking on a button. They can even be removed entirely (Danz et al., 2022).

(3) **But say that truth-telling is optimal.** Specifically, your decision screen can include the phrase “Remember: you maximize your overall chance of being paid when you report truthfully” or “you can secure the largest chance of winning the prize by reporting your most-accurate guess.” If you use a Multiple Price List [MPL], you can strengthen this to “Reporting truthfully is the only way to guarantee that you get your preferred choice on every single row of the choice list.”

4. Choosing What to Elicit

The following is a menu of the various beliefs and statistics of beliefs that can be elicited, with links to the relevant subsection of Section 5 that describes the mechanisms available for that statistic. Readers who are confident about which statistic they want to elicit can jump immediately to the relevant subsection, though the remainder of this section has additional thoughts and advice about that decision.

MENU

Options for Belief Elicitation

- **Abstract Events** (*"Will E Occur?" or "Is E True?"*)
  - Eliciting the Probability of an Event (See Section 5.1)
  - Eliciting the Probability of Multiple Events (See Section 5.2)
  - Eliciting the Modal Event (See Section 5.3)
  - Eliciting a Ranking of Events ("Qualitative Probabilities") (See Section 5.4)
- **The Value of a Number** (*"What is the value of X?"*)
  - Eliciting the Modal Set, Mode, or Modal Interval (See Section 5.5)
  - Eliciting the Median and other Quantiles (See Section 5.6)
  - Eliciting a Confidence Interval (See Section 5.7)
  - Eliciting the Mean (See Section 5.8)
  - Eliciting an Entire Distribution (See Section 5.9)
  - Approximating Other Statistics of a Distribution (See Section 5.10)

4.1. Beliefs about Events versus Beliefs About a Number

The first step in eliciting beliefs is determining what type of belief — or what statistic of a belief distribution — you are trying to elicit. This decision depends on the nature of the belief in question and your research goals. For outcomes that are binary in nature like "will event E happen" or "is proposition P true" the participant's entire belief is encapsulated by a single probability. These are beliefs about *events*. But, for questions *about a number* such as "What will be the value of X?" There can be many possible outcomes, and a belief about X can be best understood as a distribution over those outcomes: a "belief distribution" or "subjective distribution".

Belief distributions can be difficult to elicit in their entirety (although we offer some options in Section 5.9). Thus, in these cases, it is often convenient to elicit some statistic of the belief, such as the mean, median, or mode of X. In the menu above, we categorize the different options available. The first set of options concern beliefs about single events, or collections of events. The second set deals with beliefs about a number.

Our dichotomy of "events" versus "numbers" is meant to follow distinct strands of elicitation methodology, but is admittedly not precise. Some beliefs about numbers are, in fact, beliefs about an event. For example, whether a student gets a score in the range $[90\%, 100\%]$ is an event that depends on the realization of a number but the outcome is still binary, the student either does or does not score in that range. In cases like this, the belief is simply a probability and the "beliefs about events" sections are appropriate.

In addition, almost all of the "beliefs about a number" methods can also be applied to non-numeric ordinal categories, such as beliefs about whether the weather tomorrow will be judged to be cold, warm, or hot by a third party observer. The mode, median, and quantiles of beliefs over these categories are well defined and can be elicited exactly as described. Only the mean (and other moments) cannot be elicited, since it does not exist without a nominal scale.

OUR RECOMMENDATION #2

For beliefs about a number, avoid eliciting the mean.
Consider eliciting the mode (or modal interval) instead.

There are some instances where the mean may be the most appropriate statistic to elicit (see Section 5.8). However, in our experience, researchers often just want some measure of the "center" of a belief distribution. and are not interested in (and will not have information about) possible differences between the mean, median, mode, or other statistics. When surveying incentive compatibility conditions for the various mechanisms, it becomes clear that eliciting the mean requires extra assumptions on the participant's preferences. Roughly, this occurs because, to our knowledge, the mean cannot be defined from some indifference point between observable events and objective lotteries. If the goal is to minimize those assumptions, then we recommend considering another statistic such as the mode or median.

We believe that the mode is the easiest and most reliable statistic to elicit: Simply ask the participant to guess the value of the number (for example, guess someone's score on an exam) and pay them a fixed prize if and only if their guess is correct. Not only is this easy to explain, but it is also incentive compatible under much weaker assumptions than any of the mechanisms to elicit other statistics (see Section 5.5).

The downside of this procedure is that, if there are many possible values of the unknown number, the participant is unlikely to be paid. A simple solution is to pay the fixed prize if and only if their guess is within a certain distance (say, d) from the true value. This elicits the (center of the) participant's belief about the most likely interval of length $2d$. Although this may be slightly different from the modal point, in practice the difference is small, especially when d is small and the beliefs are roughly symmetric.

If the mode is not a suitable measure of the central tendency for the belief in question, we suggest eliciting the median or other quantiles (see Section 5.6).

4.2. An Important Distinction: Frequencies versus Probabilities

It is important to distinguish between population frequencies and individual probabilities because in many settings it would be possible to elicit beliefs about either quantity.

For example, if 100 people have taken a pass/fail exam, you could either elicit beliefs about the fraction (out of 100) that passed the exam or you could elicit a single probability that one randomly selected person passed the exam. These are different objects (see Figure 4 for an illustration), and the relationship between them may not be so cut and dry, so if the distinction matters for your research question then it becomes important to be clear about which you are eliciting and how.

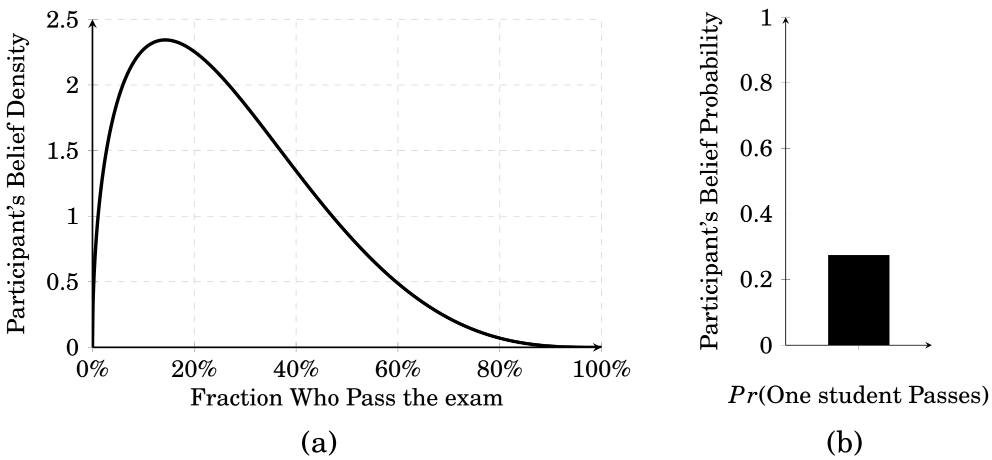


FIGURE 4. An illustration of the difference between (a) beliefs over the frequency of successes and (b) the probability of a single success.

Furthermore, beliefs about the frequency are an entire distribution over $[0,100]$, not just a single probability, so eliciting such beliefs forces you to decide which statistic of that distribution is relevant for your results. Is it the mean? The median? The mode? Sometimes researchers will simply ask the participant to make a "guess" of the actual fraction, but how they incentivize that guess affects which statistic is being elicited. For example, paying if their guess is correct elicits the mode (see Section 5.5), giving them a quadratic penalty elicits the mean (see Section 5.8), and giving them a linear penalty elicits the median (see Section 5.6.

Importantly, it is often possible to choose between a frequency belief and a probability belief. If your data consists of a population of test-takers, for example, you can switch to eliciting a single probability if you ask instead "what is the probability that a single, randomly chosen person passed the exam?" And sometimes you can also switch from eliciting a probability to eliciting a frequency if the event can be repeated. As Charness et al. (2021) illustrates, instead of asking for the probability that a red ball will be drawn from an urn, you could instead draw the ball N times with replacement and ask for their modal belief about the number of draws that are red. Assuming N is reasonably large and that participants' beliefs respect both independence of draws and the weak law of large numbers, their modal frequency should be close to their probability in the single draw.

One satisfying solution is to switch from eliciting information about the population frequency and instead elicit the probability of a single, randomly chosen student passing the exam. This is now a single event whose belief is characterized entirely by one probability, as in panel (b) of Figure 4. It is relatively easy to elicit, easy to explain, and unambiguous to interpret.

Moreover, if desired, this probability of a random student passing the exam can be extrapolated to the entire population. Since the target of the elicitation is a randomly chosen student, it is reasonable to assume that the belief about them passing is identical and independent across test-takers (or, more precisely, across the identity of the test-taker randomly chosen for payment).[13] In that case their beliefs about the population frequency should form a binomial distribution. And the mean of that distribution should equal their reported probability about the single individual passing the exam.[14] Thus, the probability report can be interpreted as a population-level belief if desired. We discuss mechanisms for eliciting a single probability in Section 5.

Schlag and Tremewan (2021) suggest going in the opposite direction: rather than eliciting a probability of a single student passing the exam, elicit the *mode* (or, more precisely, a modal interval) of the participant's beliefs about the frequency. This is done by asking them to predict the number of students who pass (out of 20) and rewarding them with two euros if and only if that prediction is correct. When comparing this to an elicitation of a probability, they find no real difference in overall performance but do find that participants are faster and self report a better understanding when reporting a frequency. Other measures of decision confidence yielded no significant differences.

[13]Specifically, if the participant believes that different students have different probabilities of passing the exam, then their probability report about a single student should represent the average of those individual probabilities.

[14]Interestingly, Leo and Stelnicki (2024) show that this is often not the case: in their data the actual distribution over population frequencies has fatter tails than the binomial distribution implied from their single elicited probability.

OUR RECOMMENDATION #3

For beliefs about a frequency, consider eliciting the probability of a single, randomly chosen outcome instead.

As the preceding discussion highlights, eliciting the probability of a single student passing an exam is a simple sufficient statistic for the population percentage of students who pass the exam. This avoids the sometimes-arbitrary decision about which statistic of the distribution over population frequencies to elicit.

4.3. Coarse Elicitation

Often, it is possible to measure a statistic both precisely and coarsely. A wonderful example of coarse beliefs was the pioneering weather forecasting method of Cooke (1906), who categorized each weather forecast into one of five possible bins that ranged from "the barest possibility" to "absolute certainty." Although Cooke was not explicitly incentivized, we could imagine asking a participant in which bin their belief lies, and trying to find a way to incentivize that report to be truthful. Wallsten et al. (1993) show that coarse elicitation can also be performed verbally by having subjects choose from statements such as "highly probable," each of which maps to a given probability number using a mapping generated from previous studies (Wallsten et al., 1986, e.g.). Participants are aware of the mapping and know that ultimately they are paid as if they submitted the corresponding probability to a given mechanism.

Coarse elicitation often has the benefit of strengthening incentives. For instance, in the Multiple Price List [MPL] mechanisms discussed in Sections 5.1.1, 5.6.1, 5.8.1, coarse elicitation can be achieved by using fewer rows of the MPL. This, in turn, increases the probability that any particular row is chosen for payment, and thus magnifies the incentives involved in the remaining rows. For the scoring rules presented in Sections 5.1.4 and 5.8.3 which are suitable for use when eliciting coarsely, limiting participants to a small set of possible reports means that they cannot deviate marginally from their true belief, and therefore any deviation from truth-telling will be "large", improving incentives against misreporting.

OUR RECOMMENDATION #4

Consider Coarse Elicitation

In many cases, a precise probability is not necessary for the research question at hand. Often it is sufficient to know whether the participant believes the probability is above or below some threshold. This can be achieved simply by asking participants if they would rather receive a prize if the event occurs or be paid a prize with a fixed probability $X\%$ (the threshold). When a single threshold is not enough, it may be enough to categorize the belief into a small number of categories; this can be achieved either by using an MPL with a small number of rows or the more recently developed ternary lists which reduce the number of questions participants must respond to achieve the same categorization. Coarse elicitation can also be used when eliciting quantiles (see Section 5.6) by reducing the number of rows and identifying coarse ranges that a quantile must fall within.

At a higher level, entire belief distributions can also be elicited coarsely by using less precise approximations (see Section 5.9). When eliciting the CDF, a coarser approximation can be achieved by eliciting fewer quantiles. When eliciting the PDF, a coarser approximation can be achieved by eliciting probabilities from larger bins.

When appropriate, in Section 5, we discuss how to use a mechanism to elicit coarsely below the pop-out box for that mechanism.

5. The Mechanisms

This section contains all of the belief elicitation mechanisms organized into subsections by the statistic they elicited. Refer back to Section 4 for a complete list of the statistics covered. When there are several mechanisms for the same statistic, the subsection for that statistic includes a menu linking to each one.

5.1. Eliciting the Probability of an Event

MENU

Eliciting the Probability of an Event

- Multiple Price Lists (See Section 5.1.1)
- Ternary Price Lists (See Section 5.1.2)
- Single Response BDM Mechanisms (See Section 5.1.3)
- Scoring Rules (See Section 5.1.4)

The problem of eliciting a participant's belief about the probability of an event is by far the most studied problem in the belief elicitation literature. There are many mechanisms for eliciting the probability of an event, including (1) multiple price lists [MPL], (2) the related ternary price lists [TPL], (3) single-response Becker–DeGroot–Marschak [BDM] mechanisms, and (4) scoring rules.

We can start comparing these options by looking at their incentive-compatibility requirements. As discussed in Section 2.3, incentive compatibility of the MPL / BDM, and here also the TPL, rely on an assumption known as *Statewise Monotonicity* (See Axiom 7 in Section 2.3). This is a relatively weak assumption and is exactly the same condition that is required for an experiment with multiple decisions to be incentive compatible (Azrieli et al., 2018, 2020). Scoring rules rely, at least, on a form of compound lottery reduction called *Subjective-Objective [S-O] Reduction* (See Axiom 8 in Section 2.4). Healy and Kagel (2023) show that statewise monotonicity is strictly weaker than S-O reduction and so, in this sense, the MPL, TPL, and BDM are superior to binarized scoring rules (and in turn dollar-denominated scoring rules) in terms of incentive-compatibility.

Another advantage of the MPL and BDM mechanisms is that they may help reduce hedging incentives (Möbius et al., 2022). Suppose for example a participant will take a difficult exam and is also asked their probability that they will pass the exam. Suppose if they try hard, then they could achieve a 50% chance of passing. If they truthfully report that belief, their overall probability of winning a prize from the elicitation is 62.5%. Now they may be tempted to fail the exam on purpose so that they can confidently report a 0% chance of passing. However, this reduces their probability of winning a prize to 50%. Thus, the MPL and BDM remove hedging incentives.[15] This property is not true for many scoring rules, including the binarized quadratic scoring rule [BQSR].

The MPL and BDM have one theoretical disadvantage to the BQSR: lower incentives. Healy and Leo (2024) show that the BQSR has double the incentive strength of the MPL and BDM. To address this, they devise a new type of MPL, the Ternary Price List [TPL] that maintains the strong incentive properties of the MPL but has the exact same strength of incentives (both marginal and absolute incentives) as the BQSR.

