Rank-Based Elicitation Schemes for Relative Likelihoods of Events

Vivek Bhattacharya and Benjamin N. Roth

Department of Economics, Massachusetts Institute of Technology

March 8, 2016

Abstract. Rank-based elicitation mechanisms, which elicit *relative* likelihoods of events rather than numerical probabilities, are becoming common in development economics. We consider rank-based scoring rules, in which agents rank events by their success probabilities. Though some rank-based scoring rules are sensitive to risk preferences, we provide two schemes to approximately elicit a large class of probability distributions regardless of risk preferences. We discuss a simple, indirect implementation which utilizes only the language of relative likelihoods already used in practice. Finally we conduct numerical experiments to demonstrate its usefulness in eliciting probabilistic bounds.

Keywords. Belief elicitation, development economics, rank order elicitation, scoring rule.
JEL Codes. C60, D81, D83.

1. Introduction

Uncertainty—from the fluctuation of daily income to the expression of civil unrest—is a fundamental characteristic of life in the developing world. The efficient distribution of aid often relies on knowledge of the distribution of uncertain events. To distribute targeted cash transfers development practitioners may like to know the likelihood a villager's income falls below a certain threshold. To distribute microloans a microfinance institution might like to know the likelihood a potential borrower repays her loan or the distribution of marginal returns to capital for her primary business.

Development economists expend considerable effort eliciting subjective expectations of the outcomes of uncertain, future events. In doing so, they face two main problems. First, many respondents have low levels of numerical literacy and may have difficulty grasping the notion of an exact probability. Second, when asking community members about sensitive information, they may wish to distort their reports to shed favorable light on themselves and their friends. Alternatively, respondents may become fatigued during the course of an exercise and consquently provide inaccurate reports.[1] Delavande, Giné, and McKenzie (2011) provide a review of the use of

Contact. bhattacv@mit.edu and benroth@mit.edu. This material is based on work supported by the NSF Graduate Research Fellowhip under Grant No. 1122374. We thank Itai Ashlagi, Abhijit Banerjee, Glenn Ellison, Scott Kominers, David Parkes, Parag Pathak, and Rob Townsend for helpful comments and discussions. Any errors are our own.

[1] Alatas, Banerjee, Hanna, Olken, and Tobias (2012) document such fatiguing effect in their PRA whereby respondents in that reports that are provided later in the session are less accurate than those provided earlier.

elicitation mechanisms in development economics and discuss that practitioners face such issues in the field.

In this paper, we study a class of mechanisms that we call *rank-based scoring rules* which have the potential of addressing both these problems and can be feasibly implemented in practice in the field. First, even when participants have low numeracy, they can gauge the intuitive notion of *relative* likelihoods (the concept that event A is more likely than event B). Indeed, elicitation mechanisms based on *ranks* rather than probabilities are already common in the field.[2] Second, a rank-based scoring rule also specifies a payment scheme that rewards respondents for the ex post accuracy of their reports to counteract any incentives to distort reports. Unlike rank-based schemes themselves, these sorts of incentives have been slow to gain traction in development economics, and Delavande, Giné, and McKenzie (2011) attribute this to a dearth of payment rules that are both robust to varying degrees of risk aversion and that can be coupled with simple elicitation schemes. We study one such payment rule in this paper.

We explore the extent to which rank-based scoring rules can be used to elicit both relative and absolute probabilities of events, both from a theoretical and a practical perspective. We aim to bring insights from the literature on proper scoring rules—solutions to the problem of eliciting features of an unknown probability distribution from utility maximizing agents—to the task of incentivizing truthful rankings of uncertain events.[3]

Our first observation provides a cautionary tale for practitioners: without the assumption of risk neutrality, it is not possible to guarantee that arbitrary events will be ranked truthfully in such schemes. Delavande, Giné, and McKenzie (2011), among other papers, have made this point in the context of eliciting probabilities directly, but our observation is that risk-averse individuals may even interchange the *relative* probabilities of a set of events.

Our second observation—and indeed the main theoretical contribution of this paper—is that by careful construction of auxiliary events one can elicit the *probabilities* of *arbitrary* events via a rank-based scoring rule. We provide two theoretical constructions that elicit these probabilities to arbitrary precision. First, we discuss a mechanism that directly refers to events with objective probabilities. Second—and perhaps more interestingly—we provide one which only uses relative comparisons between events. We then discuss how this mechanism can be implemented in practice to establish informative bounds on probabilities of events when the researcher desires more information than relative rankings alone would provide. To elicit the relative likelihoods of uncertain events, we have respondents rank the likelihood of events of interest interesected with auxiliary events with known probability. With carefully chosen auxiliary events we can obtain arbitrarily tight bounds on the events of interest while maintaining incentive compatibility from agents with arbitrary risk preferences.

Our most substantial practical contribution is a natural indirect implementation of our mechanism

[2]Examples include Alatas, Banerjee, Hanna, Olken, and Tobias (2012), Bandiera, Burgess, Das, Gulesci, Rasul, and Sulaiman (2013), and Banerjee, Duflo, Chattopadhyay, and Shapiro (2011) who all utilize Participatory Rural Appraisals (PRAs) to target transfers of cash and business assets to those most in need. PRAs elicit this information by asking community members to rank one another based on who is most in need or who is most likely to fall below a certain level of wealth or income.

[3]Note we do not refer to a setting where the *outcome* is a rank ordering. Rather, we refer to the task of eliciting the relative likelihoods of the probabilities of events.

that explicitly elicits relative magnitudes rather than numerical probabilities. That is, it elicits responses to questions of the form "Is event A at least x times more likely than event B?" truthfully from agents with arbitrary risk preferences. We show that asking just a few questions of this form can elicit probabilities to a remarkable degree of precision, and we provide an algorithm to determine the order in which to ask these questions that provides an optimal worst-case guarantee. To our knowledge this is the first paper to provide a method to incentivize relative elicitation of this form.

The elicitation of relative rather than absolute likelihoods mirrors the types of questions already utilized in the field: development researchers usually ask respondents to rank events that are similar in nature, e.g., whether various clients will repay their loans, or whether various villagers will be in financial distress within the next year. Respondents may be able to compare the probability of an event to that of a similar one but not necessarily to that of a dissimilar one—or even to an objective probability. For a stylized example, a villager may be able to evaluate whether her sister is at least twice as likely as her neighbor to have more than \$50 of business income in the following month, without being able to evaluate whether either event has probability larger than 1/2. This reasoning may be especially apt for respondents with low numeracy and may be one of the driving motivations for the already widespread use of rank order elicitation mechanisms in development, which by definition elicit only relative magnitudes and not numerical probabilities.

Aside from the empirical literature discussed above, our paper is closely related to the theoretical literature on proper scoring rules. Much of this literature focuses on the case in which agents are risk-neutral or their utility functions are otherwise known.[4] However, there has been a recent effort to generate scoring rules that are robust to risk preferences. Karni (2009) provides an elegant payment rule resembling the mechanism in Becker, DeGroot, and Marschak (1964) that elicits the probability of an event from any expected utility maximizer, and Qu (2012) extends this rule to elicit an arbitrary distribution from any expected utility maximizer. Following Savage (1971) and Roth and Malouf (1979), Hossain and Okui (2013) note that expected utility maximizers behave as if they were risk-neutral over lotteries and show that many of the standard scoring rules used in the literature under the assumption of risk neutrality are robust to risk aversion when they pay lottery tickets rather than money. An alternate method, studied by Offerman, Sonnemans, Van De Kuilen, and Wakker (2009), is to elicit features of the agent's utility function and use these to adjust the agent's report accordingly. Holt and Smith (2016) study a menu version of the Becker-Degroot-Marschak lottery mechanism which is robust to risk preferences and show that it performs well in the lab. Schlag, Tremewan, and van der Weele (2014) provide a comprehensive survey of both theoretical and experimental results. In a similar light, this paper can be viewed as providing a novel scoring rule to elicit a broad class of distributions to arbitrary precision regardless of the respondent's degree of risk aversion.

Our method also relates to a literature studying mechanisms that ask for simpler reports to approximate a probability distribution, in contrast to the mechanisms that elicit direct reports in the papers listed above. Friedman (1983) considers eliciting histogram approximations of general probability distributions. More recently, Schlag and Tremewan (2014) present a mechanism that

[4]See, for instance, Brier (1950), Gneiting and Raftery (2007), Lambert, Pennock, and Shoham (2008), or Witkowski and Parkes (2012).

recovers bounds on beliefs rather than exact probabilities. Similarly, we propose eliciting a rank-order list and using it to approximate a probability distribution. Our indirect implementation, which we call a *question based mechanism* relies on direct elicitation of responses to questions about relative magnitudes. It too is thus a method of approximating a probability distribution based on simpler reports.

The rest of this paper proceeds as follows. Sections 2 and 3 contain our main theoretical results. Practitioners interested in the implementation of our payment scheme to elicit relative likelihoods can skip directly to Section 4. In Section 2 we describe the class of rank-based scoring rules, show that they do not elicit rankings of arbitrary events truthfully, and then provide a set of natural correlation structures under which they are indeed truthful. In Section 3 we show how these results can be extended to elicit more general probability distributions, and we provide two mechanisms that do so. Section 4 discusses practical implementation, including how to use rank-based scoring rules to incentivize relative comparisons, which relative comparisons to elicit, the order in which to ask relative comparisons, and numerical experiments. Section 5 concludes. Appendix A discusses technical details associated with the practical implementation, and Appendix B collects proofs omitted from the body of the paper.

2. A Rank-Based Scoring Rule

A *rank-based scoring rule for a collection of events* $\mathcal{A} \equiv \{A_i\}_{i=1}^N$ is a set of rewards $\{a_j\}_{j=1}^N$ such that $a_j > a_{j-1} \geq 0$ that are delivered to the agent in the following manner. The agent, who is assumed to know the probability of each of these events, is asked to rank the events in increasing order of probability. He reports a permutation σ so that $A_{\sigma(j)}$ is the event that the agent ranks in position j. The state of the world is then realized, so that some subset of events are realized as successes. If $A_{\sigma(j)}$ is a success, then the agent is paid a_j, so that his total payoff is $\sum_j \mathbb{1}_{A_{\sigma(j)}} a_j$, where $\mathbb{1}_A$ is the indicator of the event A.

Rank-based proper scoring rules by themselves are not a novel concept; see Lambert (2011), for instance. Nevertheless, they are an interesting class of mechanisms to study not just because of the similarity with mechanisms already in use by development economists but also because the mechnism and payment scheme are especially natural, simple to explain, and quite flexible.

Of course, these benefits of rank-based scoring rules are only relevant if we can identify circumstances under which the incentives provided by the mechanism induce agents to report the ranking truthfully. We say a rank-based scoring rule for a particular set of events is *proper* if any expected utility maximizer with a strictly increasing utility function will strictly prefer to rank the events in this set truthfully. While we focus on expected utility maximizers, all results hold for any agent whose preferences over monetary lotteries are monotone with respect to first order stochastic dominance.

Throughout the paper, we will be concerned with whether rank-based proper scoring rules can be used to elicit the distribution of events $\{E_i\}_{i=1}^n$. Note that to elicit this distribution, we could consider a rank-based scoring rule for the events $\{E_i\}_{i=1}^n$ themselves. Alternatively, we could consider a rank-based scoring rule for a *different* set of events $\{A_i\}_{i=1}^N$ which are constructed from the $\{E_i\}$, and then use the information about $\{A_i\}$ to determine information about $\{E_i\}$. In this

section, we will be concerned with the case where the $\{A_i\} = \{E_i\}$ and we are designed rank-based scoring rules directly for the set of events of interst. We will relax this restriction in Section 3.