Unfortunately, the local incentives provided even by the BQSR or TPL can be quite weak. For example, in the BQSR, a participant who believes that an event will occur with probability 0.5 would expect to win the prize with a chance of 0.75 if they say their belief is 0.5, but the probability drops only to 0.6875 if they say their belief is 0.75 or 0.25. This point was emphasized by Danz et al. (2022). Understanding how to navigate these weak incentives and design rules that provide incentives for precision where it is required is a worthwhile goal. In the absence of this, boosting incentives by scaling up the prize that results from any of the mechanisms discussed below will always improve the incentive for precision. We suggest that when measuring beliefs is an important part of your research question, *incentivize them generously*.

For empirical comparisons, there are surprisingly few papers that test across these mechanisms. Hollard et al. (2010) find that the single-response BDM gives more accurate beliefs about one's own performance than the dollar-denominated quadratic scoring rule [QSR], a result that could be driven by risk aversion.[16] Trautmann and van de Kuilen (2015) compare the MPL, dollar-denominated QSR, and a QSR adjusted for risk aversion via the Offerman et al. (2009) method of using other elicitations with a known "correct" probability to identify bias in reports. The participants reported the probabilities that their opponent would choose different strategies in the ultimatum game.[17] They find no real differences between the mechanisms, both in terms of the accuracy of the beliefs and in using those beliefs to predict strategy choices. They do find that

[15] If, however, the participant feels that high effort would introduce more ambiguity, and if they are very ambiguity averse, then they may still prefer to hedge. In that case, however, they do not have a true probability belief; see Section 2.1

[16] Huck and Weizsäcker (2002) compare the dollar-denominated QSR to a single-response BDM that elicits an indifferent dollar amount instead of an indifference probability. Neither is incentive compatible under risk aversion. In the latter a mean frequency is elicited, while in the former they elicit the probability of a single randomly chosen trial. They find that the dollar-denominated QSR for frequencies outperforms the dollar-valued BDM for a probability in terms of accuracy.

[17] This opens the door for hedging between the elicitation and the strategy choice. They attempt to minimize this by paying for one choice at random, following the guidance of Blanco et al. (2010).

beliefs about complementary events often sum to more than 100%, but again this is true for all three mechanisms. Holt and Smith (2016) compare the dollar-denominated QSR and the single-response BDM to a two-stage MPL in which a coarse MPL with grid $\{0\%, 10\%, 20\%, \ldots, 100\%\}$ is used in the first stage and then a second MPL "zooms in" to elicit the probability at 1% intervals. They find that, compared to the other two, the MPL gave probabilities slightly closer to both an induced prior and the correct Bayesian posterior in an updating task. The MPL also had fewer incidences of 0% or 100% reports, which are never correct in their experiment. Finally, Healy and Kagel (2023) compare the MPL with the BQSR and use chat transcripts between teams of agents to explore whether participants are consciously and intentionally misreporting their beliefs. They find that both mechanisms perform well when eliciting an induced belief, and chat transcripts reveal almost no evidence of intentional misreports in either mechanism. They do find that participants talk more about altering their report in the BQSR but ultimately choose to report truthfully.

Focusing specifically on the BQSR, results are mixed about its absolute performance. When participants are given a simple event with an objective probability, will they report back the correct belief under the BQSR? Erkal et al. (2020) and Danz et al. (2022) find substantial misreporting when the true probability is not 50% and participants are given an on-screen payoff calculator, although Danz et al. (2022) show that the misreports are greatly reduced if the payoff calculator is removed. Healy and Kagel (2023) find low rates of misreporting in the BQSR regardless of the true probability and whether or not a calculator is given.

OUR RECOMMENDATION #5

When Eliciting a Probability, Use a Price List

In terms of incentive compatibility, the MPL and BDM mechanisms are superior to binarized scoring rules. Although it has not been tested extensively, thus far we are not aware of any papers showing that another mechanism yields more accurate beliefs than the MPL. The only downsides to the MPL are that it can take a bit more time and screen space to implement (which is why we recommend not putting it on the decision screen itself) and the incentives are flatter than the binarized quadratic scoring rule. If the latter is a concern, consider the ternary price list described in Section 5.1.2.

5.1.1. Multiple Price Lists

As a summary of our more extensive discussion in Section 2.3, the multiple price list (MPL) is an important variation of the BDM (Becker et al., 1964) mechanism for eliciting beliefs. Although the use of single-response BDM mechanisms dates back to at least Ducharme and Donnell (1973) and Grether (1981), to our knowledge it was not displayed as a choice list until Holt and Smith (2016).

The idea behind the MPL is simple: A participant's belief about an event E is defined as the indifference point between the bet "you win a prize if E occurs" and the lottery "you win a prize with probability p." Thus, our goal is to find that indifference point directly by offering this binary choice for many different values of p and seeing at which point they switch from choosing the bet to choosing the lottery.

If we present the binary choices in increasing order of the probability in the objective lottery, then it is easy to identify the tightest interval that must contain the participant's belief. Let p' be the probability in the highest row where the subjective bet was chosen. Let p'' be the probability in the lowest row where the objective lottery was chosen. The participant's belief p must be in the interval $[p', p'']$. Under *Stochastic Monotonicity over Pure Lotteries* (See Axiom 2 in Section 2.1), there is only one place they will switch from choosing the subjective bet to the objective lottery and this can be used to streamline decisions if desired by asking for this switch-point directly.

EXAMPLE **ELICIT A PROBABILITY**

Multiple Price List

Construct a list of binary choices. Each row contains a subjective bet that pays a prize if E occurs and an objective lottery that pays the same prize with probability p. The value of p increases on each subsequent row. To simplify the procedure, ask only for their switch-point. To elicit more coarsely, reduce the number of rows.

What is your belief about the chance that a randomly selected student passes this exam? To answer this, you will choose either Option A or Option B in each row of the table below. One row will then be chosen at random and you will be paid according to your choice in that row. Clearly you have no reason to lie on any row, because if that row is chosen for payment then you'd end up with the option you like less.

Row #	Option A	Option B
0	☐ \$10 if they Pass	☐ \$10 with 0% chance
1	☐ \$10 if they Pass	☐ \$10 with 1% chance
2	☐ \$10 if they Pass	☐ \$10 with 2% chance
3	☐ \$10 if they Pass	☐ \$10 with 3% chance
⋮	⋮	⋮
98	☐ \$10 if they Pass	☐ \$10 with 98% chance
99	☐ \$10 if they Pass	☐ \$10 with 99% chance
100	☐ \$10 if they Pass	☐ \$10 with 100% chance

Optional: If using single-response.

Presumably you'd choose Option A on the first row (Row 0) and at some point switch to choosing Option B. The row where you'd switch is where you're indifferent; it indicates what you think is the % chance (out of 100) that they will Pass the exam. To save time, simply report your belief below and we'll fill out the list for you, switching from A to B at the % chance you indicate. Again, you get your favorite option on every row if you report your belief truthfully.

I believe the chance they pass the exam is: ______ % (out of 100).

Incentives & Incentive Compatibility.
The MPL is incentive compatible for eliciting an indifference between an objective lottery and subjective bet under *Statewise Monotonicity* (See Axiom 7 in Section 2.3). This indifference can be used to infer beliefs under Definition 2 (See Section 2.1). Statewise monotonicity is a relatively weak assumption and is exactly the same condition required for an experiment with multiple decisions to be incentive compatible (Azrieli et al., 2018, 2020).

Coarse Elicitation.
If you do not need to elicit probabilities precisely, you can simply use an MPL with fewer rows. For example, if $\{10\%, 20\%, \dots, 90\%\}$ are the probabilities used in the price list, then a participant who switches at 70%, for example, reveals that their belief lies in the interval $[60\%, 70\%]$. Note that rows with 0% and 100% can be omitted since, for example, a subject who never switches in this example reveals that they have a belief in the interval $[90\%, 100\%]$. By reducing the number of rows, the incentive for each remaining row is magnified.

An MPL with a single row can be used to ask whether a belief is above or below a given threshold. For example, Dienes and Seth (2010) ask participants only the question on Row 50: \$10 if they Pass versus \$10 with 50% chance. While this obviously elicits less information, it is extremely simple and no longer requires any assumptions for incentive compatibility.

Implementation Details.
For a discussion of additional implementation details when using an MPL—such as whether to allow multiple switch-points—refer now to Section 6.1.

5.1.2. Ternary Price Lists

The ternary price list (TPL) was created by Healy and Leo (2024) as a way to elicit beliefs using an MPL but with fewer questions. The intuition is as follows. Under the assumption that beliefs about an event and its complement sum to one, we can infer beliefs about E by observing beliefs about its complement. Under this assumption, one of these beliefs must be at least 0.5. Thus, we can allow participants to choose whether they would rather be paid for, an event or its complement, and then use an MPL to elicit the belief about that event *but with objective lotteries that start at* 0.5. In this way, half-as-many rows can be used to elicit the same belief.

The TPL has the advantage of giving more powerful marginal incentives, equal to those of the BQSR, while maintaining the MPL's desirable property that it is incentive compatible assuming only statewise monotonicity.[18] The downsides of the TPL are that it is clearly more complex than the MPL, assumes that the beliefs about the event E and its complement sum to 100%, and to date it has not been tested in the laboratory, so we do not yet know how well it performs in practice.[19]

If we present the binary choices in increasing order of the probability in the objective lottery, then it is easy to identify the tightest interval that must contain the participant's belief. Suppose that we observe that the participant prefers to be paid conditional on E over the complement of E. Let p' be the probability in the highest row where the subjective bet was chosen. Let p'' be probability in the lowest row where the objective lottery was chosen. The participant's belief p must be in the interval $[p', p'']$. If the participant prefers to be paid conditional on the complement of E then $1-p \in [p', p'']$ or by complementarity $p \in [1-p'', 1-p']$.

Furthermore, under *stochastic monotonicitiy* (see Section 2.3) there will be a single switch-point and this can be used to streamline decisions if desired by asking for this switch-point directly.

[18]See Axiom 7 in Section 2.3

[19]Scoring rules (binarized or not) also assume beliefs add to 100%. The original MPL and BDM mechanisms do not.

EXAMPLE **ELICIT A PROBABILITY**

Ternary Price List: Single List

What is your belief about the chance that a randomly selected student passes this exam? To answer this, you will choose either Option A, Option B, or Option C in each row of the table below. One row will then be chosen at random and you will be paid according to your choice in that row. Clearly you have no reason to lie on any row, because if that row is chosen for payment then you'd end up with the option you like less.

Row #	Option A	Option B	Option C
50	☐ \$10 if they Pass	☐ \$10 if they Fail	☐ \$10 with 50% chance
51	☐ \$10 if they Pass	☐ \$10 if they Fail	☐ \$10 with 51% chance
52	☐ \$10 if they Pass	☐ \$10 if they Fail	☐ \$10 with 52% chance
⋮	⋮	⋮	⋮
99	☐ \$10 if they Pass	☐ \$10 if they Fail	☐ \$10 with 99% chance
100	☐ \$10 if they Pass	☐ \$10 if they Fail	☐ \$10 with 100% chance

Optional: If using single-response.

Presumably you'd choose either Option A or Option B on the first row (Row 50), depending on whether you think it's more likely that they'd Pass or Fail. Then at some point you'd switch to choosing Option C. The row where you'd switch is where you're indifferent; it indicates what you think is the % chance (out of 100) that they will either Pass or Fail the exam (whichever you thought was more likely). To save time, simply report (1) which you think is more likely (Pass or Fail), and (2) your belief that it will happen. We'll then fill out the list for you, switching from either A or B to Option C at the % chance you indicate. Again, you get your favorite option on every row if you report your belief truthfully.

I believe it's more likely that they: ○Pass ○Fail.

and the chance of that is: ______ % (out of 100).

EXAMPLE **ELICIT A PROBABILITY**

The Ternary Price List for a Probability: Two Binary Lists

Same as above, except now it is presented as a choice between two binary lists.

What is your belief about the chance that a randomly selected student passes this exam? To answer this, you will first tell us whether you think it's more likely that they Pass or Fail the exam. Then, depending on your answer, we'll give you a table of choices where you'll choose between Option A and Option B in each row of the table. One row will then be chosen at random from that list and you will be paid according to your choice in that row. Clearly you have no reason to lie on any row, because if that row is chosen for payment then you'd end up with the option you like less.

If you think Pass is more likely:

Row #	Option A	Option B
50	\$10 if Pass	\$10 w/ 50% chance
51	\$10 if Pass	\$10 w/ 51% chance
52	\$10 if Pass	\$10 w/ 52% chance
⋮	⋮	⋮
98	\$10 if Pass	\$10 w/ 98% chance
99	\$10 if Pass	\$10 w/ 99% chance
100	\$10 if Pass	\$10 w/ 100% chance

If you think Fail is more likely:

Option A	Option B
\$10 if Fail	\$10 w/ 50% chance
\$10 if Fail	\$10 w/ 51% chance
\$10 if Fail	\$10 w/ 52% chance
⋮	⋮
\$10 if Fail	\$10 w/ 98% chance
\$10 if Fail	\$10 w/ 99% chance
\$10 if Fail	\$10 w/ 100% chance

Optional: If using single-response.

For the table you choose, presumably you'd choose Option A on the first row (Row 50) and at some point switch to choosing Option B. The row where you'd switch is where you're indifferent; it indicates what you think is the % chance (out of 100) that they will either Pass or Fail the exam (whichever you thought was more likely). To save time, simply report (1) which you think is more likely (Pass or Fail), and (2) your belief that it will happen. We'll then fill out the appropriate list for you, switching from A to B at the % chance you indicate. Again, you get your favorite option on every row if you report your belief truthfully.

I believe it's more likely that they: ○Pass ○Fail.

and the chance of that is: _____ % (out of 100).

Incentives & Incentive Compatibility.
The TPL is incentive compatible for eliciting an indifference between an objective lottery and subjective bet under *Statewise Monotonicity* (See Axiom 7 in Section 2.3). This indifference can be used to infer beliefs under Definition 2 (See Section 2.1). Statewise monotonicity is a relatively weak assumption and is exactly the same condition required for an experiment with multiple decisions to be incentive compatible (Azrieli et al., 2018, 2020).

The TPL also makes an additional assumption that is not made in the MPL (Section 5.1.1) or the BDM (Section 5.1.3). When a participant prefers to be paid when an event does not happen, their belief about the probability of the event is determined as the complement of their belief about the probability it does not happen. This assumes that a participant's belief about E and its complement must sum to 1.

Coarse Elicitation.
If you do not need to elicit probabilities precisely, you can simply use a TPL with fewer rows. A TPL with a single row can be used to partition the possible beliefs into three ranges. For example, if only the 75% row is given, then choosing Option A reveals that $p \geq 0.75$, choosing Option B reveals that $p \leq 0.25$, and choosing Option C reveals that $p \in [0.25, 0.75]$. While this obviously elicits less information, it is fairly simple and no longer requires any assumptions for incentive compatibility.

Implementation Details.
Implementation details for the TPL are essentially the same as for the MPL, so refer to Section 6.1 to continue.