Suppose we are interested in events $\{E_i\}$ and would like to design rank-based proper scoring rules for these events. It is easy to see that a risk-neutral agent will always rank the events truthfully; this observation is a simple application of the rearrangement inequality and a reflection of the fact that the correlation between events does not enter the agent's utility. However, a general utility maximizer would not ignore the correlation structure and may have an incentive to misreport the true ranking.

As an example, suppose a microfinance institution (MFI) would like to screen a set of three clients for repayment capacity. Each client owns a business that succeeds with some probability known to the community but unknown to the MFI. Clients A and $\bar{A}$ are direct competitors and thus only one can succeed. Suppose A succeeds with probability $4/5$ and $\bar{A}$ succeeds with complementary probability $1/5$. Client B is in an independent industry and succeeds with probability $1/2$ independent of A and $\bar{A}$. A risk neutral agent in the community asked to rank his peers and paid according to a rank-based scoring rule (for the events $\{A, \bar{A}, B\}$, which correspond to clients A, $\bar{A}$, and B succeeding, respectively) would rank his peers truthfully. However a risk averse agent may not. Reporting $(B, \bar{A}, A)$ instead of the truthful ranking $(\bar{A}, B, A)$ allows the agent to face a lottery with lower mean but lower variance as well. A sufficiently risk-averse utility maximizer would prefer to misreport. For instance, if we consider the rank-based scoring rule for these events with $a_1 = 0$, $a_2 = 1$, and $a_3 = 2$, then if the agent has a utility function $u(x) = x^\alpha$, the agent will prefer to report $(B, \bar{A}, A)$ instead of $(\bar{A}, B, A)$ for $\alpha \lesssim 0.422$.

While arbitrary events are not necessarily ranked truthfully, we can easily nevertheless check that if the events in $\mathcal{E}$ are known to satisfy some standard correlation structures, then they are indeed ranked truthfully by any expected utility maximizer (and also regardless of the specific payoffs). The correlation structures we study are mutual exclusivity, nestedness, and independence. Note that a set of events $\mathcal{E}$ is *nested* if for any E_i and E_j in $\mathcal{E}$, either E_i happens whenever E_j happens, or vice versa.

Lemma 1. *Suppose that we have events $\{E_1, \ldots, E_n\}$ such that either (i) the E_i are mutually exclusive, (ii) the E_i are nested, or (iii) the E_i are independent. Then any rank-based scoring rule $\{a_j\}_{j=1}^n$ for these events $\{E_i\}$ is proper.*

Lemma 1, whose proof is provided in Appendix B, notes that if the events of interest are known to satisfy any of the above correlation structures, simple incentive schemes are sufficient to elicit their relative likelihoods truthfully regardless of risk preferences.[5] Indeed, since Lemma 1 holds for any (increasing) sequence of payoffs a_j, a researcher also has a great deal of flexibility in designing a rank-based scoring rule while still ensuring that it is proper.

Perhaps more importantly, while the correlation structures in Lemma 1 seem rather special, the designer can *construct* the events that the agent is asked to rank from the events of interest

[5]That we can elicit the ranks of certain events truthfully is not on its own a surprising result. For instance, Peysakhovich and Plagborg-Møller (2012) and Kadane and Winkler (1988) show that for a large class of scoring rules (e.g., the Brier scoring rule) that ask for probabilities of mutually exclusive events, the probabilities are misreported but the ranking remains truthful. In these scoring rules, however, the payoffs of the agent are determined directly by the probabilities reported, and the designer does not necessarily have much control over these payoffs.

in such a way that the ranked events do satisfy these correlation structures. In Section 3, we will show that the ability to truthfully elicit the rankings of mutually exclusive events is quite powerful: rankings themselves can be used to elicit *probabilities* to arbitrary precision (regardless of the correlation structure of the underlying events of interest). Straightforward simplifications of this mechanism, which essentially amount to asking for a rank followed by some questions about the *relative* magnitudes of the underlying probabilities, then allow us to recover informative bounds on the beliefs.

3. Eliciting Probabilities from Ranks

In this section, we discuss two separate schemes to elicit probabilities of events of interest. The innovation in both schemes involves noting that the events $\{E_i\}$ of interest need not coincide with the events $\{A_i\}$ that are part of the rank-based scoring rule. Indeed, by careful choice of events $\{A_i\}$ to be ranked, it is possible to elicit the *probabilities* of $\{E_i\}$ to arbitrary precision.

In Section 3.1 we discuss one scheme to elicit the probabilities of arbitrary events, and an especially simple indirect implementation. However, this scheme makes reference to numerical probabilities which, as argued in the introduction, may constitute a significant disadvantage in contexts where respondents have a low degree of numerical literacy. Section 3.2 discusses an alternative elicitation scheme for events with arbitrary correlation structure that forgoes reference to numerical probabilities. This mechanism forms the basis for our discussion of incentivizing relative elicitation in Section 4. Section 3.3 discusses an extension of the mechanisms in Sections 3.1 and 3.2 for the approximate elicitation of many continuous distributions.

3.1. Incentivizing Comparisions to Explicit Probabilities

Suppose we are interested in eliciting the approximate probabilities of the events $\{E_i\}_{i=1}^n$ using a rank-based scoring rule. A natural way to elicit these probabilities would be to have the respondents rank these events $\{E_i\}$ of interest, along with events with known probabilities of success. Specifically, consider a series of independent weighted coin flips $\{C_i\}_{i=1}^m$ with probability of success q_i, with $q_i < q_{i+1}$ for all i. If the relative rankings of the $\{C_i\}$ and the $\{E_i\}$ can be elicited, then by choosing q_i and q_{i+1} to be sufficiently close to each other, one can elicit the probabilities of the $\{E_i\}$ to arbitrary precision.

Consider, therefore, a rank-based scoring rule for the set of events $\{E_i\}_{i=1}^n \cup \{C_i\}_{i=1}^m$. These events collectively need not satisfy any of the correlation structures discussed in Lemma 1, and there is no guarantee that an arbitrary expected utility maximizer will rank these events truthfully. Indeed, our example above shows that two mutually exclusive events A and $\bar{A}$, and an event B independent from both A and $\bar{A}$ need not be ranked truthfully by a risk averse respondent. However, we provide a construction below that yields the intuitive indirect implementation described in the above paragraph. Namely it elicits the ranking of $\{E_i\}_{i=1}^n \cup \{C_i\}_{i=1}^m$ truthfully regardless of risk preferences.

Scheme 1. *Consider an arbitrary set of events* $\{E_i\}_{i=1}^n$

- *Construct a set of weighted coin flips* $\{C_i\}_{i=1}^m$ *with probability of success* q_i*, with* $q_i < q_{i+1}$ *for all* i*.*

- *Construct a set of* $n+m$ *mutually exclusive events* $\{R_{Ei}\}_{i=1}^n \cup \{R_{Ci}\}_{i=1}^m$*, each of which has probability* $1/(n+m)$ *and is independent of* $\{E_i\} \cup \{C_i\}$*.*

- *Implement a rank-based scoring rule for* $\{E_i \cap R_{Ei}\}_{i=1}^n \cup \{C_i \cap R_{Ci}\}_{i=1}^m$*.*

We have the following proposition related to Scheme 1.

Proposition 1. *The rank-based scoring rule in Scheme 1 for the events* $\{E_i \cap R_{Ei}\}_{i=1}^n \cup \{C_i \cap R_{Ci}\}_{i=1}^m$ *is proper. Further, for any* $\epsilon > 0$*, there exists a set of weighted coin flips* $\{C_i\}_{i=1}^m$ *such that the rank elicited from such a scoring rule will allow one to recover each* p_i *to within* ϵ *of the true value, for any set of events* $\{E_i\}$*.*

Proof. Because the events $\{R_{Ei}\}_{i=1}^n \cup \{R_{Ci}\}_{i=1}^m$ are mutually exclusive, the events $\{E_i \cap R_{Ei}\}_{i=1}^n \cup \{C_i \cap R_{Ci}\}_{i=1}^m$ are mutually excusive as well, and thus by Lemma 1, the events $\{E_i \cap R_{Ei}\}_{i=1}^n \cup \{C_i \cap R_{Ci}\}_{i=1}^m$ will be ranked truthfully by an expected utility maximizer. We clearly have $\Pr(E_i \cap R_{Ei}) = \Pr(E_i)/(n+m)$, and similarly $\Pr(C_i \cap R_{Ci}) = q_i/(n+m)$, and therefore relative rankings of $\{E_i \cap R_{Ei}\}$ and $\{C_i \cap R_{Ci}\}$ will allow us to bound the probabilities of the events $\{E_i\}$ between consecutive q_i. We can choose the coin flips to make the q_i as close to each other as desired. □

While the construction in Proposition 1 requires agents to rank synthetic events, there is another, more natural, interpretation of this mechanism. At most one of the events $\{E_i \cap R_{Ei}\}_{i=1}^n \cup \{C_i \cap R_{Ci}\}_{i=1}^m$ will occur, and thus the agent will be paid for at most one success. To implement this mechanism, therefore, we can first ask the agent to rank the events $\{E_i\} \cup \{C_i\}$. Upon eliciting this ranking, we can randomly choose an integer k between 1 and $n+m$ and then pay the agent exactly a_k from the scoring rule if the event in position k succeeds (even if other events succeed as well).

The main concern with the mechanism in Proposition 1, however, is that it requires agents to be able to compare probabilities of dissimilar events, and perhaps even to numerical probabilities. Indeed, in most settings in the field, the events $\{E_i\}$ of interest are likely similar—such as the chance that different community members repay their loans—and can be easily compared by the agents. They may have considerably more difficulty in comparing the probabilities of these events to the numerical probabilities embedded in the $\{C_i\}$.

3.2. Incentivizing Relative Comparisons Between Events

In this section we discuss an alternative payment scheme that truthfully elicits the probabilities of arbitrary events to arbitrary precision. Importantly, this scheme admits an indirect implementation that eschews mention of numerical probabilities, instead relying exclusively on relative comparisons of the likelihoods of the events of interest, i.e., by eliciting the ratios of probabilities of the events of interest. We elicit these relative probabilities with the following construction.

Scheme 2. *Consider an arbitrary set of events* $\{E_i\}_{i=1}^n$*.*

- *Construct an independent set of mutually exclusive events $\{\{R_j^i\}_{j=1}^m\}_{i=1}^n$, with objective, known probabilities r_j^i such that $r_j^i = r_j^{i'} = r_j$ for all i, i', $r_1 \geq r_2 \geq \cdots \geq r_m$ and $\sum_{i,j} r_j^i = 1$.*[6]

- *Use a rank-based scoring rule for the events $\{E_i \cap R_j^i\}_{i,j}$ to elicit the relative likelihoods of these events.*

The scoring rule in Scheme 2 operates on a simple logic. By incentivizing the respondent to rank the events $\{E_i \cap R_j^i\}_{i,j}$ we are able to bound the *relative* magnitudes of the p_i's. The benefit is demonstrated in the following simple example. Suppose we are interested in the probability p_E of event E. We can use a rank-based scoring rule for the events E and $\overline{E}$ (the complement of E), which are mutually exclusive. The respondent will thus inform us of the relative ranking of p_E and $1 - p_E$, or, in other words, whether $p_E \in [0, \frac{1}{2}]$ or $p_E \in [\frac{1}{2}, 1]$. Now suppose that we introduce an auxiliary event C (a coin flip) that is independent of E and succeeds with probability 1/3. This can be implemented by having the designer create a lottery independent of the events in the community with 1/3 probability of success. Since the events $\{E \cap C, E \cap \overline{C}, \overline{E} \cap C, \overline{E} \cap \overline{C}\}$ are still mutually exclusive, we can use a rank-based scoring rule based for these four events to elicit their ranks. This will give us the ranking of not only p_E and $1 - p_E$, by comparing the rank of $E \cap C$ with $\overline{E} \cap C$, but also $p_E \cdot 1/3$ and $(1 - p_E) \cdot 2/3$, by comparing the rank of $E \cap C$ with $\overline{E} \cap \overline{C}$. By having the respondent rank these four events we learn whether $p_E \in [0, \frac{1}{3}]$, $p_E \in [\frac{1}{3}, \frac{1}{2}]$, $p_E \in [\frac{1}{2}, \frac{2}{3}]$, $p_E \in [\frac{2}{3}, 1]$.