5.1.3. Single Response BDM Mechanisms

As a summary of our more extensive discussion in Section 2.3, the single response BDM mechanism is based on the value elicitation mechanism of Becker et al. (1964); Marschak (1964). There, a good is presented to the participant and they are asked for the dollar amount that makes them indifferent between the money and the good. Here, the good is a bet on an event, and instead of eliciting indifference in terms of dollars we elicit it in terms of probabilities.

For pedagogical purposes we believe it is useful to imagine the BDM as an MPL but with the list hidden from view.[20] The participant reports their belief p, and then a random number r is drawn from $[0,1]$. Think of r as the randomly chosen row of the "hidden" MPL. If $r \le p$ then on row r the participant prefers to bet on the event ("Option A" in the MPL) rather than receive a lottery that pays with probability r ("Option B"). Thus, they receive the bet. If $r > p$ then they prefer the objective lottery and thus are paid the fixed prize with probability r.

EXAMPLE **ELICIT A PROBABILITY**

Single-Response BDM Mechanism

Ask a participant to provide their belief p about an event or fact. Uniformly choose a probability $r \in [0,1]$. If $r \le p$ pay them a fixed prize $\$x$ if the event occurs; otherwise pay them $\$x$ with probability r.

What is your belief about the probability a randomly chosen student passes the exam? Let's say your belief is p% (out of 100). We will randomly pick a probability between 0% *and* 100%*. If the randomly chosen probability is below your belief p then you will be paid $10 if the student passes the exam. If it is above p then you will be paid $10 with that randomly chosen probability. This method is designed so that, regardless of the randomly chosen probability, you get your most-preferred option if you report your belief truthfully.*

I believe the chance they pass the exam is: ______ % (out of 100).

Incentives & Incentive Compatibility.

The single-response BDM is incentive compatible for eliciting an indifference between an objective lottery and subjective bet under *Statewise Monotonicity* (See Axiom 7 in Section 2.3). This indifference can be used to infer beliefs under Definition 2 (See Section 2.1). Statewise monotonicity is a relatively weak assumption and is exactly the same condition required for an experiment with multiple decisions to be incentive compatible (Azrieli et al., 2018, 2020).

Coarse Elicitation.

The single-response BDM is inherently precise. If you do not need precise information

[20] To be clear, our suggested implementation of the MPL in Section 5.1.1 reduces it to a "single response" mechanism by asking the participant for their switch-point. The difference is that, in the MPL, that single response is then translated into a choice on every row of the table whereas in this section it is not.

about participants' beliefs, we suggest using an MPL (see Section 5.1.1) or ternary list (see Section 5.1.2 with a limited number of rows. These are more natural for eliciting coarsely.

Implementation Details.
Researchers have struggled to find ways to present the single-response BDM mechanism to participants in a way that maximizes understanding. We discuss the many options in Section 6.2.

5.1.4. Scoring Rules

Using scoring rules is by far the oldest and most well-studied problem in this chapter. Because of that and because of how simple probabilistic beliefs are (with a binary outcome), there are many scoring rules and ways of adapting those scoring rules to account for risk preferences. In Section 2.4, we discuss three broad classes of scoring rules: *dollar-denominated*, *risk-adjusted*, and *binarized* and compare the incentives of particular rules. Among all of these options, the binarized quadratic scoring rule (BQSR) is the most commonly used scoring rule for eliciting a belief in experimental economics and the one we present here.

OUR RECOMMENDATION #6

When using a scoring rule, use a binarized scoring rule. We recommend the BQSR.

For the QSR, both risk-adjusting and binarizing improve performance. However, binarized rules are able to account for risk-preferences without additional overhead. Although we recommend using an MPL or BDM for eliciting the probability of an event (see Recommendation 5), if you are going to use a scoring rule, we suggest using a binarized scoring rule. Of these, the binarized quadratic scoring rule [BQSR] has the strongest uniform incentives (See Section 2.4).

EXAMPLE **ELICIT A PROBABILITY**

Binarized Quadratic Scoring Rule (BQSR)

Ask a participant to provide their belief $p \in [0,1]$ about an event or fact. If the event occurs or if the fact is true, give them a lottery that pays a high prize with probability $1-(1-p)^2$ (and a low prize otherwise). If the event does not occur or the fact is false, give them a lottery that pays the high prize with probability $1-(0-p)^2$.

What is your belief p (out of 100) that a randomly chosen student passes this exam? Given your belief p, you will be paid \$10 with probability $1-(1-p)^2$ *if the student passes the exam and* $1-p^2$ *if they do not pass. These formulas make it so that you maximize your overall probability of getting \$10 when you report your belief truthfully.*

I believe the chance they pass the exam is: [   ] % (out of 100).

Incentives & Incentive Compatibility.
Incentive compatibility for the BQSR or any binarized scoring rule relies on *S-O reduction*. This is discussed further in Section 2.4 (See Axiom 8).

Coarse Elicitation.
If desired, beliefs can be elicited coarsely simply by imposing a grid of possible responses. For example, you could require reports to be from the set $\{10\%, 20\%, \ldots, 90\%\}$. Arguably, one advantage of coarse elicitation is that small deviations become impossible, so that the smallest allowable deviation will become relatively more costly. Whether this has an impact on actual behavior—and whether it compensates for the resulting loss in precision—is currently an open question.

If a coarse grid $\{10\%, 20\%, 30\%, 40\%, \ldots, 90\%\}$ is used, then intuitively one would expect that anyone with beliefs in the interval $[25\%, 35\%]$ would report $q = 30\%$. In other words, the person who is exactly indifferent between reporting $q = 30\%$ and $q = 40\%$ is the person whose true belief is $p = 35\%$. We refer to this as the "midpoint property" for coarse elicitation. However, not all scoring rules have this property. In Healy and Leo (2024) we show that the only scoring rules with the midpoint property are those with a quadratic penalty, such as the BQSR and the Reduced MPL/BDM. This gives an additional reason to consider using the BQSR when eliciting coarse beliefs.

Implementation Details.

For additional details on how to implement binarized scoring rules in practice, see Section 6.3.

5.2. Eliciting the Probability of Multiple Events

As with any of the methods discussed in this chapter, it is possible to combine multiple probability elicitations by conducting each separately (using any of the mechanisms discussed in Section 5.1) and then, at the end of the experiment, randomly picking one of these probability reports for payment. Paying one randomly is incentive compatible as long as the participant's preferences respect statewise dominance (an assumption known as *Statewise Monotonicity*).[21] Here, the "state" is the report chosen for payment; see Azrieli et al. (2018, 2020) for more details. Since this assumption is also key for incentive-compatibility of the BDM/MPL mechanisms presented in this chapter, it is discussed in more detail in Section 2.3.

Eliciting multiple probabilities is discussed further in Section 5.9 in the context of eliciting an entire/approximate subjective belief distribution.

5.3. Eliciting the Modal Event

Modes are easy to elicit because they are a purely ordinal concept. Let $\mathscr{E}$ be a set of events. Building on Definition 1 in Section 2.1, E is the modal event from some set of events $\mathscr{E}$ if for all $E' \in \mathscr{E}$, the participant prefers a subjective bet that pays if E over one that pays the same amount conditional on E'. This leads to the following simple elicitation procedure.

[21] See Axiom 7 in Section 2.3

EXAMPLE **ELICIT THE MODAL EVENT**

Menu of Choices

Ask the participant to pick an event E from a set of events $\mathscr{E}$. If E occurs, pay them $\$x$.

We randomly selected someone who took the exam. Which of the following do you think is most likely? If the option you choose is correct, you will be paid \$10 for your response. Otherwise, you will be paid \$0.

Select One	Grade Received
○	They receive a grade of A on the exam.
○	They receive a grade of B on the exam.
○	They receive a grade of C on the exam.
○	They receive a grade of D on the exam.
○	They receive a grade of F on the exam.

Incentives & Incentive Compatibility.
Since this mechanism involves only a single choice without randomization, it elicits a participant's most preferred option without any additional assumptions and thus elicits a participant's belief about the most likely event under belief Definition 1 (see Section 2.1).

Implementation Details.
Normally, we refer the reader to Section 6 for a further discussion of implementation details, but in this case there is nothing further to discuss because the modal event is so simple to elicit.

5.4. Eliciting a Ranking of Events ("Qualitative Probabilities")

Similar to the mode, relative likelihood is a purely ordinal concept. From Definition 1 in Section 2.1, a participant believes that event E is more likely than event E' if they prefer a subjective bet that is conditional on E over a bet that pays the same amount conditional on E'. To elicit the complete ordering of some set of events $\mathscr{E}$, have the participant rank the events, then randomly choose two events and pay the participant the bet they ranked higher. In other words, pay a prize if their higher-ranked event occurs. This incentivizes the truthful ordering of the events for the following reason: If the ordering is incorrect, there is at least one pair of events that is swapped. If that pair is chosen as the random pair, the participant will receive a subjective lottery they prefer less than the lottery they would have received under the correct ordering.[22]

EXAMPLE **ELICIT A RANKING OF EVENTS**

Choice from Pairs

Ask the participant to rank order a set of events $\mathscr{E}$. Randomly select a pair of events E and E' from $\mathscr{E}$. If the participant ranked E above E' pay them a prize if E occurs. If the participant ranked E' above E pay them a prize if E' occurs.

*We randomly chose one student who took the exam. Rank the following in order from what you think is most likely to what you think is least likely (*1 *=most likely,* 5 *=least likely). We will randomly pick two of these options and look at the one you ranked as more likely of the two. If that is the student's actual grade then you will be paid $10. If not, then you will be paid $0. Thus, you maximize your chance of winning $10 if you report your ranking truthfully.*

My Rank (1 to 5)	**Grade Received**
	They receive a grade of A on the exam.
	They receive a grade of B on the exam.
	They receive a grade of C on the exam.
	They receive a grade of D on the exam.
	They receive a grade of F on the exam.

Incentives & Incentive Compatibility.
This method is incentive compatible for eliciting a participant's true rank ordering of the relevant subjective bets under the assumption of *Statewise Monotonicity*.[23] This rank ordering can be used to infer a likelihood ranking through belief Definition 1 (see Section 2.1). Statewise Monotonicity is a relatively weak assumption and is exactly the same condition required for an experiment with multiple decisions to be incentive compatible (Azrieli et al., 2018, 2020). Here, the "state" is the pair of options chosen for comparison.

[22] Dustan et al. (2024) introduce a similar rank-order mechanism for the elicitation of probabilistic beliefs, providing an alternative interface for the incentives of the MPL discussed in Section 5.1.1.
[23] See Axiom 7 in Section 2.3

Intuitively, if a participant submits an ranking besides their true ranking, then at least some pair of options is swapped. If any incorrectly ordered pair of options is randomly chosen as the pair for comparison, the participant will receive the option they prefer less. Since this assumption is also key for incentive-compatibility of the BDM/MPL mechanisms presented in this chapter, it is discussed in more detail in Section 2.3.

Implementation Details.
In the example above, we have shown a numeric rank input where the options are shown in a static list. Another option is to use a dynamic drag-and-drop interface to allow the participants to order the options themselves. Crockett and Oprea (2012) implement such a drag-in-drop list in the context of eliciting rank orderings of bundles and Dustan et al. (2024) implement a similar procedure in the context of eliciting probabilistic beliefs.

5.5. Eliciting the Modal Set, Mode, or Modal Interval.

Modes permit notably simple elicitation procedures, since they depend only on ordinal comparisons between bets on events and can be inferred from a single choice. Because of this, the mechanisms in this section have the weakest incentive-compatibility requirements of any presented in this chapter. See Section 2.1 for a more formal discussion of ordinal and cardinal / probabilistic beliefs.

5.5.1. Modal Set from an Arbitrary Collection

When eliciting beliefs about a random variable with only a small number of outcomes, it is possible to elicit the mode using the method from Section 5.3 on the event space of the random variable. Treat each outcome as a separate event and elicit the mode of those events. If the support is large, it is still possible to use this method by partitioning the support into a small number of sets and treating each set as an event. In fact, this method can be used to elicit the modal set from any arbitrary collection of subsets of the support even if that collection is not a partition.

EXAMPLE **ELICIT THE MODAL SET**

Menu Choice

Ask the participant to pick a set S from a partition of the support X. If S occurs, pay them \$$x$.

We randomly chose one student who took the exam. Which of the following score ranges do you think is most likely? If the range you choose is correct, you will be paid \$10.

Pick One	Score
◦	0 – 19 Points
◦	20 – 39 Points
◦	40 – 59 Points
◦	60 – 79 Points
◦	80 – 100 Points

Incentives & Incentive Compatibility.

Since this mechanism involves only a single choice without randomization, it elicits a participant's most preferred option without any additional assumptions and thus elicits a participant's belief about the most likely event under Definition 1 (see Section 2.1).

Coarse Elicitation.

This method can be made more coarse or precise by choosing finer/coarser partitions to elicit the modal set from. For instance, a very coarse procedure may elicit whether [0 – 49] or [50 – 100] is most likely. This would offer an alternative to eliciting the probability of either of these intervals relative to 0.5.

Implementation Details.

Normally, we refer the reader to Section 6 for a further discussion of implementation details, but in this case there is nothing further to discuss because the modal event is so simple to elicit.

5.5.2. Mode or Modal Interval

When you are interested in the actual mode and the support is large, it is more convenient to simply ask for the most likely value directly rather than using a menu.

EXAMPLE **ELICIT THE MODE**

Pay-if-True

Ask the participant to pick a value x from the support X. If x occurs, pay them $\$y$.

We randomly chose one student who took the exam. What exam score do you think is most likely? If the score you choose is correct, you will be paid $10.

Your Belief: [ ] Points (out of 100)

When eliciting the mode of a random variable with large support, the probability of being paid can be very small. In this case, it may be better to elicit a modal interval. We do this by paying the participant if their "guess" is within some distance d of the true value. This incentivizes the participant to reveal their belief about the center of the modal interval of length $2d$.[24]

EXAMPLE **ELICIT THE MODAL INTERVAL OF FIXED SIZE**

Pay-if-Within-Threshold

To elicit the modal interval of size $2d$, ask the participant to pick a value v from the support X. If $|x - v| \leq d$, pay them $\$y$.

We randomly chose one student who took the exam. What score do you think is most likely to be within 5 *points of a randomly chosen student's actual exam score? If the score you choose is within* 5 *points (above or below) of their actual score, you will be paid $10.*

Your Belief: [ ] Points (out of 100)

Incentives & Incentive Compatibility.
Since these mechanisms involve only a single choice without randomization, they elicit a participant's most preferred option without any additional assumptions and thus elicit a participant's belief about the most likely value/interval under Definition 1 (see Section 2.1).

[24]We note that, for a random variable with support $[a,b]$, reports in the range $[a,a+d)$ are dominated (weakly) by the report of $a+d$. Similarly, the range $(b-d,b]$ is dominated (weakly) by $b-d$.

Implementation Details.
Normally, we refer the reader to Section 6 for a further discussion of implementation details, but in this case there is nothing further to discuss because the mode (or modal interval) is so simple to elicit.

5.6. Eliciting the Median and other Quantiles

MENU

Eliciting the Median or other Quantiles

- Multiple Price Lists (See Section 5.6.1)
- Single Response BDM Mechanisms (See Section 5.6.2)
- (Binarized) Scoring Rules (See Section 5.6.3)

Eliciting quantiles are a useful way to pinpoint "parts" of a distribution relevant to a particular research question. They can be used to uncover information about the shape of a distribution that might be hidden by eliciting a mean or mode. Compared to the mean, they are also simpler to elicit since they are event-based rather than being a summary of an entire distribution. The elicitation of quantiles also forms the foundation of methods for confidence interval elicitation presented in Sections 5.7 as well as for CDF approximation presented in Section 5.9.2.

5.6.1. Multiple Price Lists

When eliciting a probability with a price list, we fix an event E and elicit the probability of that event by comparing a bet that pays conditional on E to lotteries that pay conditional on objective probabilities. This allows us to identify which objective lottery is indifferent to the bet that pays if E to determine the belief about p through Definition 2 (See Section 2.1). In this case, the E is fixed and we look for a probability that makes the participant indifferent.

For continuous distributions, a quantile is a value q such that $F(q) = p$. Finding this q is equivalent to looking for which event E of the form $X \leq q$ has probability p. Thus, eliciting a quantile is the inverse of eliciting a probability: E is unknown and p is fixed. Because of this, quantiles can be elicited with a sort of inverse probability price list where the objective lottery is fixed and the subjective lotteries change in each row. This elicitation was introduced by Leo and Stelnicki (2024).

To construct this list, we create a sequence of events $X \leq x$ and ask the participant to compare being paid a subjective bet conditional on each of these to being paid a fixed objective lottery that pays with p (the quantile being elicited). If we present these choices in increasing order of x, then it is easy to identify the tightest interval that must contain the participant's quantile belief q. Let x' be the value of x in the highest row where the objective lottery was chosen. Let x'' be the value of x in the lowest row where the subjective bet was chosen. The participant's belief $q \in [x', x'']$. Under *Statewise Monotonicity over Pure Bets* (See Axiom 3 in Section 2.1) there is only one place they will switch from choosing the objective lottery to the subjective bet and this can be used to streamline decisions if desired by asking for this switch-point directly.

EXAMPLE **ELICIT A QUANTILE**

Quantile MPL

Construct a list of binary choices. Each row contains a subjective bet that pays a prize if $X \leq x$ (for various x) and an objective lottery that pays with a probability p (the quantile being elicited). The value of x increases (or decreases) in each subsequent row. Randomly choose a pair and pay the participant their chosen option from the pair. To simplify the procedure, ask only for their switch-point.

What is your belief about the score such that there is a 50% chance that a randomly selected student will achieve a score at or below that value? This is your belief about the median score. To answer this, you will choose either Option A, Option B in each row of the table below. One row will then be chosen at random and you will be paid according to your choice in that row. Clearly you have no reason to lie on any row, because if that row is chosen for payment then you'd end up with the option you like less.

Row #	Option A	Option B
0	☐ \$10 with 50% chance	☐ \$10 if Score ≤ 0
1	☐ \$10 with 50% chance	☐ \$10 if Score ≤ 1
2	☐ \$10 with 50% chance	☐ \$10 if Score ≤ 2
⋮	⋮	⋮
98	☐ \$10 with 50% chance	☐ \$10 if Score ≤ 98
99	☐ \$10 with 50% chance	☐ \$10 if Score ≤ 99
100	☐ \$10 with 50% chance	☐ \$10 if Score ≤ 100

Optional: If using single-response.

Presumably you'd choose Option A on the first row (Row 0) and at some point switch to choosing Option B. The row where you'd switch is where you're indifferent; it indicates your belief about the median score. To save time, simply report your belief below and we'll fill out the list for you, switching from A to B at the row chance you indicate. Again, you get your favorite option on every row if you report your belief truthfully.

I believe the median score is: ☐ points (out of 100).

Incentives & Incentive Compatibility. Like the MPL for eliciting a probability discussed in Section 5.1.1, this MPL is incentive compatible for eliciting an indifference between an objective lottery and subjective bet under *Statewise Monotonicity* (See Axiom 7 in Section 2.3). This indifference can be used to infer a quantile belief under Definition 2 (See Section 2.1). Statewise monotonicity is a relatively weak assumption and is exactly the same condition required for an experiment with multiple decisions to be incentive compatible (Azrieli et al., 2018, 2020). See also the discussion of incentive compatibility in Leo and Stelnicki (2024).

Coarse Elicitation.
Like the MPL for eliciting a probability (See Section 5.1.1), this procedure's precision can be adjusted by reducing/increasing the number of rows. Since the switch-point identifies an interval that must contain the desired quantile of the participant's belief, decreasing the number of rows and increasing the gap between the values in "Option B" of the list increases the size of the ranges that are identified as containing the quantile.

Implementation Details.
For a discussion of additional implementation details when using an MPL—such as whether to allow multiple switch-points—refer now to Section 6.1.

5.6.2. *Single Response BDM Mechanisms*

For a brief history of BDM mechanisms for eliciting probabilistic beliefs, refer to Section 2.3. We are unaware of any literature discussing the use of single-response BDM mechanisms for eliciting quantiles; however, the methodology follows naturally from the MPL mechanism for quantiles proposed by Leo and Stelnicki (2024).

As mentioned in Section 5.1.3, we believe it is useful, for pedagogical purposes, to imagine the BDM as an MPL but with the list hidden from view.[25] The participant reports their belief q in $[a,b]$, and then a random number r is drawn from $[a,b]$. Think of r as the randomly chosen row of the "hidden" MPL. If $r \leq q$ then on row r the participant prefers the objective lottery ("Option A" in the MPL) rather than receive a bet that pays if $x \leq r$ ("Option B"). Thus, they receive the lottery. If $r > q$ then they prefer the bet that $x \leq r$ and thus are paid if this event is true.

[25]To be clear, our suggested implementation of the MPL in Section 5.6.1 reduces it to a "single response" mechanism by asking the participant for their switch-point. The difference is that, in the MPL, that single response is then translated into a choice on every row of the table whereas in this section it is not.

EXAMPLE **ELICIT A QUANTILE**

Single-Response BDM Mechanism

Ask a participant to provide their belief q about the value of quantity X such that there will be a probability p (the quantile) that $X \leq q$. Limit the reports to $[a,b]$. Uniformly choose a value $r \in [a,b]$. If $r \leq q$ pay them \$$x$ with probability p. Otherwise, pay them \$$x$ if $x \leq r$.

What is your belief about the score such that there is a 50% *chance that a randomly selected student will achieve a score at or below that value? This is your belief about the median score. Let's say your belief is q (out of 100). We will randomly pick a value between* 0 *and* 100*. If the randomly chosen value is larger than your belief q then you will be paid \$10 if the student's score is below that randomly chosen value. If it is below q then you will be paid \$10 with a* 50% *chance. This method is designed so that, regardless of the randomly chosen value, you get your most preferred option if you report your belief about the median truthfully.*

I believe the median score is: [ ] points (out of 100).

Incentives & Incentive Compatibility.
Like the single-response BDM for eliciting a probability discussed in Section 5.1.3, this MPL is incentive compatible for eliciting an indifference between an objective lottery and subjective bet under *Statewise Monotonicity* (See Axiom 7 in Section 2.3). This indifference can be used to infer a quantile belief under Definition 2 (See Section 2.1). Statewise monotonicity is a relatively weak assumption and is exactly the same condition required for an experiment with multiple decisions to be incentive compatible (Azrieli et al., 2018, 2020).

Coarse Elicitation.
The single-response BDM is inherently precise. If you do not need precise information about participants' beliefs, we suggest using an MPL (see Section 5.6.1) with a limited number of rows which are more natural for eliciting coarsely.

Implementation Details.
Researchers have struggled to find ways to present the single-response BDM mechanism

to participants in a way that maximizes understanding. We discuss the many options in Section 6.2.

5.6.3. (Binarized) Scoring Rules

Although less studied than the scoring rules for probabilities discussed in Section 5.1.4, quantiles can also be elicited with scoring rules. Here, we will focus on binarized scoring rules rather than dollar-denominated rules. See Section 2.4 for a discussion of binarized and dollar-denominated rules and additional theory of scoring rules.

The goal of eliciting a quantile of probability with a scoring rule is to design a score $S_x(q)$ that maps a participant's report q and the actual realization x into a probability of being paid a fixed prize. The rule is said to be strictly proper for eliciting quantile p if a participant who wishes to maximize the expected value of the score $S_x(q)$ does so by reporting a / the quantile p of their belief distribution truthfully.

Theorem 6 of Gneiting (2011) provides a sufficient condition for a rule to be strictly proper for a quantile. Using our notation, a rule is strictly proper for eliciting the p quantile if it has the form $S_x(q) = ps(q) + (s(x) - s(q))\mathbb{1}\{x \leq q\} + h(x)$ where s is nondecreasing and h is arbitrary.[26] This is a broad class of rules. However, existing literature focuses mainly on the linear rules established by Cervera and Munoz (1996). For example, Eyting and Schmidt (2021); Schlag and Van der Weele (2013); Hossain and Okui (2013) discuss the binarization of these linear rules.

To generate a binarized linear scoring rule, we can take $s(x) = x$ and $h(x) = 1 - px$ in the sufficient condition above to get the scoring rule:

$$S_x(q) = \begin{cases} 1 - p(x-q) & if\, q < x \\ 1 - (1-p)(q-x) & if\, q \geq x \end{cases}$$

. In general, this is not a binarized rule since $S_x(q)$ may take values outside of $[0,1]$, and thus it cannot be interpreted as the probability of being paid a fixed prize. However, a normalization (shown in the box below) ensures that the image of this function (for $x, q \in [\underline{Q}, \overline{Q}]$) is the entire unit interval $[0,1]$. This normalization depends on whether $p \geq 0.5$.

Notice that under this rule, the probability of winning the prize decreases linearly in the distance of the realization from the elicited value (contrast this with the quadratic penalty used for eliciting a mean in Section 5.8.3). The slope of this linear penalty

[26]See also Schervish et al. (2012) for a related characterization.

depends on the quantile being elicited and whether $x > q$ or $x < q$. For quantiles above 0.5, the penalty is higher for errors where $x > q$ than for where $x < q$. The opposite is true when eliciting quantiles below 0.5.

For some intuition on this, note that a random variable is just as likely to take a value above the median as it is below the median. However, a random variable is ten times more likely to take a value below its 0.9 quantile than above. Thus, to ensure that participants have an incentive to properly locate the 0.9 quantile of their beliefs, we penalize errors above their stated quantile q ten times more than errors below. Eyting and Schmidt (2021) utilize this in the presentation for the rule to participants. For example, in eliciting the 0.75 quantile, they ask: “What do you say is y if underestimation is three times less costly than overestimation?”.

EXAMPLE **ELICIT A QUANTILE**

Binarized Linear Scoring Rule

Choose a minimum value $\underline{Q}$ and a maximum value $\overline{Q}$ that the participants can submit. Ask a participant to provide their belief q about the p-quantile of the value (0.5 for the median). Compare q to the actual outcome x and pay the participant a prize with probability $s_x(q)$ defined below.

$$\text{For } p \geq 0.5\text{: } s_x(q) = \begin{cases} 1 - \dfrac{q-x}{\overline{Q}-\underline{Q}}, & \text{if } q > x \\ 1 - \dfrac{1-p}{p}\dfrac{x-q}{\overline{Q}-\underline{Q}}, & \text{if } q < x \end{cases}$$

$$\text{For } p < 0.5\text{: } s_x(q) = \begin{cases} 1 - \dfrac{p}{1-p}\dfrac{q-x}{\overline{Q}-\underline{Q}}, & \text{if } q > x \\ 1 - \dfrac{x-q}{\overline{Q}-\underline{Q}}, & \text{if } q < x \end{cases}$$

We will select a random student's exam. What do you believe is the score such that there is a 50% chance that the student has earned more than that score and 50% chance that they have earned less than that score. This is your belief about the median score. To incentivize you to truthfully report your belief about the median score, we will pay you in the following way based on your guess q. Suppose that the randomly chosen student's score is x. If x is below q or equal to q, you will be paid \$10 *with probability* $1-(q-x)$. *If x is above q, you will be paid* \$10 *with probability* $1-(x-q)$. *This procedure has been carefully designed so that you maximize your chance of being paid the \$10 when you report your belief about the median truthfully.*

I believe the median score is: [ ] points (out of 100).

Incentives & Incentive Compatibility.

Like for scoring rules eliciting a probability discussed in Section 5.1.4, incentive compatibility of the binarized linear scoring rule requires a form of reduction of compound lotteries called *Subjective-Objective Reduction* (See Section 2.4 Axiom 8 for further discussion of this assumption). Unlike probability elicitation, this reduction must be applied across possibly infinite values in the support of X. Thus, the resulting compound lotteries are much more complex than those arising from scoring rules for probabilities. Nevertheless, the structure is the same and subjective-objective reduction remains the key incentive-compatibility requirement.

Coarse Elicitation.

We are not aware of any papers studying coarse elicitation of quantiles using scoring rules in general. However, the linear scoring rule is not suitable for coarse belief elicitation in the sense that, when possible reports are limited, participant might not have incentive to pick the report that is closest to their true belief.[27] For this reason, we suggest using the MPL presented in Section 5.6.1 when you need less precise information about quantile beliefs.

Implementation Details.

For additional details on how to implement binarized scoring rules in practice, see Section 6.3. Quantile elicitation has received far less attention than probability elicitation, and although Section 6.3 is more focused on probabilistic scoring rules, many of the details will apply to scoring rules generically.

One important detail specific to eliciting quantiles directly is how we explain what we are eliciting to participants. Although participants are generally familiar with probabilities and we can ask "what do you think the probability is that...?" it is harder to succinctly communicate a quantile. In the example above, we have offered some example language, but we do not have nearly enough evidence to build a "best practice" for this.

Explaining explicitly what you are eliciting is most important when you are ensuring participants that it is in their best interest to report that thing truthfully. Eyting and Schmidt (2021) opt not to mention what is being elicited, instead relying on the incentives themselves to direct participants to the correct quantity. Dustan et al. (2022), on the other hand, mention that they are explicitly eliciting the median and attempt to explain the median with language similar to that above. One benefit of using an MPL

[27]For instance, suppose we elicit the median about a quantity with natural support $[0,1]$ from a participant with a subjective belief distribution that has density $f(x) = 3x^2$. The participant's expected probability of winning the prize is higher when reporting 0.55 over 1 despite the fact that 0.55 is further from the median of $\frac{1}{\sqrt[3]{2}}$.

for eliciting a quantile is that it is not as necessary to explain that a quantile is the target, this is because the quantile is inferred through the preferences rather than being elicited directly.

5.7. Eliciting a Confidence Interval

Any continuous distribution contains many intervals that occur with probability p. Thus, our first step in eliciting a confidence interval is deciding which one to elicit. The most obvious choice is an interval centered on some number. We will refer to these as *centered confidence intervals*.

A *centered* confidence interval of probability p with center c is an interval $[c-d, c+d]$ such that a participant believes there is a p chance that the true value of x falls within the interval. The goal here is to elicit the d.

If there is not a clear "center" that you want to elicit an interval around, you can instead elicit what we call a *balanced confidence interval*. Formally, a balanced confidence interval of probability p is a pair of numbers l and h such that $1-F(h) = \frac{p}{2}$ and $F(l) = \frac{p}{2}$. That is, there is equal probability of being "above" and "below" the interval.