By choosing the auxiliary events carefully, we have the following result. Note that while part (ii) of the result is stated for mutually exclusive events, we will extend this result to arbitrary events of interest.

Proposition 2. *(i) The rank-based scoring rule for the events $\{E_i \cap R_j^i\}_{i,j}$ in Scheme 2 is proper. (ii) Suppose further that $\{E_i\}_{i=1}^n$ are mutually exclusive and exhaustive events, with probabilities p_i so that $\sum_i p_i = 1$. Moreover, for any $\epsilon > 0$, there exists a set $\{\{R_j^i\}_{j=1}^{m_i}\}_{i=1}^n$ such that the ranks elicited from the scoring rule in Scheme 2 will allow one to recover each p_i to within ϵ of the true value.*

Proof. Part (i) follows directly from Lemma 1(i) and the fact that the events $\{E_i \cap R_j^i\}_{i,j}$ are mutually exclusive. We thus focus on (ii). Note that the event $E_i \cap R_j^i$ has probability $p_i r_j^i$. It then remains to choose the r_j^i in a particular fashion so as to get a tight bound on the probabilities of each event. Let

$$r_j^i \equiv r_j \equiv \left(\frac{2}{m} - \frac{1}{m^2} - \frac{2(j-1)}{m^2}\right)\frac{1}{n} \tag{1}$$

for $j \leq m$ and $r_{m+1}^i = 0$.

Suppose for concreteness that the most likely event is E_1.[7] Then, we will have a series of inequalities of the form $r_{k_i} p_1 \geq r_1 p_i \geq r_{k_i+1} p_1$ for each i: the first inequality is guaranteed since

[6] We can easily extend this scheme to the case where the probabilities of the events $\{R_j^i\}$ and $\{R_j^{i'}\}$ do not coincide, but we avoid doing so to simply make the notation—and the comparison to question-based mechanisms in Section 4 cleaner.

[7] This will be revealed by the agent's report and is thus without loss of generality.

$r_1 p_1 \geq r_1 p_i$ for all i, and the latter inequality is guaranteed since $r_1 p_i \geq 0$. Then, we have that

$$p_i \in \left[p_1 \frac{r_{k_i+1}}{r_1}, p_1 \frac{r_{k_i}}{r_1} \right].$$

The width of this interval is $(p_1/r_1)(r_{k_i} - r_{k_i+1}) \leq (r_{k_i} - r_{k_i+1})/r_1$. Now, we have that

$$\frac{r_{k_i} - r_{k_i+1}}{r_1} = \begin{cases} \frac{2/m^2}{2/m - 1/m^2} = \frac{2}{2m-1} & \text{if } k_i < m \\ \frac{1/m^2}{2/m - 1/m^2} = \frac{1}{2m-1} & \text{if } k_i = m \end{cases}.$$

It remains to find a bound on p_1. Note that $p_1 = 1 - \sum_{j=2}^{n} p_j$. But, we know each p_j is within an interval of length at most $2/(2m-1)$. Then, p_1 must lie in an interval of length $2(n-1)/(2m-1)$. Now simply choose m large enough such that $2(n-1)/(2m-1) < \epsilon$. □

Note that our ability to elicit relative (i.e., ratios of) probabilites is insensitive to the correlation structure of $\{E_i\}_{i=1}^n$. The assumption of mutual exclusivity in Proposition 2 allows us to map relative magnitudes to actual probabilities; many other types of information (e.g., knowledge of the expected number of events that will succeed, or knowledge about the probability of a specific event) will also allow us to recover actual probabilities. However, the assumption that $\{E_i\}_{i=1}^n$ are mutually exclusive is of course not restrictive from a theoretical perspective: we may begin with events $\{E_i\}_{i=1}^n$ that are not known to be mutually exclusive. We can construct 2^n mutually exclusive events by simply partitioning the outcome space fully. Then, the probability of each E_i will be the sum of the probabilities of 2^{n-1} of these synthetic events. Thus, to back out the probabilities of E_i to within ϵ, we can use the result in Proposition 2 to back out the probabilities of the synthetic events to within $\epsilon/2^{n-1}$.

The construction above is crude, and we make no claims about the optimality of our choice of r_i in the proof of Proposition 2.[8] From a conceptual standpoint, the important point of the construction is that r_i can be chosen independently of the E_i, and as such the implementation of this scoring rule does not depend on the designer having any special information about the events. In Section 4, we provide an indirect implementation of Scheme 2 to highlight that eliciation need not require directly ranking a large number of synthetic events.

3.3. Eliciting Continuous Distributions

The construction for discrete distributions in Propositions 1 and 2 can be extended to arbitrary distributions by binning the support. In this section, we formalize this approximation. Let $\mathscr{P}$ be a finite partition of some subset S of $\mathbb{R}^N$ and let $\mathbf{p}$ be a probability distribution over $\mathscr{P}$. Let λ be a probability distribution over S. We denote the *continuous extension of* $\mathbf{p}$ *to* S (with respect to some fixed measure λ) as $\nu(\mathbf{p}, \mathscr{P})$. For any set $A \subseteq S$ such that there exist an element $\mathscr{P}(A)$ of

[8] However, Section 4 does compare our choice to other natural choices to show that arbitrary choices of r_i do not provide much information about probabilities, even as the set of events becomes infinite.

the partition $\mathscr{P}$ with $A \subseteq \mathscr{P}(A)$,[9] we define

$$\nu(\mathbf{p}, \mathscr{P})(A) \equiv \frac{\lambda(A)}{\lambda(\mathscr{P}(A))} \cdot p_{\mathscr{P}(A)}, \tag{2}$$

where $p_{\mathscr{P}(A)}$ is the probability that $\mathbf{p}$ assigns to the set $\mathscr{P}(A)$.[10] We can define the measure on any other set A' as

$$\nu(\mathbf{p}, \mathscr{P})(A') \equiv \sum_{P \in \mathscr{P}} \nu(\mathbf{p}, \mathscr{P})(P \cap A'),$$

where $\nu(\mathbf{p}, \mathscr{P})(P \cap A')$ is defined via (2) since $\mathscr{P}(P \cap A') = P$ if $P \in \mathscr{P}$.

Suppose μ is an unknown probability distribution on $\mathbb{R}^N$ that the designer wishes to elicit. We place no restrictions on μ: it can be continuous, discrete with unknown support, or a mixture of the two. Suppose that the support of μ is known to be contained in some set S.

Proposition 3. *Suppose S is compact. Then for any $\epsilon > 0$, there is a partition $\mathscr{P}$ and a rank-based scoring rule such that the continuous extension (to S) of the distribution elicited from this rank-based scoring rule is within ϵ of μ with respect to the Prokhorov metric.*[11]

Proposition 3 formally connects two approximation results: a continuous distribution can be approximated by a discrete distribution, and this discrete distribution can in turn be approximated by the one elicited by the rank-based scoring rule. Once again, the conceptual appeal of this construction is that a partition $\mathscr{P}$ can be chosen to approximate *any* distribution, and the designer does not need any a priori information about the distribution.

Unfortunately, elicitation is no longer possible if S is not known to be compact. Intuitively, if S is not compact then at least one element of $\mathscr{P}$ must be unbounded, and it will be difficult to approximate distributions that have most of their mass concentrated in this element of the partition. We formally encode this logic in the following result.

Proposition 4. *Suppose S is not compact. Then there exists sufficiently small $\epsilon > 0$ such that for any partition $\mathscr{P}$, we can always find a distribution μ such that the continuous distribution elicited from any rank-based scoring rule is farther than ϵ away from μ.*

The concern in the above argument is that most of the mass of the distribution lies in the large set of the partition: a histogram of the age distribution of a population that bins ages five and

[9] That is, with a slight abuse of notation, we let $\mathscr{P}$ denote both the partition as well as the function that takes a set A to the element of the partition that contains A.

[10] We suppose for simplicity that all elements of $\mathscr{P}$ are measurable and have positive measure with respect to λ. It is easy to conceptualize this as λ being a multivariate normal distribution—or a uniform distribution if S is compact—and the elements of $\mathscr{P}$ being nontrivial intervals (or products of intervals, more generally) so that ν bears resemblance to a histogram. It is in this manner that a discrete distribution (the probability in each bin of the histogram) is transformed into a continuous one over the whole space (e.g., subsets of each bin). It is possible to extend this definition in a natural way to cases in which elements of $\mathscr{P}$ are measure zero (e.g., finite collections of points), but doing so simply clutters notation and is tangential to the main goal of the paper.

[11] For two probability distributions μ and ν, the Prokhorov distance between μ and ν is defined to be

$$d_P(\mu, \nu) \equiv \inf\{\epsilon > 0 : \mu(B) \leq \nu(B^\epsilon) + \epsilon \text{ for all Borel sets } B\},$$

where $B^\epsilon \equiv \{x : \inf_{y \in B} \|x - y\| \leq \epsilon\}$. See Huber (2004).

above into one group will not be especially informative of the true age distribution. It is of course possible to impose simple assumptions to relax the compactness in Proposition 3 but avoid the impossibility result of Proposition 4. For instance, suppose the designer has some knowledge of the tail behavior of the distribution; that is, he knows f such that $\mu(\{x : \|x\| \geq R\}) \leq f(R)$ with $f(R) \to 0$ as $R \to \infty$. Then, the partition can be created by choosing R large enough so that $f(R) < \epsilon$, partitioning the set $\{x : \|x\| \leq R\}$ as in Proposition 3, and including $\{x : \|x\| > R\}$ as the final set in $\mathscr{P}$. However, the main message of Proposition 4 is that there are limits to how well *arbitrary* continuous distributions can be elicited purely via a rank-based scoring rule, and this inherent tradeoff must be taken into consideration if a researcher wishes to elicit such distributions.

4. Practical Implementation

4.1. Question-Based Mechanisms

One may object that the construction presented in Section 3.2 loses part of the practical appeal of the rank-based scoring rules discussed in Section 2. The agent is now required to internalize the probabilities of the auxiliary events R^i_j, comprehend the laws of probability to compute the probabilities of the synthetic events $E_i \cap R^i_j$, and rank a much larger set of events. Of course, one way elicit this rank via a direct mechanism: instead of asking the agent to rank a large set of synthetic events, we can ask him for the probabilities of the events E_i. Proposition 2 can then be interpreted as saying that for any expected utility maximizer, all reports that are weakly dominant will be within ϵ of the true distribution (as measured by the sup norm in $\mathbb{R}^n$).[12] However eliciting probabilities directly seems to eliminate the motivating appeal of a rank-based mechanism, and as such, we discuss a more intuitive elicitation mechanism in this section.

The mechanism in Scheme 2 can be implemented in an especially simple way by asking the respondent to rank the events E_i and then answer a short series of yes/no questions. Importantly, these questions only involve comparing the *relative* probabilities of the events E_i. For respondents with low levels of numeracy, this may be a substantial advantage. It may be easy for respondents to evaluate the relative likelihoods of default for their peers holding microloans without being able to place a meaningful numerical probability on the likelihood of any individual peer's default. As argued in the introduction, this logic may motivate the prevalence of rank-based elicitation in existing development efforts and is thus a natural criterion for desirability.