By definition, the balanced confidence interval is a pair of quantiles: the $\frac{p}{2}$ and $1-\frac{p}{2}$ quantiles. Thus, it can be determined by eliciting this pair of quantiles. These quantiles can be measured by two separate elicitations using any of the methods appropriate for eliciting quantiles discussed in Section 5.6, with one elicitation being chosen at random for payment. Gneiting and Raftery (2007) discuss the specific case of eliciting a balanced interval with a scoring rule and propose a single scoring rule which is the reduced form of using two quantile scoring rules and choosing one at random for payment.

The centered confidence interval can also be reduced to a quantile elicitation. Notice that for any interval centered at c and for any number d, $P(c-d \le x \le c+d) = P(|x-c| \le d)$. Thus, if $[c-d, c+d]$ is a p confidence interval for x, then d is a p quantile of $|x-c|$. Thus, any centered confidence interval can be elicited with a single quantile elicitation of the transformed variable $|X-c|$.

EXAMPLE	**ELICIT A CENTERED CONFIDENCE INTERVAL**

Elicit a Quantile $|X - c|$

Choose a mechanism from Section 5.6. To elicit a confidence interval of probability p centered at c for random variable X, elicit the p quantile of $|X-c|$. That is, a quantile of the participants belief about the distance between x and c.

EXAMPLE **ELICIT A BALANCED CONFIDENCE INTERVAL**

Elicit a Pair of Quantiles

Choose a mechanism from Section 5.6. To elicit the centered confidence interval of probability p, use the mechanism to elicit the $\frac{p}{2}$ and $1-\frac{p}{2}$ quantile. Then randomly choose one of these elicitations for payment.

Incentives & Incentive Compatibility.
Refer to the relevant incentive compatibility conditions for your chosen quantile elicitation in Section 5.6. Eliciting a balanced confidence requires eliciting two quantiles and then paying for one randomly. This whole procedure is incentive compatible as long as the individual elicitations are incentive compatible and preferences obey *Statewise Monotonicity* (See Axiom 7 in Section 2.3). Here, the "state" here is which report is chosen for payment. However, since statewise monotonicity is already needed for the incentive compatibility of the MPL or single-response BDM, and since it is weaker than S-O reduction (See Axiom 8 in Section 2.4) which is needed for the incentive compatibility of the binarized linear scoring rule, no additional assumptions are needed to elicit a balanced confidence interval on top of those needed by the chosen mechanism.

Implementation Details.
Refer to the relevant implementation details for your chosen quantile elicitation in Section 5.6.

5.8. Eliciting the Mean

MENU

Eliciting the Mean

- Multiple Price Lists (See Section 5.8.1)
- Single Response BDM Mechanisms (See Section 5.8.2)
- (Binarized) Scoring Rules (See Section 5.8.3)

Since the mean, in particular, is commonly used for analysis in statistics, moments can be a tempting target for belief elicitation. However, because moments are summaries of an entire distribution, they can be difficult to elicit.[28] For this reason, in Recommendation 2 we suggest avoiding eliciting means when possible.

However, in some instances, the mean may be most appropriate statistic to elicit. Consider a hiring experiment where a "manager" can hire from two groups of test-takers. The manager's payment is proportional to the score of a randomly chosen person from the chosen group. For a risk-neutral manager, the mean of each group is the belief most relevant to their hiring decision. Eliciting another statistic can be misleading. For instance, suppose each group consists of three people and group 1 scores are 60, 60, 10 while group 2 scores are 20, 20, 70. The mean score of group 1 is higher than the mean score of group 2 even though there is a $\frac{5}{9}$ chance that a randomly chosen person from group 2 outscores a randomly chosen person from group 1.

5.8.1. Multiple Price Lists

MPLs are designed to elicit an indifference. The beliefs that define probabilities and quantiles lead to simple indifferences of the form presented in Definition 2 of Section 2.1 "What is a Belief?" However, constructing a suitable indifference that defines the belief about a mean is a little more difficult.

To start, note that a risk-neutral expected utility maximizer would be indifferent between receiving the value of x and the mean of $\mu = E(X)$. However, we cannot rely on risk neutrality for elicitation. However, we can extend this indifference to be appropriate for any expected utility maximizer by transforming X to always be between 0 and 1 and then defining the indifference over lotteries that pay some prize with a probability that depends on the realization of this transformed variable.

Suppose X has support $[a,b]$. Consider the random variable $\tilde{X} = \frac{X-a}{b-a}$ (the distance x is on the line between a and b). This random variable always takes a value between 0 and 1. Any expected utility maximizer is indifferent between a lottery that pays with probability $\frac{X-a}{b-a}$ and one that pays with their belief about the mean of this variable $\mu_{\tilde{X}}$. Thus, a mean can be identified by looking for this indifference.

[28]It is possible to elicit other moments as well. For example, Gneiting and Raftery (2007) propose a scoring rule for eliciting a variance. However, it requires two independent samples of the random variable in question. In this sense, it is not suitable for eliciting variance of beliefs about quantities that are not inherently random. If other moments are important to your research, we suggest approximating the participant's entire belief distribution and calculating approximate moments. See Section 5.9.

As with the MPL for eliciting a probability, we ask the participant to compare a fixed (complex) subjective bet that pays with a probability equal to the realized value of the transformed variable and a sequence of objective lotteries. If we present the binary choices in increasing order of the probability in the objective lottery, then it is easy to identify the tightest interval that must contain the participant's mean belief about the transformed variable. Let p' be the probability in the highest row where the subjective bet was chosen. Let p'' be the probability in the lowest row where the objective lottery was chosen. The participant's belief $\mu_{\tilde{X}}$ must be in the interval $\in [p', p'']$. This can then be de-normalized to the the participant's belief about the mean of the original random variable X. *Stochastic Monotonicity over Pure Lotteries* (See Axiom 2 in Section 2.1), there is only one place they will switch from choosing the subjective bet to the objective lottery and this can be used to streamline decisions if desired by asking for this switch-point directly.

EXAMPLE **ELICIT A MEAN**

Multiple Price List

Construct a list of binary choices. Each row contains a subjective bet that pays a prize with probability euqal to the realization of $\frac{X-a}{b-a}$ and an objective lottery that pays the same prize with probability p. The value of p increases (or decreases) on each subsequent row. Randomly choose a pair and pay the participant their chosen option from the pair. To simplify the procedure, ask only for their switch-point.

What is your belief about the average score (as a percentage) of a randomly selected student? To answer this, you will choose either Option A or Option B in each row of the table below. One row will then be chosen at random and you will be paid according to your choice in that row. Clearly you have no reason to lie on any row, because if that row is chosen for payment then you'd end up with the option you like less.

Row #	Option A	Option B
0	☐ \$10 with probability Score%	☐ \$10 with 0% chance
1	☐ \$10 with probability Score%	☐ \$10 with 1% chance
2	☐ \$10 with probability Score%	☐ \$10 with 2% chance
3	☐ \$10 with probability Score%	☐ \$10 with 3% chance
⋮	⋮	⋮
98	☐ \$10 with probability Score%	☐ \$10 with 98% chance
99	☐ \$10 with probability Score%	☐ \$10 with 99% chance
100	☐ \$10 with probability Score%	☐ \$10 with 100% chance

Optional: If using single-response.
Presumably you'd choose Option A on the first row (Row 0) and at some point switch to choosing Option B. The row where you'd switch is where you're indifferent; it indicates what you think is the % chance (out of 100) that they will Pass the exam. To save time, simply report your belief below and we'll fill out the list for you, switching from A to B at the % chance you indicate. Again, you get your favorite option on every row if you report your belief truthfully.

I believe the mean score is: ☐ points (out of 100).

Incentives & Incentive Compatibility.
The incentive requirements for eliciting a mean using an MPL are more stringent than those needed when using MPLs for eliciting probabilities discussed in Section 5.1.1 or the MPLs for eliciting quantiles discussed in 5.6.1. Like those other MPLs, since this procedure uses multiple binary choices, the entire procedure requires *Statewise Monotonicity* to be incentive compatible *to truthfully elicit the participants' preferences in each row*.[29] These preferences can then be used to identify an indifference point. However, notice that in the introduction to this section that indifference point that identifies a mean belief is constructed under the assumption that the participant is an expected utility maximizer. Thus, unlike for other MPLs, this procedure requires reduction of compound lotteries to be incentive compatible for eliciting a mean belief. Because of this, using an MPL to elicit a mean belief does not come with the usual benefit of weak incentive requirements.

Coarse Elicitation.
Like the MPL for eliciting a probability (See Section 5.1.1), this procedure's precision can be adjusted by reducing/increasing the number of rows. Since the switch-point identifies an interval that must contain the mean of the participant's belief distribution, decreasing the number of rows and increasing the gap between the values in "Option B" of the list increases the size of the ranges that contain the mean.

Implementation Details.
For a discussion of additional implementation details when using an MPL—such as whether to allow multiple switch-points—refer now to Section 6.1.

5.8.2. Single Response BDM Mechanisms

For a brief history of BDM mechanisms for eliciting probabilistic beliefs, refer to Section 5.1.3. We are unaware of any literature discussing the use of single-response BDM mechanisms for eliciting means more generally; however, the methodology follows naturally from the MPL mechanism discussed in the previous subsection.

As mentioned in Section 5.1.3, we believe it is useful, for pedagogical purposes, to imagine the BDM as an MPL but with the list hidden from view.[30] As in the previous

[29] See Axiom 7 in Section 2.3 for a formal definition of statewise monotonicity.

[30] To be clear, our suggested implementation of the MPL in Section 5.6.1 reduces it to a "single response" mechanism by asking the participant for their switch-point. The difference is that, in the MPL, that single response is then translated into a choice on every row of the table whereas in this section it is not.

subsection, we begin by transforming the variable in question so that it has natural support $[0,1]$. For X with natural support $[a,b]$, let $\tilde{X} = \frac{X-a}{b-a}$. Instead of eliciting beliefs about the mean of X, we elicit the belief about $\tilde{X}$.

The participant reports their belief μ in $[0,1]$, and then a random number r is drawn from $[0,1]$. Think of r as the randomly chosen row of the "hidden" MPL. If $r \leq \mu$ then on row r the participant prefers the lottery that pays the prize with a probability equal to the transformed value of the variable in question ("Option A" in the MPL) rather than receive the objective lottery that pays with probability r ("Option B"). Thus, they receive the lottery that pays with a probability based on the realization of the random variable. If $r > p$ then they prefer the objective lottery that pays with probability r and are paid that lottery.

EXAMPLE	**ELICIT A MEAN**

Single-Response BDM Mechanism

Transform the random variable in question to have support $[0,1]$, for X with natural support $[a,b]$, elicit beliefs about the transformed variable $\tilde{X} = \frac{X-a}{b-a}$. Ask a participant to provide their mean belief μ about the value $\tilde{X}$. Uniformly choose a value $r \in [0,1]$. If $\mu \leq r$ pay them a lottery that pays some amount of money with probability equal to the realization of the transformed variable $\tilde{X}$. Otherwise, pay them the same amount with probability r%.

What is your belief about the average score of randomly selected student? Let's say your belief is m (out of 100). We will randomly pick a number between 0 *and* 100. *Call this r. If the randomly chosen value r is larger than your belief m then you will be paid \$10 with probability r otherwise, you will be paid \$10 with probability x% where x is the score of the randomly chosen student. This method is designed so that, regardless of the randomly chosen value, you get your most preferred option if you report your belief truthfully.*

I believe the mean score is: ______ points (out of 100).

Incentives & Incentive Compatibility.

The incentive requirements for eliciting a mean using a single-response BDM are more stringent than those needed when using a signle-response BDMs for eliciting probabilities discussed in Section 5.1.3 or quantiles discussed in 5.6.2. Like those other single-response BDMs, since this procedure induces multiple binary choices, the entire procedure requires *Statewise Monotonicity* to be incentive compatible *to truthfully elicit the participants' preferences in each row*.[31] These preferences can then be used to identify an indifference point. However, notice that in the introduction to this section that indifference point that identifies a mean belief is constructed under the assumption that the participant is an expected utility maximizer. Thus, unlike for other BDM mechanisms, this procedure requires reduction of compound lotteries to be incentive compatible for eliciting a mean belief. Because of this, using a BDM to elicit a mean belief does not come with the usual benefit of weak incentive requirements.

Coarse Elicitation.
The single-response BDM is inherently precise. If you do not need precise information about participants' beliefs, we suggest using an MPL (see Section 5.8.1) with a limited number of rows which are more natural for eliciting coarsely.

Implementation Details.
Researchers have struggled to find ways to present the single-response BDM mechanism to participants in a way that maximizes understanding. We discuss the many options in Section 6.2.

5.8.3. (Binarized) Scoring Rules

An uncertain event can be understood to have a Bernoulli distribution taking value 1 if the event is true and 0 if it is false. The probability of that event is the mean of this distribution. In this sense, the problem of eliciting a probability is a special case of eliciting a mean.

Like for eliciting a probability (See Section 5.1.4), eliciting the mean of a distribution can also be accomplished with a quadratic scoring rule. Here, the elicited value μ is compared to the realization x and participants are paid with a lottery that pays a prize with a probability decreasing quadratically in error $\mu - x$. This is identical to the case of eliciting probabilities with a BQSR, except here the realization can be numbers other than 0 and 1.

[31]See Axiom 7 in Section 2.3 for a formal definition of statewise monotonicity.

EXAMPLE **ELICIT A MEAN**

Binarized Mean (Quadratic) Scoring Rule

For a quantity with natural support $[a,b]$, ask the participant to provide their mean belief μ. Compare μ to the actual outcome x. Pay the participant a prize with probability $1-\left(\frac{\mu-X}{b-a}\right)^2$.

What is your mean μ of your belief about a randomly chosen student's score on the exam? We will choose a random student and look at their score x. You will be paid $10 with probability $1-\left(\frac{\mu-x}{100}\right)^2$.

Your Belief:

Incentives & Incentive Compatibility.

Like for scoring rules for eliciting a probability discussed in Section 5.1.4, incentive compatibility of the binarized quadratic scoring rule for eliciting a mean requires a form of reduction of compound lotteries called *Subjective-Objective Reduction* (See Section 2.4, Axiom 8 for further discussion of this assumption). Unlike probability elicitation, this reduction must be applied across the possibly infinite values in the support of X. Thus, the resulting compound lotteries are far more complex that those arising from scoring rules for probabilities. Nevertheless, the structure is the same, and subjective-objective reduction remains the key incentive-compatibility requirement.

Coarse Elicitation.

Like the quadratic scoring rule for eliciting beliefs discussed in Section 5.1.4, the quadratic rule for a mean also has the "midpoint property" for coarse elicitation (see Proposition 6 in Appendix Section A.1). That is, when given a limited set of possible reports, the participant has incentive to pick the value that is closest to their true mean belief. Given a grid of possible reports, one can easily infer an interval that contains a participants true belief using this midpoint property. This may improve incentives by forcing participants to consider "large" changes in their reported mean over the marginal changes possible when eliciting precisely.

Implementation Details.

For additional details on how to implement binarized scoring rules in practice, see Section 6.3. Although Section 6.3 is more focused on probabilistic scoring rules, many of the details will apply to scoring rules generically.