This mechanism can be illustrated via the example discussed at the beginning of Section 3.2. Suppose that the respondent is asked to rank the event E (which he knows has probability 1/4) and its complement $\overline{E}$; the auxiliary events are C (which succeeds with known probability 1/3) and its complement $\overline{C}$. As discussed above, a direct application of the mechanism in Section 3.2 would be to ask the agent to rank all four synthetic events generated by the intersection of these events, and the respondent would order them as $\overline{E} \cap \overline{C}$, $\overline{E} \cap C$, $\overline{C} \cap E$, and $E \cap C$ from most to least likely. Instead, we can first ask the respondent to rank the initial events of interest (i.e., answer that $\overline{E}$ is more likely than E). We can then ask "Is $\bar{E}$ at least twice as likely as E?" and again receive an

[12] The same logic can be applied for a similar interpretation for Proposition 3. One can ask for a distribution function, and all weakly dominant reports will be within ϵ of the true distribution function as measured by the Prokhorov metric.

answer of "yes." This single question following the elicited rank is rather simple to understand, and the response is enough to rank all the synthetic events and use the payment scheme outlined in Section 3.2.

More generally, consider arbitrary events E_i and $E_{i'}$ of interest and auxiliary events R_j and $R_{j'}$ (with probabilities r_j and $r_{j'}$). Instead of explicitly asking for a ranking between $E_i \cap R_j$ and $E_{i'} \cap R_{j'}$, one can ask the respondent the yes/no question "Is E_i at least $q = r_j/r_{j'}$ times as likely as E_i'?" Thus, the decision of which questions to ask can be reduced to deciding which values $\{q_k\}_{k=1}^K$ to ask about, with each $q_k \in (0, \infty)$. We will refer to such an implementation as a *question-based elicitation mechanism*. Given a set of events $\{E_i\}$ and a set $\{q_k\}$, a question-based elicitation mechanism is one in which the agent is asked (1) to rank the events and (2) answer a series of yes/no questions comparing the ratio $\Pr(E_i)/\Pr(E_j)$ to some value q_k. The mechanism terminates when each ratio is bounded between two consecutive elements of $\{q_k\}$, and payment is based on the elicited bounds and one realization of the events.

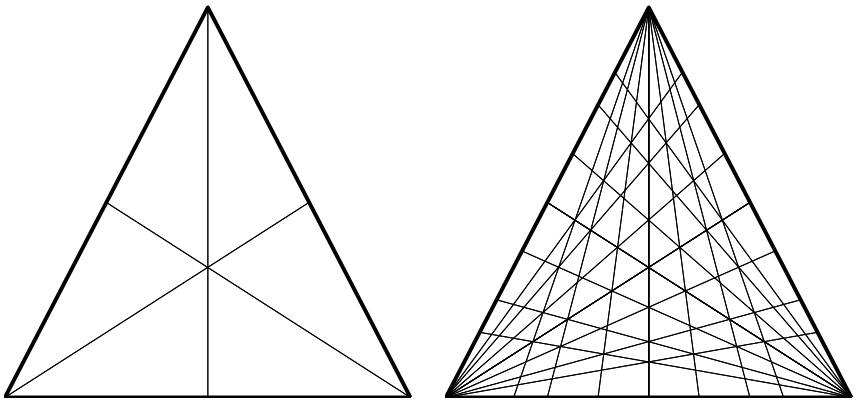
Note that there is a clear connection between the rank-based scoring rule proposed in Scheme 2 and question-based mechanisms. We encode this connection in the following proposition, although the formal equivalence requires a slight generalization of rank-based scoring rules, which we call *multiple list rank-based scoring rules*, and we leave the details to Appendix A.1.

Proposition 5. *(i) For a rank-based scoring rule described in Scheme 2, there exists a set $\{q_k\}$ such that the answers elicited from the question-based mechanism with this set can be used to recover the desired ranking of the events $\{E_i \cap R_j^i\}$ in the rank-based scoring rule. (ii) For a question-based mechanism $\{q_k\}$, there exists a multiple list rank-based scoring rule such that there is a one to one correspondence between a response in the question based mechanism and a ranking in the multiple list rank-based scoring rule.*

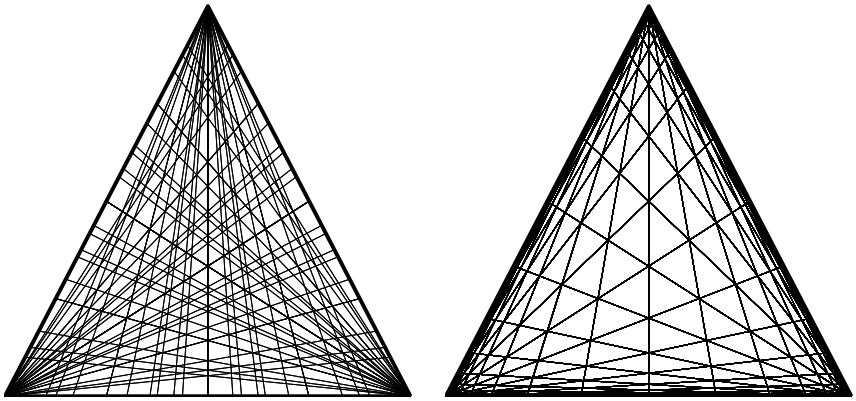
The proof of Proposition 5 involves noting the connection between q_k and ratios of different r_i's. While the result is natural, it is nevertheless quite useful. In particular, Proposition 5(i) formally justifies that a question-based mechanism is an indirect implementation of the rank-based scoring rule in Scheme 2, and the question-based mechanism is a much more intuitive method of eliciting the required information from agents with low numeracy. Secondly, Proposition 5(ii) justifies a focus directly on question-based mechanisms: for any question-based mechanism, we can determine a payoff scheme so that the responses to the questions will be elicited truthfully from the agent. This payoff scheme is based on (a generalization of) a rank-based scoring rule, and strategyproofness will thus follow directly from the results developed in Sections 2 and 3. For the remainder of this section, therefore, we will focus on how to choose the $\{q_k\}$ in a question-based mechanism (Section 4.2) as well as on the efficient order in which to ask the questions (Section 4.3). We will conclude with some numerical examples which illuminate the practicality of our scheme (Section 4.4).

4.2. Which Questions Should One Ask?

One way to analyze the choice of $\{q_k\}$ is to note that it corresponds to a particular partition of the $(n-1)$-simplex (if there are n events). The main restriction on this partition is that it must be generated by hyperplanes of the form $p_i - q_k p_j = 0$, along with the faces of the simplex. For $n = 3$



(a) $m = 1$ (b) $m = 3$



(c) $m = 5$ (d) $m = 15$, $q_i = 2^i$

Figure 1: Partitions of the simplex induced by the mechanism proposed in Scheme 2, for various m using both (1) and an alternate formula for q_i (corresponding to $r_i = 2^{-i}$).

events, the simplex is simply a triangle, and the partition is generated by line segments emanating from each vertex.

Figure 1 plots the simplex that contains all probability distributions over three events, and (a)–(c) illustrate the construction in Proposition 2. Panel (a) depicts the basic mechanism, where only the ranks are elicited; here, the report will allow the designer to place the distribution in one of the six triangles. Panels (b) and (c) depict the mechanism with $m = 3$ and $m = 5$, where q_k are chosen according to ratios of r_i given by (1). The elements of the partition—which correspond to the uncertainty of the final probability elicited—are rather small even for $m = 3$. If using the direct rank-based mechanism for these events, these choices of m correspond to ranking nine and fifteen events, respectively, and depending on the report, the probability distribution will be placed in one of the elements of the partition of the simplex. However, note that the number of questions required to find the appropriate partition of the simplex may not be large at all and can indeed be as few as three for $m = 3$. This number is investigated in more detail in the computational experiments in Section 4.4.

With the choice of $\{q_k\}$ given in Panels (a)–(c), as $k \to \infty$ (which corresponds to taking $m \to \infty$ auxiliary events) the elements of the partition become progressively smaller. This serves as an illustration of Proposition 2. By contrast, Panel (d) illustrates that arbitrary choices of $\{q_k\}$ will not be as informative even asymptotically: taking $q_i = 2^i$ does not allow one to precisely elicit distributions near the middle of the simplex, even for much larger k. Informally viewed through the lens of Figure 1, the hyperplanes generated by the corresponding $\{q_k\}$ tend to bunch near the edges of the simplex as $k \to \infty$ instead of being distributed more evenly throughout the simplex. However, the choice of $q_i = 2^i$ does provide more precision for probability distributions near the edges of the simplex.

Summarizing the discussion, there are two main observations from Figure 1. First, Proposition 2 serves more than simply a conceptual purpose: it provides guidance as to which choices of questions will be informative about probabilities *in the limit*. Moreover, careful choices of these questions can elicit precise estimates across all potential probability distributions without sacrificing much simplicity. The scheme in (b) involves asking *many* fewer questions than the one in (d), but it elicits probabilites to approximately the same level of precision. The second observation is that a researcher with a strong prior about the probability distribution can choose the questions accordingly: if she believes that the distribution is likely to lie near the boundary of the simplex, she may choose to use a set of questions corresponding to $q_i = 2^i$ even though increasing the number of questions will not increase the accuracy if the distribution happens to lie near the middle of the simplex.

We should mention a final consideration when choosing the set of questions $\{q_k\}$. The main motivation for a rank-based elicitation scheme is the ease with which it can be explained to respondents with low levels of numerical literacy. Respondents may have a good sense of the relative probabilities of events and may be able to easily answer whether an event is *twice* or *thrice* as likely as a similar event. However, they may have a much harder time answering whether an event is 1.15 times more likely than a different one, say. As such there may be a tradeoff between the difficulty questions induced by the ranking exercise and the precision of the inferred bounds.

4.3. How Should One Ask the Questions?

The scheme suggested in Section 4.2 is simple. First, elicit the relative ranks of the events E_i by asking the respondent. Second, based on these ranks, ask a series of yes/no questions about the relative probabilities of pairs of events adaptively so as to recover the rankings of the synthetic events introduced in the mechanism in Section 3.2. The first step is straightforward. The second step requires more exploration.

The natural goal in the second step, given the desire to apply this mechanism in the field, is to minimize the number of questions that must be asked upon fixing the set $\{q_k\}$.[13] If the researcher has a well-founded prior on the probability distribution to be elicited, the natural "optimal" sequence of questions would be the one that minimizes the expected number of questions asked. In practice, the researcher may not have such a prior. As such, our goal will be to find a "policy" function of questions—i.e., the choice of question to ask as a function of the answers received so far—to

[13]Such a consideration can be traced back to the development literature, as discussed in the Introduction. Delavande, Giné, and McKenzie (2011) note that fatigue is a concern when applying elicitation methods in the field.

minimize the worst-case number of questions that must be asked.

We find this optimal policy as the solution to a discrete dynamic programming problem. We relegate the technical details of the algorithm to Appendix A.2, but we provide an outline of the algorithm here. Fix a set $\{q_k\}_{k=1}^K$ as a parameter of the problem, and let $q_0 \equiv 0$ and $q_{K+1} \equiv \infty$ for notational convenience. A *state* of the algorithm is a set $\{(\underline{k}_{ij}, \overline{k}_{ij})\}_{1 \le i < j \le n}$ with $0 \le \underline{k}_{ij} < \overline{k}_{ij} \le K+1$. This state corresponds to the knowledge that $q_{\underline{k}_{ij}} \le p_i/p_j \le q_{\overline{k}_{ij}}$. At each state, the researcher can ask a number of *questions*, and a question from a state $\{(\underline{k}_{ij}, \overline{k}_{ij})\}$ is a pair (i, j) and a k such that $\underline{k}_{ij} < k < \overline{k}_{ij}$: this corresponds to asking "Is p_i/p_j greater than q_k?" Note that a question necessarily leads to one of two possible states, depending on the answer. The algorithm to find the optimal question from each state essentially proceeds via backward induction. There are a set of states from which there are no possible questions (i.e., we have recovered all ratios of probabilities to the precision allowed by $\{q_k\}$). We then find all states from which we can ask a question that necessarily leads to one of these terminal states: it is of course optimal to ask this question from such states, since that guarantees that only one question will be needed. We next find all states from which we can ask a question that necessarily leads to states from which at most one question will be needed. We proceed until we have exhausted all the states.

4.4. Numerical Experiments

In this section, we study the performance of the algorithm in situations that resemble those in the field. We consider small sets of events (three or four mutually exclusive events) whose probabilities are distributed on the simplex following a Dirichlet distribution. We vary the parameters of the Dirichlet distribution as well as both the fineness and the spacing of the partition.

We draw probabilities from a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$, varying α from 0.01 to 100. Note that $\alpha = 1$ corresponds to the uniform distribution on the simplex. Smaller values of α correspond to distributions that are concentrated around the faces of the simplex (e.g., one event is much more likely than other events) and large values of α correspond to distributions concentrated near the center of the simplex. For concreteness, Figure 2 plots the average value of the order statistics for $n = 3$ and $n = 4$ events, as a function of α.

We also use two separate sets of $\{q_k\}$, listed explicity in Table 1. The sets in the first column, labeled "Even", are chosen such that the corresponding hyperplanes are evenly spaced in the simplex. (In the 2-simplex, they correspond to line segments that subdiving the opposite edge evenly.) The sets in the second column, labeled "Double", correspond to comparing ratios of probabilities to powers of two. As discussed in Section 4.2, increasing the number of elements of such sets will not lead to identifying all possible probabilities to arbitrary precision, but such questions may be more intuitive to ask, and they may also have desirable properties in practice.

For each choice of $\{q_k\}$, we first compute the optimal policy function following the algorithm discussed in Section 4.3. We then draw 100,000 probability distributions for each value of α and use the optimal policy function to find the cell $C(p)$ in the partition in which the probability distribution p lies. We keep track of the number of questions needed to identify this element of the partition.

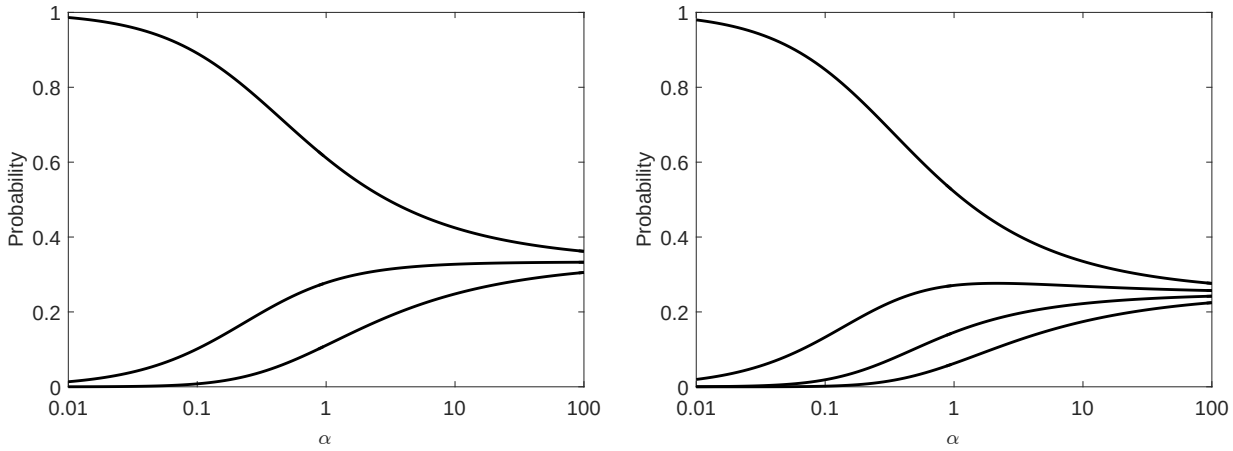


Figure 2: Mean probabilities of the most likely (top) to least likely (bottom) events for a Dirichlet distribution parameterized by α, for $n = 3$ (left) and $n = 4$ (right).

K	Even	Double
3	$\{1/3, 1, 3\}$	$\{1/2, 1, 2\}$
5	$\{1/5, 1/2, 1, 2, 5\}$	$\{1/4, 1/2, 1, 2, 4\}$
7	$\{1/7, 1/3, 3/5, 1, 5/3, 3, 7\}$	$\{1/8, 1/4, 1/2, 1, 2, 4, 8\}$

Table 1: Sets $\{q_k\}$ as a function of the total number of elements K.

We also compute a "worst-case" error for each distribution $(p_1, \ldots, p_n)$, defined as

$$\max_{\hat{p} \in C(p)} \frac{1}{n} \sqrt{\sum_{i=1}^{n} (p_i - \hat{p}_i)^2}. \tag{3}$$

The program (3) computes the largest error one could make (normalized by the number of events) if one randomly picked a point in the cell $C(p)$ after eliciting it from the respondent via the series of questions. Note first that this error is infeasible to calculate in practice: since p is not elicited exactly, the researcher will not know the actual magnitude of the error for any given realization of p. Secondly, the error could potentially be reduced in practice by choosing a point within each cell C as the best guess of p. Since the "optimal" point should depend in principle on the prior, we stack the deck against finding low errors and take a worst-case approach to analyzing the mechanism.

Figure 3 plots the mean number of questions asked by the optimal policy (after eliciting the order by asking the respondent to rank the events). For $K = 3$ and $n = 3$ exactly 3 questions are required regardless of from which distribution the probabilities are drawn. However, increasing K to 5 only requires about 1 to 1.5 more questions on average—and even fewer for larger values of α. Increasing K further to 7 can require as many as 6 questions on average and as few as 4. In general, more questions are required to elicit the specific cell of the distribution when the probabilities are concentrated around the edges of the simplex. Moreover, $\{q_k\}$ that follow the "double" pattern

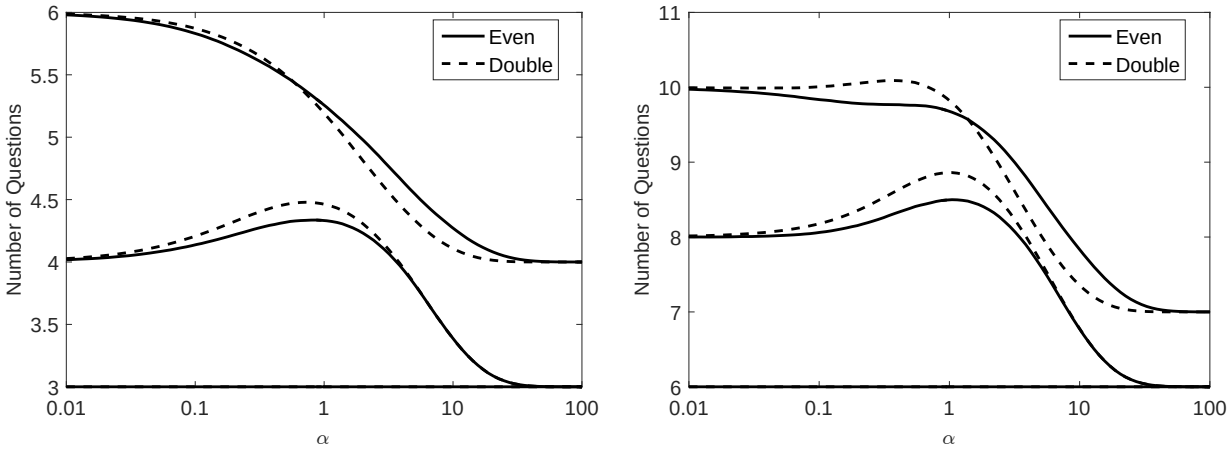


Figure 3: Expected number of questions required using the optimal policy, as a function of the of $\{q_k\}$ as well as α, for $n = 3$ (left) and $n = 4$ (right). The top lines correspond to $K = 7$, the middle lines to $K = 5$, and the (constant) bottom lines corresponds to $K = 3$.

tend to require slightly more questions for low α and, for $K = 7$, slightly fewer questions for high α. These qualitative patterns are similar for $n = 4$, with $K = 3$ requiring 6 questions instead of 3, $K = 5$ requiring around 7–9 on average, and $K = 7$ requiring between 8 and 10 (and possibly fewer for distributions that are especially concentrated around the center of the simplex).

Figure 4 plots the mean value of (3) over the draws of p. For $n = 3$, the error at $\alpha = 1$ is about 0.037 for $K = 3$, 0.028 for $K = 5$, and 0.021 for $K = 7$ when using the "even" $\{q_k\}$. For lower α, the error increases (since the cells of the partition near the edges of the simplex have larger diameter on average) but always stays below 0.11. When using the "double" $\{q_k\}$, the mean error is about 0.049 for $K = 3$, 0.026 for $K = 5$, and 0.019 for $K = 7$—and it always stays below 0.15. Figure 4 allows for some observations comparing errors with the evenly spacing $\{q_k\}$ with errors when using the "double" $\{q_k\}$. Evenly spaced sets tend to have smaller mean errors when distributions are concentrated near the edges for low K. It is only when K is larger ($K = 7$ here) that the "double" $\{q_k\}$ show an advantage for distributions concentrated around the edges of the simplex and a disadvantage for distributions concentrated around the center of the simplex. This corresponds to the discussion in Section 4.2 that choosing a partition of the simplex that corresponds to $q_i = 2^i$ creates finer cells at the edges of the simplex at the expense of coarser cells near the middle. Moreover, one can note that increasing K does not reduce the mean error for high α when using the "double" $\{q_k\}$: increasing K for such sets does not reduce the size of the cells near the center of the simplex, which is where most of the mass is concentrated for high α. The patterns are, once again, similar for $n = 4$ and the worst-case errors tend to be slightly *lower* on average.

The main takeaway from these figures is that asking the relatively small number of questions is often sufficient to elicit probabilities to within a reasonably small error—possibly eliciting each probability to within a few hundreths. We envision that researchers who wish to implement this mechanism in the field can use simulations like these, along with considerations like the ones discussed in Section 4.2, to decide on the set $\{q_k\}$.

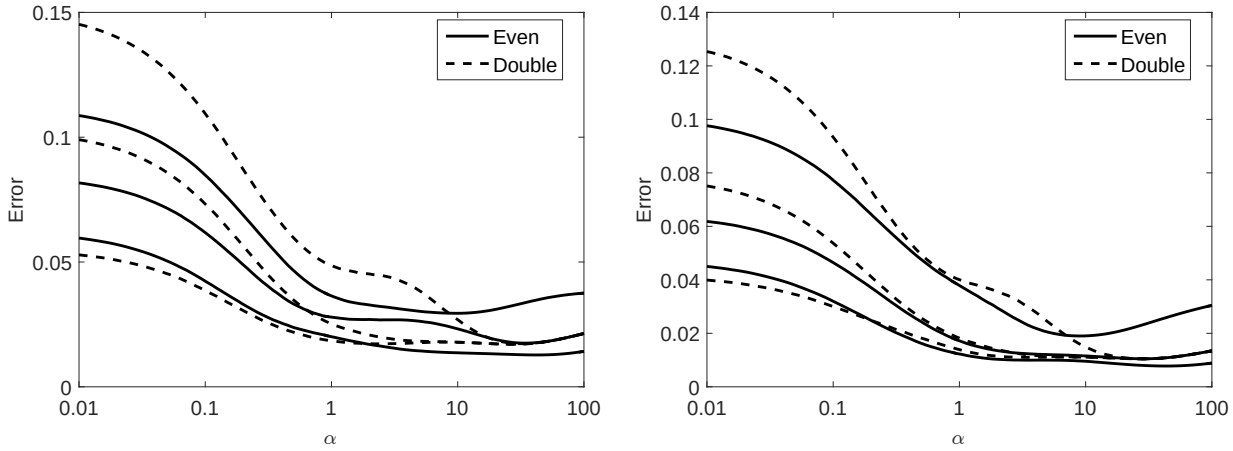


Figure 4: Mean worst-case error (3) achieved using the optimal policy, as a function of the of $\{q_k\}$ as well as α, for $n = 3$ (left) and $n = 4$ (right). The top lines correspond to $K = 3$, the middle lines to $K = 5$, and the bottom lines to $K = 7$.