5.9. Eliciting an Entire Distribution

MENU

Eliciting the Median or other Quantiles

- Eliciting the PDF/PMF (See Section 5.9.1)
- Eliciting the CDF (See Section 5.9.2)

A distribution can be elicited either through its PDF or CDF. Eliciting the PDF or PMF requires eliciting many probabilities, while eliciting the CDF requires eliciting either many quantiles or probabilities.

5.9.1. Eliciting the PDF/PMF

Eliciting an entire distribution from the PDF or PMF can be achieved by eliciting the probability of many events. For example, if the possible outcomes are finite, the distribution is characterized by the probability of each of those outcomes. In this case, the entire PMF can be elicited by applying one of the mechanisms in Section 5.1 to each of the outcomes, then randomly selecting an outcome and implementing the incentives associated with the participant's report for that outcome.

If eliciting the probability of all the outcomes is infeasible because there are too many outcomes (or an infinite number of outcomes), the support of the distribution can be "binned" into sets. Then, you can elicit the probability of each of these sets. This provides an approximate PMF/PDF.

EXAMPLE **ELICIT A (APPROXIMATE) PMF/PDF**

Elicit Many Probabilities

To elicit the PDF/PMF of a random variable X, Elicit the probability of every value in the support or some partition of these values. This can be achieved using any of the mechanisms presented in Section 5.1. Randomly pick a value/set of values and implement the incentives associated with the mechanism used to elicit that probability.

We randomly chose one student who took the exam. For each of the ten ranges below, please provide your belief about the probability that the randomly chosen student's score falls in that range. We will randomly pick one range and pay you based on your belief for that range. Your payment will be determined based on the procedure discussed earlier in the instructions. However, remember that you maximize your chance of winning $10 by reporting your belief truthfully.

Probability	**Score Range**
________	0-9
________	10-19
________	20-29
________	30-39
________	40-49
________	50-59
________	60-69
________	70-79
________	80-89
________	90-100

Incentives & Incentive Compatibility.
Since this procedure uses multiple probability elicitations and pays one randomly, this whole procedure is incentive compatible as long as the individual elicitations are incentive compatible and preferences obey *Statewise Monotonicity* (See Axiom 7 in Section 2.3). Here, the "state" here is which report is chosen for payment. However, since statewise monotonicity is already needed for the incentive compatibility of the MPL or single-response BDM, and since it is weaker than S-O reduction (See Axiom 8 in Section 2.4) which is needed for the incentive compatibility of the binarized scoring rule, no

additional assumptions are needed to elicit a PDF/PMF on top of those needed by the chosen probability elicitation mechanism.

Implementation Details.
If you are willing to assume that probabilities sum to 1, it is possible to omit the elicitation of a single value / set. Depending on the number of elicitations, this may improve incentives marginally by increasing the probability any of the other elicitations are chosen for payment.

5.9.2. Eliciting the CDF

Another approach to elicit an entire (or an approximation of an entire) distribution is to elicit points on the CDF. This can be done by eliciting several probabilities of events of the form $X \leq q$ according to the procedures in Section 5.1 or by eliciting several quantiles according to the procedures in Section 5.6. Specifically, the participant reports a cdf G (or an approximation of G restricted to some grid). The mechanism then randomly picks one q from the support of the random variable. Since $G(q)$ is the reported probability that $X \leq q$, we can apply any of the mechanisms for a single probability (Section 5.1) on $G(q)$. Qu (2012) provides an example that uses the single-response BDM.

When the participant is only asked to report values of the CDF $G(q)$ on a finite grid then the actual CDF can be approximated via interpolation. For example, Leo and Stelnicki (2024) provide a procedure for fitting a complete CDF through these points using the principle of Maximum Entropy.

EXAMPLE **ELICIT A (APPROXIMATE) CDF**

Elicit Many Probabilities

To elicit the (approximate) CDF of a random variable X, elicit the probability of X being below every value in the support or some subset of the values in the support. This can be achieved using any of the mechanisms presented in Section 5. Randomly pick a value/set of values and implement the incentives associated with the mechanism used to elicit that probability.

We randomly chose one student who took the exam. For each of the ten ranges below, please provide your belief about the probability the randomly chosen student's score falls in that range. We will randomly pick one range and pay you based on your belief for that range. Your payment will be determined based on the procedure discussed earlier in the instructions. However, remember that you maximize your chance of winning $10 by reporting your belief truthfully.

Probability	Score Range
___________	0-9
___________	0-19
___________	0-29
___________	0-39
___________	0-49
___________	0-59
___________	0-69
___________	0-79
___________	0-89

EXAMPLE **ELICIT AN APPROXIMATE CDF**

Eliciting Several Quantiles

To elicit the (approximate) CDF of a random variable X, select several quantiles and elicit each of these quantiles using any of the mechanisms presented in Section 5.6. Randomly pick a quantile and implement the incentives associated with the mechanism used to elicit that quantile.

We randomly chose one student who took the exam. For each of the probabilities below, provide the number q for which you believe there is an x% chance (the indicated probability) the student's score is below q. We will randomly pick one of the probabilities and pay you based on your prediction for that range. Your payment will be determined based on the procedure discussed earlier in the instructions. However, remember that you maximize your chance of winning $10 by reporting your belief truthfully.

Score	**Probability**
___________	20%
___________	40%
___________	60%
___________	80%

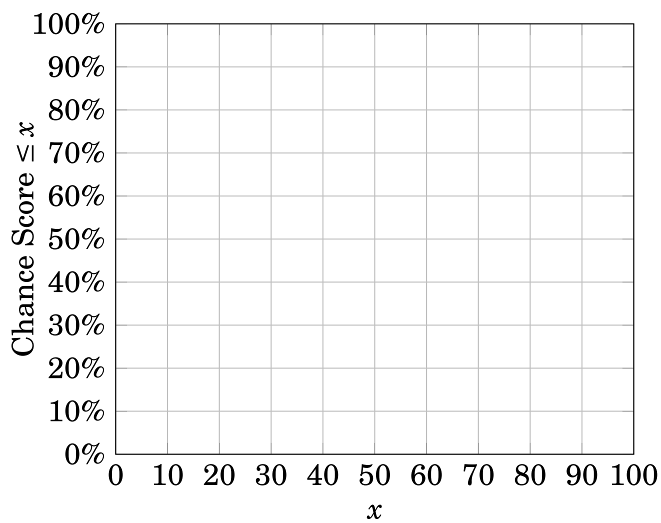
The difficulty in eliciting a precise CDF is in allowing participants to express the CDF. One option is to have them draw the CDF. The probabilities or quantiles can then be inferred from the drawing.

EXAMPLE **ELICIT A CDF**

Drawing the CDF

To elicit the precise CDF of a random variable X, have the participants draw a CDF. Randomly select a point x in the support and find the probability $F(x)$ according to their drawn CDF. Incentivize this probability using any of the mechanisms in Section 5.1.

We randomly chose one student who took the exam. Below, draw a picture that represents the probability you think this student's score is below each possible value from 0 *to* 100. *For instance, if the height or your graph is* 50% *at the score of* 75*, this means you think there is a* 50% *chance the student's score is* 75 *or lower. We will randomly pick a value x and pay you based on your prediction about the probability the score is below x. Your payment will be determined based on the procedure discussed earlier in the instructions. However, remember that you maximize your chance of winning $10 by reporting your belief truthfully.*



Incentives & Incentive Compatibility.
Since this procedure uses multiple probability or quantile elicitations and pays one randomly, this whole procedure is incentive compatible as long as the individual elicitations are incentive compatible and preferences obey *Statewise Monotonicity* (See Axiom 7 in Section 2.3). Here, the "state" here is which report is chosen for payment. However, since statewise monotonicity is already needed for the incentive compatibility of the MPL or single-response BDM, and since it is weaker than S-O reduction (See Axiom 8

in Section 2.4) which is needed for the incentive compatibility of the binarized scoring rule, no additional assumptions are needed to elicit a CDF on top of those needed by the chosen probability or quantile elicitation mechanism.

Implementation Details.
When eliciting a CDF, you may choose to require that their reported distribution is a weakly increasing function on $[0,1]$. If this is not imposed, it's possible that a participant will report a CDF that decreases over some range, which would either reveal a deep failure of incentive compatibility of the mechanism or reveal “beliefs” that are inconsistent with the fundamental axioms of probability (Kolmogorov, 1950). Whether you choose to impose an increasing function depends on whether it's important for you to be able to identify such failures and inconsistencies.

5.10. Approximating Other Statistics of a Distribution

In this chapter, we have tried to be exhaustive about the statistics for which we know or could easily construct mechanisms. If there is some other statistic that your research depends on, we suggest eliciting or approximating an entire distribution via the PDF/PMF (see Section 5.9.1 or CDF (see Section 5.9.2) and then calculating the statistic from this elicited or approximated distribution. Leo and Stelnicki (2024) provide a procedure for fitting a complete CDF through several elicited quantiles using the principle of Maximum Entropy.

6. Implementation Details

In this section we describe implementation issues that arise regardless of the particular statistic that is being elicited. We first discuess issues with the MPL, BDM, and scoring rule elicitation methods. We conclude with a brief discussion of input methods, uncertainty visualization, and hedging.

6.1. Multiple Price Lists

Many of the issues with multiple price lists are also discussed in the previous chapter on preference elicitation (Chapman and Fisher, 2024).

6.1.1. Multiple Switch Points

(Chapman and Fisher, 2024) describe three methods for handling multiple switch points in MPLs. Our suggested implementation of the MPL enforces a single switch-point. Suppose instead that you run an MPL allowing for multiple switch points and in fact observe multiple switching. How should such data be interpreted?

Proposition 5. Under transitivity, any instance of multiple switch-points imply a violation of Monotonicity over Simple Lotteries.

Proof. If there are multiple switch-points then there is a triple $p_1 < p_2 < p_3$ such that either:
$L^{p_1} \succsim f^E, L^{p_2} \precsim f^E, L^{p_3} \succsim f^E$
or
$L^{p_1} \precsim f^E, L^{p_2} \succsim f^E, L^{p_3} \precsim f^E$.
By transitivity, one of these must be true:
$L^{p_2} \succsim L^{p_3}$ or $L^{p_1} \succsim L^{p_2}$.
Since $p_1 < p_2 < p_3$, either case violates stochastic monotonicity over simple lotteries. $\square$

Given this result, one needs to analyze the data through the lens of a theory that allows for violations of monotonicity over simple lotteries. This can be done, for example, by assuming that responses are noisy and estimating the degree of noise. A downside to this approach is that the belief that is inferred depends crucially on the structure of the noise that we assume. Given the arbitrariness of these noise assumptions (absent any data on the structure of noise), we recommend simply forcing a single switch point and allowing any noise to manifest as measurement error in the observed switch point. Again, see the previous chapter (Chapman and Fisher, 2024) for further discussion.

6.1.2. *Order Effects*

One worry with the MPL is that the ordering of the list or the endpoints of the list may affect elicited switch-points. For example, Jack et al. (2022) use MPLs to elicit willingness to pay [WTP] for prepaid electricity vouchers and find that the estimated WTP is different depending on whether the list features increasing or decreasing dollar amounts. Whether this would also be true in the context of belief elicitation is an open question, but is certainly plausible. Jack et al. (2022) also vary which option appears on the left versus right side of the list, but find no significant difference in behavior. Andersen et al. (2006) show that skewing the options in the rows (having more high-valued rows versus having more low-valued rows) can also alter responses.

One suggestion is that showing each row of the MPL on a separate screen and in a randomized order would improve reports. Brown and Healy (2018) show that such a mechanism is more likely to be incentive compatible in the context of risk elicitation, though Dave et al. (2010) and Andersen et al. (2006) argue that it adds confusion and noise to the elicited beliefs, especially with less-educated populations.

Chapman and Fisher (2024) also discuss the possibility of the MPL creating reference points. This has been observed in the domain of risk preferences, but in the domain of belief elicitation there is little scope for changing the available options in the list. Thus, it seems less of a concern in this domain.

6.1.3. *Iterative MPL*

One option for implementing the MPL is to use a coarse grid initially, such as $\{10\%, 20\%, \ldots, 90\%\}$ and then, based on the elicited switch-point, "zoom in" to a finer grid such as $\{30\%, 31\%, \ldots, 39\%, 40\%\}$. (Andersen et al., 2006) find that this iterative MPL has no effect on behavior in the domain of valuations and time preferences, but does affect risk elicitation. The method has been used several times, including in experiments by Holt and Smith (2016) and Burfurd and Wilkening (2018). We note that when using an iterative MPL, you have to select a row randomly from all the rows that could have been encountered. For example, if the finest grid has 1% increments as in the example above, you have to infer a belief from the choice in the finest grid and then randomly choose a row randomly from the full 100-row non-iterative MPL. If a row is only selected from those encountered, the mechanism is not incentive-compatible since a participant might "lie" in the first grid to encounter more favorable rows in the second grid.

6.1.4. Embedded Scroll Box

Healy and Kagel (2023) show the MPL to participants in a separate embedded box on the screen with its own scroll bars. Although the entire list has 101 rows, the participant can only see about 10 rows at a time. They are given a text box in which they can report their belief, and after they type in their belief the MPL is automatically filled out for them (switching at their reported belief) and the scroll bars automatically adjust so that the reported switch-point is centered among the visible rows. In addition, if the participant clicks a row of the MPL then the text box is updated accordingly. Thus, users can either interact through the text box or through scrolling the list and clicking the row on which they are indifferent. Ideally, this method reduces order and framing effects, though its efficacy against other methods has not been tested.

6.2. Single-Response BDM

For many decades, researchers have been wrestling with how to explain the single-response BDM to participants, a fact that we think highlights its inherent complexity.

The single-response BDM has historically been presented as a payment that depends on two random devices. For example Ducharme and Donnell (1973) describe it as follows: the participant chooses an "indifference number" called IN. Then they press a key to generate a random number called FDN, or "first display number". If FDN is weakly less than IN then they are paid five cents if and only if the event of interest occurs. If FDN is greater than IN then they are paid five cents with probability FDN. The way this is operationalized is that the participant pressed a key to generate a "second display number" (SDN). If $SDN < FDN$ then the participant is paid five cents, otherwise they are paid nothing. Grether (1981) used a similar framing, though with Bingo cages.

Holt and Smith (2009) use instructions that help the participant think through their optimal report. They first tell the participant that if their belief about an event E is p then they should be indifferent between (1) a lottery that pays \$1 if E occurs and (2) a lottery that pays \$1 with probability p. Then they explain that they will use dice to generate a number N to construct a "dice lottery" that pays with probability N. Participants are told that if $N < p$ then the resulting dice lottery offers a lower chance of \$1 than (1) and (2) above, so the dice lottery will be rejected and the participant will be paid lottery (1). If $N \geq p$ then the dice lottery gives a higher chance of \$1 than (1) and (2), so the dice lottery will be accepted and used to pay the participant. Thus, Holt and Smith (2009) frame the decision very similarly to an MPL, explicitly viewing it as a binary choice between the bet on E and the lottery with payoff probability N. Although they do not tell participants that truth-telling is in their best interest, they do suggest they "think carefully" about their decision since it will determine when the dice lottery is accepted or rejected.