While computing the optimal policy function may be tedious for large sets of events of large sets of $\{q_k\}$, note that it need only be computed once for a set of $\{q_k\}$. That is, the optimal policy of how to ask questions need not be computed on the fly in the field. However, we do recognize that computing the optimal algorithm may be tedious for large sets of events. In Appendix A.3, we provide and analyze a heuristic algorithm to determine the questions to ask without relying on computing the solution to the dynamic programming problem. This algorithm also provides some bounds on the number of questions and the worst-case error which may be reasonable in practice.

5. Conclusion

Motivated by the prevalence of rank-based elicitation schemes for potentially sensitive information in development economics, we study the extent to which a rank-based scoring rule for a set of events can be used to truthfully elicit rank order information about their relative likelihood. We provide a cautionary example, echoed in much of the literature on proper scoring rules, that arbitrary events may not be truthfully ranked by individuals who are not risk-neutral. However, certain classes of events are indeed truthfully ranked. We extended this logic to show that similar, albeit substantially more complicated rank-based scoring rules can elicit a large class of distributions to arbitrary precision.

We discuss an indirect implementation of our mechanism that only utilizes questions about the relative likelihoods of events—a language already commonly used in development surveys. Thus our paper provides a field-ready mechanism to incentivize responses to the types of questions already being asked. Given a set of desired comparisons, we provide an algorithm for determining the optimal adaptive sequence of questions.

Researchers interested in recovering only the ranking of the likelihood of events can do so using simple payment rules. For those who are interested in recovering more precise bounds on the probabilities of events, we provide guidance in evaluating the tradeoff between simplicity of the

elicitation scheme and precision, and show that when the number of events is fairly small, rich information can be elicited using only a few auxiliary questions.

References

Alatas, V., A. Banerjee, R. Hanna, B. Olken, and J. Tobias (2012): "Targeting the Poor: Evidence from a Field Experiment in Indonesia," *American Economic Review*, 102(4), 1206–1240.

Bandiera, O., R. Burgess, N. C. Das, S. Gulesci, I. Rasul, and M. Sulaiman (2013): "Can Basic Entrepreneurship Transform the Economic Lives of the Poor?," IZA Discussion Paper 7386, IZA.

Banerjee, A., E. Duflo, R. Chattopadhyay, and J. Shapiro (2011): "Targeting the Hard-Core Poor: An Impact Assessment," Working paper, MIT.

Becker, G., M. DeGroot, and J. Marschak (1964): "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science*, 9(3), 226–232.

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

Delavande, A., X. Giné, and D. McKenzie (2011): "Measuring subjective expectations in developing countries: A critical review and new evidence," *Journal of Development Economics*, 94(2), 151–163.

Friedman, D. (1983): "Effective scoring rules for probabilistic forecasts," *Management Science*, 29(4), 447–454.

Gneiting, T., and A. E. Raftery (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102(477), 359–378.

Holt, C., and A. Smith (2016): "Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes," *American Economic Journal: Microeconomics*, 8(1), 110–139.

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

Huber, P. J. (2004): *Robust Statistics*. Wiley.

Kadane, J. B., and R. L. Winkler (1988): "Separating probability elicitation from utilities," *Journal of the American Statistical Association*, 83(402), 357–363.

Karni, E. (2009): "A Mechanism for Eliciting Probabilities," *Econometrica*, 77(2), 603–606.

Lambert, N., D. Pennock, and Y. Shoham (2008): "Eliciting Properties of Probability Distributions," *Proceedings of the 9th ACM Conference on Electronic Commerce*.

Lambert, N. S. (2011): "Elicitation and evaluation of statistical forecasts," *Working Paper*.

Offerman, T., J. Sonnemans, G. Van De Kuilen, and P. P. Wakker (2009): "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes," *Review of Economic Studies*, 76(4), 1461–1489.

Peysakhovich, A., and M. Plagborg-Møller (2012): "A Note on Proper Scoring Rules and Risk Aversion," *Economics Letters*, 117(1), 357–361.

Qu, X. (2012): "A Mechanism for Eliciting a Probability Distribution," *Economics Letters*, 115, 399–400.

Roth, A. E., and M. W. Malouf (1979): "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, 86(6), 574–594.

Savage, L. J. (1971): "Elicitation of personal probabilities and expectations," *Journal of the American Statistical Association*, 66(336), 783–801.

Schlag, K., and J. Tremewan (2014): "Simple Belief Elicitation," Working paper, University of Vienna.

Schlag, K., J. Tremewan, and J. van der Weele (2014): "A Penny for Your Thoughts: A Survey of Methods for Eliciting Beliefs," Working Paper 1401, University of Vienna.

Witkowski, J., and D. Parkes (2012): "A Robust Bayesian Truth Serum for Small Populations," *In Proceedings of the 26th AAAI Conference on Artificial Intelligence*.

A. Further Discussion of Practical Implementation

In this appendix, we first define a multiple list rank-based scoring rule and state a simple lemma about incentive compatibility. We then discuss the method of implementing a question based mechanism in more detail.

A.1. Map Between Question-Based Mechanisms and Rank-Based Scoring Rules

As mentioned in Section 4.2, it is not the case that arbitrary choices of $\{q_k\}$ correspond to a choice of $\{r_i\}$ that would allow us to directly apply the payment scheme from the rank-based scoring rule in Proposition 2. For that we need a generalization of rank-based scoring rules defined below.

A *multiple list rank-based scoring rule for a collection of events* $\bigcup_{\ell=1}^{L}\{E_j^\ell\}_{j=1}^{J_\ell}$ is a set of rewards $\bigcup_{\ell=1}^{L}\{a_j^\ell\}_{j=1}^{J_\ell}$ such that $a_j^\ell > a_{j-1}^\ell \geq 0$ that are delivered to the agent in the following manner. The agent, who is assumed to know the probability of each of these events, is asked to form L rank order lists, where the ℓ^{th} list is a ranking of the events $\{E_j^\ell\}_{j=1}^{J_\ell}$ in increasing order of probability. He reports L permutations $\{\sigma^\ell\}_{\ell=1}^{L}$ so that $A_{\sigma^\ell(j)}$ is the event that the agent ranks in position j in list ℓ. The state of the world is then realized, so that some subset of events are realized as successes. If $A_{\sigma^\ell(j)}$ is a success, then the agent is paid a_j^ℓ, so that his total payoff is $\sum_{j,\ell} \mathbb{1}_{A_{\sigma^\ell(j)}} a_j^\ell$, where $\mathbb{1}_A$ is the indicator of the event A. We say a multiple list rank-based scoring rule for a particular set of events is *proper* if any expected utility maximizer with a strictly increasing utility function will strictly prefer to rank the events in this set truthfully.

We now state a simple result for multiple list rank-based scoring rules that is analogous to Lemma 1. The proof is written out for completeness in Appendix B.

Lemma 2. *Suppose that we have events* $\bigcup_{\ell=1}^{L}\{E_j^\ell\}_{j=1}^{J_\ell}$*, all of which are mutually exclusive. Then any multiple list rank-based scoring rule* $\bigcup_{\ell=1}^{L}\{a_j^\ell\}_{j=1}^{J_\ell}$ *for these events is proper.*

Together with Proposition 5, Lemma 2 shows that truthtelling can be incentivized in question-based mechanisms by using the payment scheme from the associated multiple list rank-based scoring rule.

A.2. Details of the Algorithm in Section 4.3

We first repeat some notation from Section 4.3. The set $\{q_k\}_{k=1}^K$ is fixed, and $q_0 \equiv 0$ and $q_{K+1} \equiv \infty$. A *state* a set $\{(\underline{k}_{ij}, \overline{k}_{ij})\}_{1 \leq i < j \leq n}$ with $0 \leq \underline{k}_{ij} < \overline{k}_{ij} \leq K+1$. This state corresponds to the knowledge that $q_{\underline{k}_{ij}} \leq p_i/p_j \leq q_{\overline{k}_{ij}}$. A state is *admissible* if these implied inequalities (for each pair (i,j)) can be satisfied for some p_i such that $\sum p_i = 1$. Restrict attention only to admissible states. A *question* from a state $\{(\underline{k}_{ij}, \overline{k}_{ij})\}$ is a pair (i,j) and a k such that $\underline{k}_{ij} < k < \overline{k}_{ij}$: this corresponds to asking "Is p_i/p_j greater than q_k?" Let $Q(s)$ denote the set of questions that can be asked at a state s. Let $A(q)$ denote the possible answers to a question. Note that it suffices to restrict our attention to *irreducible* states, i.e., ones such that $|A(q)| = 2$ for all $q \in Q(s)$. If for a particular state s, there exists a q such that $A(q) = \{s'\}$, then being in the state s automatically implies being in s', so we need not ask this question (which explains the term irreducible). Let S denote the set of admissible and irreducible states; we will restrict attention to this set and often call them "states," dropping the qualifying adjectives unless necessary. A state is *terminal* if $\overline{k}_{ij} - \underline{k}_{ij} = 1$ for all (i,j); if we are at a terminal state, then we have found the element of the partition that contains the probability distribution and thus have enough knowledge to award the respondent the appropriate payoffs. Denote the set of terminal states by S_0.

As discussed in Section 4.3, we search for a policy that optimizes the worst-case scenario of the number of questions that will be asked, starting from each state. A policy from a starting state $\bar{s}$ is a map $\mathbf{q}(\bar{s})$ from states to questions such that $\mathbf{q}(\bar{s})(s) \in Q(s)$ for each s. Let $\#(\mathbf{q}, \bar{s}, p)$ denote the number of questions that must be asked to arrive at a terminal state if the probability distribution is p, the policy function is $\mathbf{q}$, and the initial state is $\bar{s}$. An optimal policy starting at a state is

$$\mathbf{q}^*(\bar{s}) \in \arg\min_{\mathbf{q}} \left[\max_p \#(\mathbf{q}, \bar{s}, p) \right]. \tag{4}$$

While we only need an optimal policy starting at the initial state, which corresponds to $p_1 \geq p_2 \geq \cdots \geq p_n$ (since we will have elicited the rank of the events from the respondent before asking the questions), it will help to find an optimal policy for all states s since the program in (4) is solvable via a dynamic programming algorithm rather than brute force. To understand the connection with dynamic programming, we essentially wish to show that a policy that is optimal at some state $\bar{s}$ remains optimal at all states s that "follow" $\bar{s}$ on path.

Consider some $\bar{s}$ with

$$\mathbf{q}^*(\bar{s}) \in \arg\min_{\mathbf{q} \in Q(\bar{s})} \left[\max_p \#(\mathbf{q}, \bar{s}, p) \right] \text{ and } \bar{p} \in \arg\max_p \#(\mathbf{q}^*(\bar{s}), \bar{s}, p).$$

For a probability distribution p and state s, the question $\mathbf{q}^*(\bar{s})(s)$ has answer $\mathbf{q}^*(\bar{s})(s)_p \in A(\mathbf{q}^*(\bar{s})(s))$. We denote the restriction of the line of questioning $\mathbf{q}^*(\bar{s})$ to states following s as $\mathbf{q}^*(\bar{s})_s$. Since we want to establish that the optimal line of questioning at state $\bar{s}$ remains optimal

at all future states, we only need to consider any state s such that $\bar{p}$ coupled with $\mathbf{q}^*(\bar{s})$ actually leads to s (else $\mathbf{q}^*(\bar{s})(s)$ is irrelevant as it plays no role in the worst case analysis associated with probability distribution $\bar{p}$). Let the worst case probability distribution at state s be $\tilde{p}$.

Now consider the path $\bar{s}, \ldots, s$ of all states reached when the line of questioning is $\mathbf{q}^*(\bar{s})$ and the underlying probability distribution is $\bar{p}$. For all states s' along this path, and for all probability distributions $p, p' \in s$ we have $\mathbf{q}^*(\bar{s})(s')_p = \mathbf{q}^*(\bar{s})(s')_{p'}$. That is, for any probability distribution consistent with state s, the answer to any question along the path will be the same. This follows because the question $\mathbf{q}^*(\bar{s})(s')$ partitions the probability distributions consistent with state s' into two subsets, one of which contains all probability distributions consistent with s. Furthermore, starting from s, $\tilde{p}$ must result in at least as many questions under the path $\mathbf{q}^*(\bar{s})_s$ as it does from $\mathbf{q}^*(s)$ (by the definition of $\mathbf{q}^*(s)$).

Thus, modifying $\mathbf{q}^*(\bar{s})$ to $\mathbf{q}^{*'}(\bar{s})$ such that it agrees with $\mathbf{q}^*(s)$ at all states following s would be a weak improvement. That is if at all states s' along the path from $\bar{s}$ to s when the line of questioning is $\mathbf{q}^*(\bar{s})$ and the underlying probability distribution is $\tilde{p}$, $\mathbf{q}^{*'}(\bar{s})(s') = \mathbf{q}^*(\bar{s})(s')$, and if at all states s' along the path from s to the terminal state s^t reached when the line of questioning is $\mathbf{q}^*(s)$ and the underlying probability distribution is $\tilde{p}$, $\mathbf{q}^{*'}(\bar{s})(s') = \mathbf{q}^*(s)(s')$, then

$$\#\left(\mathbf{q}^{*'}(\bar{s}), \bar{s}, \tilde{p}\right) \leq \#\left(\mathbf{q}^*(\bar{s}), \bar{s}, \tilde{p}\right).$$

The improvement is strict if $\mathbf{q}^*(\bar{s})_s \notin \arg\min_{\mathbf{q} \in Q(s)} \left[\max_p \#(\mathbf{q}, s, p)\right]$. Since

$$\mathbf{q}^*(\bar{s}) \in \arg\min_{\mathbf{q} \in Q(\bar{s})} \left[\max_p \#(\mathbf{q}, \bar{s}, p)\right],$$

such an improvement is impossible, and thus we have $\mathbf{q}^*(\bar{s})_s \in \arg\min_{\mathbf{q} \in Q(s)} \left[\max_p \#(\mathbf{q}, s, p)\right]$.

This proves that the optimal line of questioning at $\bar{s}$ is also optimal at all states that follow it. Moreover, since modifying $\mathbf{q}^*(\bar{s})$ to $\mathbf{q}^{*'}(\bar{s})$ such that it agrees with any optimal line of questioning starting from a future state results in the same weak improvement, we can solve for the optimal line of questioning at state $\bar{s}$ by working upwards from terminal states. Therefore, for any initial state $\bar{s}$, suppose we have a policy $\mathbf{q}^*(\bar{s})$ that chooses a question at each state. It must be that

$$\mathbf{q}^*(\bar{s})(s) \in \arg\min_{q \in Q(s)} \left[\max_p \#(q, s, p; \mathbf{q}^*)\right],$$

where $\#(q, \bar{s}, p; \mathbf{q}^*)$ is the number of questions that will be asked if q is asked at state s and $\mathbf{q}$ prescribes which question is asked at all other states. Let $V^*(s)$ be the associated minimum. Then, we are effectively finding $\mathbf{q}^*$ such that

$$\mathbf{q}^*(\bar{s})(s) \in \arg\min_{q \in Q(s)} \left[1 + \max_{s' \in A(q)} V^*(s')\right], \tag{5}$$

with V^* the associated number of questions. Note that the initial state $\bar{s}$ is not important in (5) and is simply residual notation from the formulation in (4).

To solve for the optimal policy and the value function, we use the following algorithm.

1. Enumerate all admissible and irreducible states and all possible questions from those states. To do so, we enumerate all possible states first and then eliminate all states $(\underline{k}_{ij}, \overline{k}_{ij})$ such that there does not exist p with $\sum p_i = 1$ and $\underline{k}_{ij} \leq p_i/p_j \leq \overline{k}_{ij}$. Note that in practice this simply amounts to checking the feasibility of a system of linear inequalities. We then enumerate all possible questions from each admissible state. If an admissible state s has a question q with $A(q) = \{s'\}$ (i.e., only one admissible answer), then we reduce s to s'. We proceed until all admissble states are reduced.

2. The value function for states $s \in S_0$ is $V^*(s) = 0$.

3. Find all states s such that *either* answer to *some* possible question from such states will lead to a state in S_0. This question (which may not be unique) is the policy function $q^*(s)$, and $V^*(s) = 1$. Let the set of such states be S_1 and let $T_1 \equiv S_0 \cup S_1$

4. Find all states s such that either answer to some possible question from s leads to a state in T_1. This prescribes the policy function for s and define $V^*(s) = 2$.[14] Define S_2 and $T_2 = T_1 \cup S_2$ analogously as before.

5. Iterate the previous step until all states are exhausted.

This algorithm converges in finitely many iterations and returns a policy and a value for all states, since answers to questions are necessarily "closer" to terminal states in a manner codified in Proposition 6. The algorithm solves (5) by construction.

Proposition 6. *There exists an $N < \infty$ such that $T_{N'} = S$ for all $N' \geq N$.*

A.3. Alternate Algorithms

We envision this practical applications of the mechanisms in this paper to be ones with relatively few events and somewhat coarse sets $\{q_k\}$ (or small sets of auxiliary events). As such, computing all admissible and irreducible states and then computing the optimal minimax policy function, as discussed in Appendix A.2, should be doable before the researcher conducts the field study. However, for larger sets of events, it may be unreasonable to even store this policy function for all states. In such settings, it would be useful to have an ad-hoc algorithm to determine how to implement the question-based mechanism. We provide one such algorithm below, which we can apply after eliciting the ranks of n events.

Step 1. For $i = \{1, \ldots, n-1\}$, elicit the ratio p_i/p_{i+1} to ask much precision as the set $\{q_k\}$ allows. To start for each i, we know that $q_{k^*} \leq p_i/p_{i+1} \leq q_{K+1} = \infty$, where $k^* \equiv \max\{k : q_k \leq 1\}$. We use bisection to determine $k_{i,i+1}$ so that $q_{k_{i,i+1}} \leq p_i/p_{i+1} \leq q_{k_{i,i+1}+1}$.[15]

[14] Note that we need not defer the choice of q^* and V^* for s until we classify all answers to all questions from s. We would not later discover that there exists a question that allows $V^*(s) = 1$, since if there were such a question, both states the question leads to would be in T_1.

[15] First ask to compare p_i/p_{i+1} to $q_{k'}$ where $k' = \lfloor (k^* + (K+1))/2 \rfloor$. If $p_i/p_{i+1} < q_{k'}$, then ask to compare to $q_{k''}$ where $k'' = \lfloor (k^* + k')/2 \rfloor$, and so on.

Step j. For $i = \{1, \ldots, n-j\}$, elicit the ratio p_i/p_{i+j} to as much precision as the set $\{q_k\}$ allows. Here, the bounds elicited up to Step $j-1$ will give tighter initial bounds on the ratio p_i/p_{i+j}, i.e, that $\underline{q}_{ij} \leq p_i/p_{i+j} \leq \overline{q}_{ij}$. The bounds $\underline{q}_{ij}$ and $\overline{q}_{ij}$ can be computed via Dijkstra's algorithm.[16]

Table 2 shows the mean number of questions asked as well as the mean worst-case error for 5, 10, and 15 events using the heuristic algorithm above and the $\{q_k\}$ listed in Table 1. For $n = 5$ events, relatively few questions (usually much fewer than 10) can elicit probabilities to a remarkable degree of precision (often a few hundredths per event) when using sets with $K = 3$. Using "double" $\{q_k\}$ tends to require fewer questions than using evenly spaced $\{q_k\}$, at the cost of slightly higher errors. Increasing the number of events to 10 still requires a reasonably small number of questions when the probability distributions lie near the edges of the simplex but a rather large number when considering distributions near the center of the simplex; about 40–45 questions are required on average when $\alpha = 10$. Average errors are still remarkably low—and indeed lower on average when normalized by n than when $n = 5$. When asking respondents to rank $n = 20$ events, the average number of questions is still somewhat reasonable (around 24 on average if using the "double" sets) if the distributions are once again concentrated near the edges of the simplex, with the probabilities of the different events fairly disperse.

Enlarging the set $\{q_k\}$ to $K = 7$ improves the precision, as expected. It also often does not increase the expected number of questions needed appreciably. For instance, moving from $K = 3$ to $K = 7$ when $n = 5$ and using an evenly spaced set only requires about 3–4 more questions on average, although the increase is much more substantial for other parameter values. The reason the increase is so small for certain parameters is that the initial comparisons elicited in the algorithm (between p_i and p_{i+1}, say) contain a significant amount of information about later comparisons when the grid is finer and thus allow for fewer questions in later steps. Researchers implementing this algorithm in the field can decide whether the possibly small number of additional questions (and perhaps larger increase in the complexity of the mechanism) is worth the perhaps small improvement in the precision.

B. Omitted Proofs

B.1. Proof of Lemma 1

Proof of Lemma 1(i). Suppose E_j is less likely than E_{j+1} for all j. Let the truthful order be O and denote by $\tilde{O}$ an ordering where $E_{\sigma(j)}$ is placed in the j^{th} spot, where σ is some permutation. The probability of having a payoff of at least a_j is $\sum_{j' \geq j} p_{j'}$ under O and $\sum_{j' \geq j} p_{\sigma(j')}$ under $\tilde{O}$. Of course, $\sum_{j' \geq j} p_{j'} \geq \sum_{j' \geq j} p_{\sigma(j')}$ for all j. The inequality is strict for some i as long as there exists an i' with $p_{i'} > p_{\sigma(i')}$, i.e., if the ordering is incorrect. Thus, the distribution of payoffs induced by

[16] The basic idea is that if $a \leq p_1/p_2 \leq b$ and $c \leq p_2/p_3 \leq d$, then $ac \leq p_1/p_3 \leq bd$. The bookkeeping becomes more tedious when we are asking for the tighest bounds between p_1 and p_9, say. To compute the upper bounds at Step j, consider a directed graph where the vertices are $\{1, \ldots, n\}$ and the weight on the edge from i to $i+s$ is $\overline{q}_{is}$ for $s < j$; these upper bounds have been elicited at the previous steps. (For completeness, we can say that the weight on the edge from $i+s$ to s is $1/\underline{q}_{is}$, although it does not matter if we include those edges or not.) Then, the tighest upper bound (implied only by the comparisons elicited thus far) on p_i/p_{i+j} is given by the shortest (multiplicative) path between i and $i+j$. The tighest lower bound is computed by considering the analogous graph with lower bounds.