Möbius et al. (2022) explain the BDM as a "crossover" mechanism, similar to Holt and Smith (2009). They write: "Participants were presented with two options: (1) Receive \$3 if their score was among the top half of scores... (2) Receive \$3 with probability $x \in \{0, 0.01, 0.02, \ldots, 0.99, 1\}$, and asked for what value of x they would be indifferent between them. We then drew a random number $y \in \{0, 0.01, 0:02, \ldots, 0.99, 1\}$ and paid participants \$3 with probability y when $y > x$ and otherwise paid them \$3 if their own score was among the top half."

Karni (2009) points out that the BDM mechanism can be implemented as a sort of English clock mechanism. As described by Hao and Houser (2012), "the individual competes with a dummy bidder and knows that the dummy bidder exits the auction at (unknown) number r. The clock starts at 0 and rises continuously as long as both the individual and the truth-revealing dummy bidder are in the auction. The clock stops when at least one bidder drops out, or when the clock reaches 1, whichever occurs first. If the individual is the first to exit, they receives the lottery [that pays with probability r]; if the dummy bidder exits first, the individual receives the bet [that pays if E occurs]." Following Li (2017), it is an obviously dominant strategy for an individual with belief p to exit when $r = p$.

Hao and Houser (2012) compare the single-response BDM to the ascending clock implementation. They operationalize it via two bags: Bag A has 2 white chips and 8 black chips, while Bag B has 10 chips total but the number of white chips (denoted by q) is equally likely to be any number from one to nine. In the single-response BDM the participant submits a number p (one to nine). If $q \le p$ then they draw a chip from Bag A, otherwise they draw a chip from Bag B. In the clock version a clock starts at one and increases every five seconds. If the participant exits first then they draw a chip from Bag B, otherwise they draw a chip from Bag A. In either case the prize is \$10 if the drawn chip is white.

A downside to the clock mechanism is that many participants never get to reveal their belief because the dummy bidder exits first. Theoretically, this can be mitigated by making the dummy bidder more likely to exit at higher values, though the behavioral effect of this change has not been tested to our knowledge. Hao and Houser (2012) do find that, after controlling for censored observations, the clock mechanism yields a higher percentage of objectively-correct reports, though this difference seems to decline when participants experience the mechanism a second time.

An alternative is to present the BDM as a simultaneous-bid second-price auction. This removes the censoring problem inherent in the clock format. But Marschak (1968) found that participants overbid in value elicitation tasks when the BDM is presented as a simultaneous-bid second-price auction against a random opponent. To this, Savage (1971, p.785) remarks:

> It is no final criticism of such a method to say that participants do not automatically and instinctively understand it or that, understanding, they have psychological difficulty in doing the rational thing. Such facts do underline the need for education and training prior to, and even during, the application of elicitation devices. Incidentally, such education promises to be of great general benefit to the participant and deserves wide promulgation on its own account.

Perhaps the most thorough examination of BDM procedures was performed by Burfurd and Wilkening (2018). They compare the Holt and Smith (2009) explanation to the chips-in-a-bag version of Hao and Houser (2012), and compare these to the MPL used by Trautmann and van de Kuilen (2015) and Holt and Smith (2016). They examine beliefs in an updating task with an objectively-correct Bayesian posterior. They find that all three formats yield similar accuracy and precision, but that the Hao and Houser (2012) chips-in-a-bag framing takes significantly less time to run. In a second experiment, Burfurd and Wilkening (2018) rerun the Hao and Houser (2012) framing but without a computerized understanding quiz. They find that the understanding quiz plays an important role: removing it significantly reduces accuracy of reports. Thus, even though it is time consuming, it appears beneficial to make sure that participants receive adequate training with the mechanism.

When the BDM is used for value elicitation there is some evidence that the support of randomly-drawn prices can affect elicited values (Cason and Plott, 2014; Mamadehussene and Sguera, 2023, e.g.). With belief elicitation, the support is naturally $[0,1]$, so it seems this is less of a concern, though it raises the question of whether the distribution over $[0,1]$ could affect responses. This is one of many interesting open questions about the framing of the BDM mechanism.

6.3. Scoring Rules

Although there have been several tests of different scoring rules, both dollar-denominated and binarized, there have not been any direct comparisons of different presentation formats. However, several presentation formats have been used across studies.

Danz et al. (2022) have participants report their beliefs using a slider and pay via the BQSR. They show the state-contingent probabilities of payment ($s_1(q)$ and $s_0(q)$) as the participant varies their report. They also provide a calculator, which is presented as a second slider. They take the first report p as truthful and then consider the second slider as a report q that may deviate from p. Applying S-O reduction, they report the overall probability of being paid, which is

$$p\,s_1(q)+(1-p)\,s_0(q),$$

as a percentage on the screen. This is clearly maximized when $q=p$, though the function is clearly very flat near the maximum. With this interface they find substantial rates of misreporting.

Healy and Kagel (2023) also study the BQSR. They remove Bayesian updating and only elicit induced prior beliefs with an arguably simplified description of the prior probability. They have participants type their belief into a box, and the calculator (which also assumes S-O reduction) shows the overall probability of being paid for any misreport in a list format rather than a slider. They find very low rates of misreporting, but we're not able to provide specific guidance on which features caused the improved performance since there are several differences between their design and that of Danz et al. (2022).

6.4. Input Methods and Uncertainty Visualization

Some researchers have participants enter their beliefs using a text box while others use a slider. Is one preferred to another? Do sliders generate biases such as pull-to-center? Our knowledge on this question is limited; to date we have not seen careful tests of these questions. Thus, we believe further work in this area would be valuable.

One very important topic that we do not touch on in this chapter is how to display uncertainty to participants. Probabilities can be represented in many different ways, including pie charts, the size of a circle, or a jar filled with marbles of two different colors. There is active research in cognitive science to understand how participants perceive uncertainty given these different visualizations. One recommendation that emerges is that multiple representations may be beneficial. For a survey of recent work, see Padilla et al. (2022).

6.5. Hedging: Ability to Control the Outcome

Consider an experiment when a participant is first asked to take an exam, and then report the mean of their belief about the probability that they earned a passing grade. Will they be tempted to fail the exam on purpose so that their elicited belief is more accurate?

To analyze this question theoretically, consider a simple model of effort on the exam. The participant chooses effort $e \in [0, \bar{e}]$. Their resulting probability of passing is given by the strictly increasing function $\pi : [0, \bar{e}] \to [0, \bar{m}]$, where $\bar{m} < 1$. They are paid via a strictly proper binarized scoring rule with expected payoff at truth-telling given by

$$G(p) = p\, s_1(p) + (1-p)\, s_0(p).$$

Given that they know their own effort e and the function $\pi(\cdot)$, they report a probability of $\pi(e)$ to the mechanism.

Suppose that they are paid \$1 if they pass the exam and their elicitation payoff is scaled by α. Then, overall, they choose e to maximize

$$\pi(e) + \alpha G(\pi(e)).$$

The effect of increasing e is given by

$$\pi'(e)\left[1 + \alpha G'(\pi(e))\right].$$

We know that for strictly proper binarized scoring rules $G'(p)$ must be in $[-1, 1]$. Thus, as long as $\alpha < 1$ the participant should maximize earnings by choosing $e = \bar{e}$. However, note that scoring rules for which $G' > 0$ for all p will unambiguously provide incentives to try hard on the quiz. As Möbius et al. (2022) note, the BDM and MPL have $G' > 0$ everywhere, so these help eliminate any hedging incentive.

One extension of the model, however, might push more towards hedging. Suppose higher effort levels lead to "less certain" or "more ambiguous" beliefs, and the decision maker is averse to this kind of ambiguity. Depending on their aversion to ambiguity, and the rate at which ambiguity increases in effort, this could provide a countervailing force that causes the decision maker to choose lower levels of effort.

What do the data say? Blanco et al. (2010) and Armantier and Treich (2013) find that subjects do hedge in simple settings where the hedging opportunity is obvious, but not in more complex settings. In games the evidence is mixed: Nyarko and Schotter

(2002) don't find evidence of hedging, but Palfrey and Wang (2009) show that beliefs differ between those playing a game and those who are simply observing.

Blanco et al. (2010) suggest that paying randomly for either the task or the belief may reduce hedging. Hedging may also be detectable in post-experiment questionnaires.

ONLINE APPENDIX

Appendix A. Additional Results

A.1. The BQSR for the Mean is Suitable for Coarse Elicitation

Proposition 6. (*BQSR Satisfies Midpoint Property*) For a participant with belief $f(x)$ about a random variable X with support $[a,b]$, the expected probability of winning the prize in the the quadratic scoring rule presented in Section 5.8.3 when announcing a mean m_1 is higher than m_2 when $|m_1-\mu|<|m_2-\mu|$.

Proof. A participant with subjective belief $f(x)$ has the following expected probability of being paid the prize when announcing mean m. Below, let $\mu = E_{f(x)}(x)$.

$$u(m) = \int_a^b f(x)\left(1-\left(\frac{m-x}{b-a}\right)^2\right)dx$$

$$= 1-\frac{1}{(b-a)^2}\int_a^b f(x)(m-x)^2\,dx$$

$$= 1-\frac{1}{(b-a)^2}\left(\int_a^b f(x)m^2dx - \int_a^b f(x)2mxdx + \int_a^b f(x)x^2dx\right)$$

$$= 1-\frac{1}{(b-a)^2}\left(m^2-2m\mu+E_{f(x)}\left(x^2\right)\right)$$

Notice that $u(m)$ is maximized at $\mu=\mu$. Let's compare two values of m_1 and m_2. Expanding $u(m_1)>u(m_2)$:

$$1-\frac{1}{(b-a)^2}\left(m_1^2-2m_1\mu+E_{f(x)}\left(x^2\right)\right) > 1-\frac{1}{(b-a)^2}\left(m_2^2-2m_2\mu+E_{f(x)}\left(x^2\right)\right)$$

$$m_2^2-2m_2\mu > m_1^2-2m_1\mu$$

$$|m_1-\mu|<|m_2-\mu|$$

□

A.1.1. Proof of Lemma 1

Lemma 3. For any differentiable proper scoring rule, the strength of incentives for truth-telling is given by $G''(p)$.

Proof. Let

$$G(q|p) = p\,s_1(q) + (1-p)\,s_0(q)$$

be the expected payoff of announcing q when the true belief is p. Define $G(p) = G(p|p)$. The strength of incentives for truth-telling is given by how fast payoffs drop as q moves away from p, which is

$$\left.\frac{\partial^2 G(q|p)}{(\partial q)^2}\right|_{q=p} = p\,s_1''(p) + (1-p)\,s_0''(p).$$

Or, to make it a positive value, we measure the strength of incentives via

$$-\left.\frac{\partial^2 G(q|p)}{(\partial q)^2}\right|_{q=p}.$$

Since

$$G(p) = p\,s_1(p) + (1-p)\,s_0(p),$$

we have

$$G'(p) = s_1(p) - s_0(p) + \left[p\,s_1'(p) + (1-p)\,s_0'(p)\right] \tag{4}$$

and

$$G''(p) = 2\left[s_1'(p) - s_0'(p)\right] + p\,s_1''(p) + (1-p)\,s_0''(p) \tag{5}$$

$$= 2\left[s_1'(p) - s_0'(p)\right] + \left.\frac{\partial^2 G(q|p)}{(\partial q)^2}\right|_{q=p}. \tag{6}$$

Now, properness of the scoring rule implies that

$$\left.\frac{\partial G(q|p)}{\partial q}\right|_{q=p} = 0,$$

or

$$p\,s_1'(p) + (1-p)\,s_0'(p) = 0,$$

so that means the expression for $G'(p)$ simplifies to

$$G'(p) = s_1(p) - s_0(p). \tag{7}$$

And from this we get

$$G''(p) = s_1'(p) - s_0'(p).$$

Plugging that into 5 thus gives

$$G''(p) = 2G''(p) + \left.\frac{\partial^2 G(q|p)}{(\partial q)^2}\right|_{q=p}$$
$$G''(p) = -\left.\frac{\partial^2 G(q|p)}{(\partial q)^2}\right|_{q=p}$$

as needed. □

A.1.2. Proof of Lemma 2

For any proper binarized scoring rule, the average strength of incentives must be weakly less than two.

Proof. The average strength of incentives is

$$\int_0^1 G''(p)dp$$

This proposition follows directly from equation 7 above:

$$\begin{aligned}\int_0^1 G''(p)dp &= G'(1) - G'(0)\\ &= [s_1(1) - s_0(1)] + [s_0(0) - s_1(0)]\\ &= [s_1(1) - s_1(0)] + [s_0(0) - s_0(1)].\end{aligned}$$

But since the range of s_0 and s_1 must be $[0,1]$, and since $s_1' > 0$ and $s_0' < 0$, we have that $s_1(1) - s_1(0) \leq 1$ and $s_0(0) - s_0(1) \leq 1$, giving the result. □

References

Allen, F., 1987. Notes—Discovering Personal Probabilities When Utility Functions are Unknown. Management Science 33, 542–544. doi:10.1287/mnsc.33.4.542.

Andersen, S., Fountain, J., Harrison, G.W., Rutström, E.E., 2014. Estimating subjective probabilities. Journal of Risk and Uncertainty 48, 207–229. doi:10.1007/s11166-014-9194-z.

Andersen, S., Harrison, G.W., Lau, M.I., Rutstrom, E.E., 2006. Elicitation Using Multiple Price List Formats. Experimental Economics 9, 383–405.

Anscombe, F.J., Aumann, R.J., 1963. A Definition of Subjective Probability. The Annals of Mathematical Statistics 34, 199–205. arXiv:2991295.

Armantier, O., Treich, N., 2013. Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. European Economic Review 62, 14–40.

Azrieli, Y., Chambers, C.P., Healy, P.J., 2018. Incentives in Experiments: A Theoretical Analysis. Journal of Political Economy 126, 1472–1503. doi:10.1086/698136.

Azrieli, Y., Chambers, C.P., Healy, P.J., 2020. Incentives in experiments with objective lotteries. Experimental Economics 23, 1–29. doi:10.1007/s10683-019-09607-0.

Baillon, A., Huang, Z., Selim, A., Wakker, P.P., 2018. Measuring Ambiguity Attitudes for All (Natural) Events. Econometrica 86, 1839–1858. doi:10.3982/ECTA14370.

Beach, L.R., Wise, J.A., 1969. Subjective probability revision and subsequent decisions. Journal of Experimental Psychology 81, 561–565. doi:10.1037/h0027919.

Becker, G.M., DeGroot, M.H., Marschak, J., 1964. Measuring Utility by a Single-Response Sequential Method. Behavioral Science 9, 226–232.

Becker, G.S., 1968. Crime and Punishment: An Economic Approach. Journal of Political Economy 76, 169. doi:10.1086/259394.

Berg, J.E., Daley, L.A., Dickhaut, J.W., O'Brien, J.R., 1986. Controlling preferences for lotteries on units of experimental exchange. The Quarterly Journal of Economics 101, 281–306.

Blanco, M., Engelmann, D., Koch, A.K., Normann, H.T., 2010. Belief elicitation in experiments: Is there a hedging problem? Experimental Economics 13, 412–438.

Brier, G.W., 1950. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review 78, 1–3.

Brown, A.L., Healy, P.J., 2018. Separated decisions. European Economic Review 101, 20–34. doi:10.1016/j.euroecorev.2017.09.014.

Burfurd, I., Wilkening, T., 2018. Experimental guidance for eliciting beliefs with the Stochastic Becker–DeGroot–Marschak mechanism. Journal of the Economic Science Association 4, 15–28. doi:10.1007/s40881-018-0046-5.

Burfurd, I., Wilkening, T., 2022. Cognitive heterogeneity and complex belief elicitation. Experimental Economics 25, 557–592. doi:10.1007/s10683-021-09722-x.

Camerer, C.F., Hogarth, R.M., 1999. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor Production Framework. Journal of Risk and Uncertainty 19, 7–42.

Canen, N., Chakraborty, A., 2023. Belief elicitation in political protest experiments: When the mode does not teach us about incentives to protest. Journal of Economic Behavior & Organization 216, 320–331. doi:10.1016/j.jebo.2023.10.018.

Cason, T.N., Plott, C.R., 2014. Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing. Journal of Political Economy 122, 1235–1270.

Cervera, J., Munoz, J., 1996. Proper scoring rules for fractiles. Bayesian statistics 5, 513–519.

Chapman, J., Fisher, G., 2024. Preference Elicitation: Common Methods and Potential Pitfalls, in: Yariv, L., Snowberg, E. (Eds.), This Volume.

Charness, G., Gneezy, U., Rasocha, V., 2021. Experimental methods: Eliciting beliefs. Journal of Economic Behavior & Organization 189, 234–256. doi:10.1016/j.jebo.2021.06.032.

Cooke, E., 1906. Forecasts and verifications in Western Australia. Monthly Weather Review 34, 23–24.

Cox, J.C., Oaxaca, R.L., 1995. Inducing risk-neutral preferences: Further analysis of the data. Journal of Risk and Uncertainty 11, 65–79. doi:10.1007/BF01132731.

Crockett, S., Oprea, R., 2012. In the long run we all trade: Reference dependence in dynamic economies. University of California Santa Barbara Working Paper .

Danz, D., Vesterlund, L., Wilson, A.J., 2022. Belief Elicitation and Behavioral Incentive Compatibility. American Economic Review 112, 2851–2883. doi:10.1257/aer.20201248.

Dave, C., Eckel, C.C., Johnson, C.A., Rojas, C., 2010. Eliciting risk preferences: When is simple better? Journal of Risk and Uncertainty 41, 219–243.

De Finetti, B., 1937. La prévision: Ses lois logiques, ses sources subjectives, in: Annales de l'institut Henri Poincaré, pp. 1–68.

De Finetti, B., 1962. Does it make sense to speak of "Good probability appraisers"?, in: Good, I. (Ed.), The Scientist Speculates: An Anthology of Partly-Baked Ideas. Basic Books, pp. 257–364.

Dienes, Z., Seth, A., 2010. Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. Consciousness and Cognition 19, 674–681. doi:10.1016/j.concog.2009.09.009.

Ducharme, W.M., Donnell, M.L., 1973. Intrasubject comparison of four response modes for "subjective probability" assessment. Organizational Behavior and Human Performance 10, 108–117. doi:10.1016/0030-5073(73)90007-X.

Dustan, A., Koutout, K., Leo, G., 2022. Second-order beliefs and gender. Journal of Economic Behavior & Organization 200, 752–781.

Dustan, A., Koutout, K., Leo, G., 2024. Reduction in belief elicitation .

Erkal, N., Gangadharan, L., Koh, B.H., 2020. Replication: Belief elicitation with quadratic and binarized scoring rules. Journal of Economic Psychology 81, 102315. doi:10.1016/j.joep.2020.102315.

Eyting, M., Schmidt, P., 2021. Belief elicitation with multiple point predictions. European Economic Review 135, 103700.

Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in Disguise—An Experimental Study on Cheating. Journal of the European Economic Association 11, 525–547. doi:10.1111/jeea.12014.

Gächter, S., Renner, E., 2010. The effects of (incentivized) belief elicitation in public goods experiments. Experimental Economics 13, 364–377. doi:10.1007/s10683-010-9246-4.

Gilboa, I., Schmeidler, D., 1989. Maxmin Expected Utility with Non-Unique Prior. Journal of Mathematical Economics vol. 18, issue 2, pp. 141–153. doi:10.1016/0304-4068(89)90018-9.

Gneezy, U., 2005. Deception: The Role of Consequences. American Economic Review 95, 384–394. doi:10.1257/0002828053828662.

Gneiting, T., 2011. Making and evaluating point forecasts. Journal of the American Statistical Association 106, 746–762.

Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association 102, 359–378.

Grether, D.M., 1980. Bayes Rule as a Descriptive Model: The Representativeness Heuristic. The Quarterly Journal of Economics 95, 537–557. doi:10.2307/1885092.

Grether, D.M., 1981. Financial Incentive Effects and Individual Decisionmaking. Caltech working paper.

Grether, D.M., 1992. Testing bayes rule and the representativeness heuristic: Some experimental evidence. Journal of Economic Behavior & Organization 17, 31–57. doi:10.1016/0167-2681(92)90078-P.

Haghani, M., Bliemer, M.C.J., Rose, J.M., Oppewal, H., Lancsar, E., 2021. Hypothetical bias in stated choice experiments: Part II. Conceptualisation of external validity, sources and explanations of bias and effectiveness of mitigation methods. Journal of Choice Modelling 41, 100322. doi:10.1016/j.jocm.2021.100322.

Hao, L., Houser, D., 2012. Belief elicitation in the presence of naive participants: An experimental study. Journal of Risk and Uncertainty 44, 161–180.

Harrison, G.W., 2014. Hypothetical surveys or incentivized scoring rules for eliciting subjective belief distributions. Center for Economic Analysis of Risk Working Paper Series, Robinson College of Business, Georgia State University .

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

Harrison, G.W., Martínez-Correa, J., Swarthout, J.T., 2014. Eliciting subjective probabilities with binary lotteries. Journal of Economic Behavior & Organization 101, 128–140.

Healy, P.J., Kagel, J., 2023. Testing Elicitation Mechanisms Via Team Chat.

Healy, P.J., Leo, G., 2024. Ternary Belief Elicitation.

Hollard, G., Massoni, S., Vergnaud, J.C., 2010. Subjective Beliefs Formation and Elicitation Rules: Experimental Evidence.

Hollard, G., Massoni, S., Vergnaud, J.C., 2016. In search of good probability assessors: An experimental comparison of elicitation rules for confidence judgments. Theory and Decision 80, 363–387. doi:10.1007/s11238-015-9509-9.

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

Holt, C.A., Smith, A.M., 2016. Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes. American Economic Journal: Microeconomics 8, 110–139. doi:10.1257/mic.20130274.

Hossain, T., Okui, R., 2013. The binarized scoring rule. The Review of Economic Studies 80, 984–1001.

Huck, S., Weizsäcker, G., 2002. Do players correctly estimate what others do?: Evidence of conservatism in beliefs. Journal of Economic Behavior & Organization 47, 71–85. doi:10.1016/S0167-2681(01)00170-6.

Jack, B.K., McDermott, K., Sautmann, A., 2022. Multiple price lists for willingness to pay elicitation. Journal of Development Economics 159, 102977. doi:10.1016/j.jdeveco.2022.102977.

Karni, E., 2009. A Mechanism for Eliciting Probabilities. Econometrica 77, 603–606.

Klibanoff, P., Marinacci, M., Mukerji, S., 2005. A Smooth Model of Decision Making under Ambiguity. Econometrica 73, 1849–1892. doi:10.1111/j.1468-0262.2005.00640.x.

Kolmogorov, A.N., 1950. Foundations of the Theory of Probability. New York: Chelsea Pub. Co.

Lambert, N.S., Pennock, D.M., Shoham, Y., 2008. Eliciting properties of probability distributions, in: Proceedings of the 9th ACM Conference on Electronic Commerce, pp. 129–138.

Leo, G., Stelnicki, S., 2024. Eliciting Subjective Real-Valued Beliefs.

Li, S., 2017. Obviously strategy-proof mechanisms. American Economic Review 107, 3257–87.

Machina, M.J., Schmeidler, D., 1992. A More Robust Definition of Subjective Probability. Econometrica 60, 745–780. doi:10.2307/2951565, arXiv:2951565.

Machina, M.J., Schmeidler, D., 1995. Bayes without Bernoulli: Simple Conditions for Probabilistically Sophisticated Choice. Journal of Economic Theory 67, 106–128. doi:10.1006/jeth.1995.1067.

Mamadehussene, S., Sguera, F., 2023. On the Reliability of the BDM Mechanism. Management Science 69, 1166–1179. doi:10.1287/mnsc.2022.4409.

Marschak, J., 1964. Actual versus consistent decision behavior. Behavioral Science 9, 103–110. doi:10.1002/bs.3830090203.

Marschak, J., 1968. Decision Making: Economic Aspects. Western Management Science Institute.

Massoni, S., Gajdos, T., Vergnaud, J.C., 2014. Confidence measurement in the light of signal detection theory. Frontiers in Psychology 5. doi:10.3389/fpsyg.2014.01455.

McKelvey, R.D., Page, T., 1990. Public and private information: An experimental study of information pooling. Econometrica: Journal of the Econometric Society , 1321–1339.

Möbius, M.M., Niederle, M., Niehaus, P., Rosenblat, T.S., 2022. Managing self-confidence: Theory and experimental evidence. Management Science .

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

Offerman, T., Palley, A.B., 2016. Lossed in translation: An off-the-shelf method to recover probabilistic beliefs from loss-averse agents. Experimental Economics 19, 1–30. doi:10.1007/s10683-015-9429-0.

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

Padilla, L., Kay, M., Hullman, J., 2022. Uncertainty visualization, in: Piegorsch, W.W., Levine, R.A., Zhang, H.H., Lee, T.C.M. (Eds.), Computational Statistics in Data Science. Wiley, pp. 405–421.

Palfrey, T.R., Wang, S.W., 2009. On eliciting beliefs in strategic games. Journal of Economic Behavior and Organization 71, 98–109.

Phillips, L.D., Edwards, W., 1966. Conservatism in a simple probability inference task. Journal of Experimental Psychology 72, 346–354. doi:10.1037/h0023653.

Qu, X., 2012. A mechanism for eliciting a probability distribution. Economics Letters 115, 399–400. doi:10.1016/j.econlet.2011.12.097.

Raiffa, H., 1968. Decision Analysis: Introductory Lectures on Choices Under Uncertainty. Addison Wesley, Reading, Mass.

Ramsey, F.P., 1931. Truth and Probability, in: Braithwaite, R.B. (Ed.), The Foundations of Mathematics and Other Logical Essays. Routledge & Kegal Paul Ltd., London, pp. 156–198.

Roth, A.E., Malouf, M.W., 1979. Game-theoretic models and the role of information in bargaining. Psychological review 86, 574.

Rutström, E.E., Wilcox, N.T., 2009. Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. Games and Economic Behavior 67, 616–632. doi:10.1016/j.geb.2009.04.001.

Savage, L.J., 1954. The Foundations of Statistics. John Wiley & Sons, New York, NY.

Savage, L.J., 1971. Elicitation of Personal Probabilities and Expectations. Journal of the American Statistical Association 66, 783–801.

Schervish, M.J., 1989. A general method for comparing probability assessors. The Annals of Statistics 17, 1856–1879.

Schervish, M.J., Kadane, J.B., Seidenfeld, T., 2012. Characterization of proper and strictly proper scoring rules for quantiles. Preprint, Carnegie Mellon University, March 18.

Schlag, K., Tremewan, J., 2021. Simple belief elicitation: An experimental evaluation. Journal of Risk and Uncertainty 62, 137–155. doi:10.1007/s11166-021-09349-6.

Schlag, K.H., Tremewan, J., van der Weele, J.J., 2015. A penny for your thoughts: A survey of methods for eliciting beliefs. Experimental Economics 18, 457–490. doi:10.1007/s10683-014-9416-x.

Schlag, K.H., Van der Weele, J.J., 2013. Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality. Theoretical Economics Letters 3, 38–42.

Schotter, A., Trevino, I., 2014. Belief Elicitation in the Laboratory. Annual Review of Economics 6, 103–128. doi:10.1146/annurev-economics-080213-040927.

Selten, R., Sadrieh, A., Abbink, K., 1999. Money does not induce risk neutral behavior, but binary lotteries do even worse. Theory and Decision 46, 213–252.

Smith, C.A.B., 1961. Consistency in Statistical Inference and Decision. Journal of the Royal Statistical Society: Series B (Methodological) 23, 1–25. doi:10.1111/j.2517-6161.1961.tb00388.x.

Smith, V.L., 1982. Microeconomic Systems as an Experimental Science. The American Economic Review 72, 923–955. arXiv:1812014.

Sonnemans, J., Offerman, T., 2001. Is the Quadratic Scoring Rule Really Incentive Compatible? Technical Report. Citeseer.

Trautmann, S.T., van de Kuilen, G., 2015. Belief Elicitation: A Horse Race among Truth Serums. The Economic Journal 125, 2116–2135. doi:10.1111/ecoj.12160.

Viscusi, W.K., Evans, W.N., 1998. Estimation of Revealed Probabilities and Utility Functions for Product Safety Decisions. The Review of Economics and Statistics 80, 28–33. doi:10.1162/003465398557302.

Viscusi, W.K., Evans, W.N., 2006. Behavioral Probabilities. Journal of Risk and Uncertainty 32, 5–15. doi:10.1007/s10797-006-6663-6.

Von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behavior. 3rd ed., Princeton University Press, Princeton, NJ.

Walker, J.M., Smith, V.L., Cox, J.C., 1990. Inducing risk-neutral preferences: An examination in a controlled market environment. Journal of Risk and Uncertainty 3, 5–24. doi:10.1007/BF00213258.

Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R., Forsyth, B., 1986. Measuring the vague meanings of probability terms. Journal of Experimental Psychology: General 115, 348.

Wallsten, T.S., Budescu, D.V., Zwick, R., 1993. Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments. Management Science 39, 176–190. doi:10.1287/mnsc.39.2.176.

Wang, S.W., 2011. Incentive effects: The case of belief elicitation from individuals in groups. Economics Letters 111, 30–33. doi:10.1016/j.econlet.2010.11.045.

Wilde, L.L., 1981. On the Use of Laboratory Experiments in Economics, in: Pitt, J.C. (Ed.), Philosophy in Economics: Papers Deriving from and Related to a Workshop on Testability and Explanation in Economics Held at Virginia Polytechnic Institute and State University, 1979. Springer Netherlands, Dordrecht, pp. 137–148. doi:10.1007/978-94-009-8394-6_9.

Wright, W.F., Aboul-Ezz, M.E., 1988. Effects of extrinsic incentives on the quality of frequency assessments. Organizational Behavior and Human Decision Processes 41, 143–152. doi:10.1016/0749-5978(88)90023-4.

Wright, W.F., Anderson, U., 1989. Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. Organizational Behavior and Human Decision Processes 44, 68–82. doi:10.1016/0749-5978(89)90035-6.