	$\alpha = 0.1$				$\alpha = 1$				$\alpha = 10$			
	# Questions		Error		# Questions		Error		# Questions		Error	
n	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Even, $K = 3$												
5	5.4	[4, 9]	0.060	[0.009, 0.104]	9.2	[6, 10]	0.027	[0.005, 0.071]	10.0	[10, 10]	0.033	[0.008, 0.047]
10	16.5	[11, 26]	0.026	[0.005, 0.053]	39.1	[29, 45]	0.007	[0.002, 0.022]	45.0	[45, 45]	0.004	[0.002, 0.007]
20	52.8	[36, 76]	0.011	[0.003, 0.023]	159.2	[130, 184]	0.002	[0.001, 0.007]	190.0	[190, 190]	0.001	[0.001, 0.002]
Double, $K = 3$												
5	4.1	[4, 5]	0.078	[0.011, 0.130]	5.4	[4, 9]	0.031	[0.004, 0.122]	9.3	[7, 10]	0.011	[0.003, 0.026]
10	9.7	[9, 13]	0.033	[0.007, 0.064]	19.3	[12, 29]	0.007	[0.002, 0.024]	40.7	[32, 45]	0.003	[0.001, 0.005]
20	24.3	[19, 33]	0.012	[0.003, 0.028]	72.0	[51, 100]	0.002	[0.001, 0.007]	168.9	[144, 188]	0.001	[0.000, 0.001]
Even, $K = 7$												
5	8.5	[8, 12]	0.025	[0.003, 0.054]	13.7	[9, 18]	0.010	[0.002, 0.031]	13.4	[11, 14]	0.007	[0.002, 0.019]
10	22.8	[18, 33]	0.011	[0.002, 0.025]	48.5	[37, 59]	0.004	[0.001, 0.010]	47.3	[41, 54]	0.002	[0.001, 0.003]
20	67.1	[50, 91]	0.004	[0.001, 0.010]	160.0	[135, 188]	0.001	[0.000, 0.003]	172.2	[154, 196]	0.000	[0.000, 0.002]
Double, $K = 7$												
5	9.1	[8, 12]	0.024	[0.002, 0.050]	12.9	[10, 14]	0.011	[0.002, 0.034]	13.6	[12, 14]	0.009	[0.003, 0.020]
10	24.5	[19, 33]	0.011	[0.002, 0.024]	42.8	[36, 50]	0.004	[0.001, 0.012]	50.1	[43, 54]	0.003	[0.001, 0.005]
20	69.5	[54, 90]	0.005	[0.001, 0.010]	137.9	[121, 159]	0.001	[0.000, 0.004]	188.6	[166, 207]	0.001	[0.000, 0.001]

Table 2: Mean number of questions asked and mean error, along with 95% confidence intervals, for various parameters when using the heuristic algorithm described in Section A.3.

the correct ordering strictly first-order stochastically dominates the lottery induced by any incorrect ordering of these events. □

Proof of Lemma 1(ii). Suppose $E_j \subseteq E_{j+1}$ for all j. Let the truthful order be O and denote by $\tilde{O}$ an ordering where $E_{\sigma(j)}$ is placed in the j^{th} spot, where σ is some permutation. In the state of the world where E_k does not happen but E_{k+1} (and therefore also $E_{k'}$ for all $k' > k$) does, the payoff under the truthful order O is $\sum_{j'>k} a_{j'} \geq \sum_{j'>k} a_{\sigma^{-1}(j')}$ which is the payoff under the order $\tilde{O}$, where $\sigma^{-1}(j')$ is the spot that $E_{j'}$ is placed. Further, if $O \neq \tilde{O}$, then the inequality will be strict for at least one k. Therefore, the distribution of payoffs induced by the correct ordering O first-order stochastically dominates the one induced by $\tilde{O}$, and $u(O) \geq u(\tilde{O})$ with strict inequality when probabilities are misarranged in $\tilde{O}$. □

To prove Lemma 1(iii), we first prove a simple lemma.

Lemma 3. *If the events E_k and E_m (with $p_k < p_m$) are independent of all other events, then any expected utility maximizer will rank E_k before E_m.*

Proof. Let p_{11} be the probability that both E_k and E_m are successes, p_{10} be the probability that only E_k is a success, and let p_{01} and p_{00} be defined analogously. Suppose that in the order $\tilde{O}$, the agent ranks E_k in location k' and E_m in location m' and that $k' > m'$, for contradiction. We will show that he can improve his expected utility by interchanging these two events in his order (to form the order O, say). Let $\mathcal{F}$ be the set of outcomes for all events other than E_k and E_m, and let $\pi(f)$ for $f \in \mathcal{F}$ be the realized payoff from these events.[17] Then,

$$\begin{aligned} u(\tilde{O}) &= \sum_{f\in\mathcal{F}} \Pr(f)\left[p_{11}u(\pi(f)+a_{k'}+a_{m'}) + p_{10}u(\pi(f)+a_{k'}) + p_{01}u(\pi(f)+a_{m'}) + p_{00}u(\pi(f))\right] \\ &< \sum_{f\in\mathcal{F}} \Pr(f)\left[p_{11}u(\pi(f)+a_{k'}+a_{m'}) + p_{10}u(\pi(f)+a_{m'}) + p_{01}u(\pi(f)+a_{k'}) + p_{00}u(\pi(f))\right] \\ &= u(O). \end{aligned}$$

The inequality follows from the fact that

$$(p_{01} - p_{10})\left[(u(\pi(f)+a_{k'}) - u(\pi(f)+a_{m'})\right] > 0$$

since $a_{k'} > a_{m'}$ and $p_{11} + p_{10} = p_k < p_m = p_{11} + p_{01}$. We thus have that conditional on any realization of the other events, it is strictly better to rank two events that are independent of the rest in the correct order, so it is unconditionally better to rank them in the right order. □

Lemma 1(iii) follows directly from Lemma 3, taking the independent events one pair at a time.

B.2. Proof of Proposition 3

Proof. Since S is compact, we can partition it into finitely many sets each of diameter no more than ϵ. Call this partition $\mathscr{P}$. Using Proposition 1 or 2, we can use a rank-based scoring rule to elicit

[17] That is, add up the payoffs from the events that actually were realized, given the ordering.

$\mu(P)$ for all $P \in \mathscr{P}$ to within $\epsilon/|\mathscr{P}|$; let $\mathbf{p}$ be the elicted discrete distribution and let $\nu \equiv \nu(\mathbf{p}, \mathscr{P})$ be the continuous extension of this elicited distribution. Pick a measurable set $A \subseteq S$. Let $\mathscr{P}(A)$ denote the union of all sets in $\mathscr{P}$ that contain an element of A. We have

$$\mu(A) \leq \mu(\mathscr{P}(A)) = \sum_{P \in \mathscr{P}: P \subseteq \mathscr{P}(A)} \mu(P) \leq \sum_{P \in \mathscr{P}: P \subseteq \mathscr{P}(A)} \left(\nu(P) + \frac{\epsilon}{|\mathscr{P}|} \right) \leq \nu(\mathscr{P}(A)) + \epsilon.$$

But, $\mathscr{P}(A) \subseteq A^\epsilon \equiv \{x : \inf_{y \in A} \|x - y\| \leq \epsilon\}$ since each element of $\mathscr{P}$ has diameter no more than ϵ. This implies that $\mu(A) \leq \nu(A^\epsilon) + \epsilon$, so $d_P(\mu, \nu) \leq \epsilon$. □

B.3. Proof of Proposition 4

Proof. As before, let ν denote the continuous extension. Suppose for simplicity that the rank-based scoring rule elicits $\mu(P)$ exactly for all $P \in \mathscr{P}$; that is, $\mu(P) = \nu(P)$ for all $P \in \mathscr{P}$. If we cannot approximate arbitrary distributions in this case, we surely will not be able to approximate them when the rank-based scoring rule elicits $\mu(P)$ only approximately.

Fix an $\epsilon > 0$ sufficiently small. Since S is not compact, at least one element of the (finite) partition $\mathscr{P}$ must have infinite diameter. Call this set Q. Pick points x and x' in Q such that $\|x - x'\| > 2\epsilon$. Consider $\mu = \delta_x$ (i.e., the distribution that places mass 1 on x) and $\mu' = \delta_{x'}$. Both these distributions induce the same elicited distribution ν, with $\nu(Q) = 1$ and $\nu(A) = \lambda(A)/\lambda(Q)$ for $A \subseteq Q$. Since $d_P(\mu, \mu') > 2\epsilon$, it must be that either $d_P(\mu, \nu) > \epsilon$ or $d_P(\mu', \nu) > \epsilon$. Thus, at least one of μ or μ' is not approximated to within ϵ. □

B.4. Proof of Lemma 2

Proof. For a payoff a, let $j_\ell(a) \equiv \min\{j : a^\ell_j > a\}$. The probability of having a payoff at least j is $\sum_\ell \sum_{j_\ell(a)}^{J_\ell} \Pr(E^\ell_{\sigma_\ell(j)})$, where $\sigma_\ell(\cdot)$ is a permutation on $\{1, \ldots, J_\ell\}$ that gives which even the agent ranks in position j for list ℓ. As in Lemma 1(i), we have that setting σ_ℓ to be the identity permutation maximizes this probability. The distribution of payoffs induced by the correct ordering first-order stochastically dominates the lottery induced by any incorrect ordering of these events—and strictly so as long as there exist two events with strictly different probabilities. □

B.5. Proof of Proposition 5

Proof of Proposition 5(i). The proof proceeds by constructing a set $\{q_k\}$ such that bounding the ratio $\frac{p_i}{p_j}$ between consecutive elements q_k and q_{k+1} for all i, j is sufficient to infer the ranking of all events $\{E_i \cap R^i_j\}_{i,j}$. For each j, j' define $q_{j,j'} = \frac{r_j}{r_{j'}}$. Responses to a question based mechanism using $\{q_{j,j'}\}_{j \neq j'}$ will allow one to infer the direction of the inequality between $\frac{p_i}{p_{i'}}$ and $\frac{r_j}{r_{j'}}$ for all i, i', j, j' which is sufficient to deduce the relative ranking of the likelihoods of $E_i \cap R^i_j$ and $E_{i'} \cap R^{i'}_{j'}$ for any i, i', j, j'. □

Proof of Proposition 5(ii). The proof proceeds by constructing a set of auxiliary events and corresponding multiple list rank-based scoring rule such that there is a one to one correspondence between rankings in the rank-based mechanism and responses in the question based mechanism.

Construct a set of auxiliary events $\{\{R_j^i\}_{j=1}^m\}_{i=1}^n \cup \{\bar{R}_i\}_{i=1}^n$ that are independent from all the E_i and are mutually exclusive. Let the probability of R_j^i be αq_i and the probability of $\bar{R}_i$ be α, with α chosen so that the probabilities of all the events sum to 1. We define $\binom{n}{2}$ lists so that the events in list ℓ are $\{A_s^\ell\} \equiv \{E_i \cap R_j^i\}_{j=1}^m \cup \{E_{i'} \cap \bar{R}_{i'}\}$. (That is, each list is indexed by a specific pair (i, i') with $i < i'$ and contains $m+1$ events.) For each (i, i'), a response to the question-based mechanism will tell us that $p_i/p_{i'} \in [q_{k_{i,i'}}, q_{k_{i,i'}+1}]$ for some $k_{i,i'}$. This information is sufficient to rank the events in the list $\ell = (i, i')$, which in turn allows the researcher to back out the implied response in the rank-based scoring rule.

On the other hand, given any set of rankings for the $\binom{n}{2}$ lists, and any $i < i'$ we bound p_i between $p_{i'}q_j$ and $p_{i'}q'_j$ for consecutive j, j'. This is sufficient to infer a response to the question based mechanism. □

B.6. Proof of Proposition 6

To prove Proposition 6, it is useful to define the *level* of a state $s = \{(\underline{k}_{ij}, \overline{k}_{ij})\}$ to be

$$\mathcal{L}(s) \equiv \sum_{(i,j):i<j} \left(\overline{k}_{ij} - \underline{k}_{ij} - 1\right).$$

Note that terminal states have a level equal to 0, and all other states have a strictly larger level. Note also that for all states s and questions $q \in Q(s)$, it must be that the level of $s' \in A(q)$ is *strictly* less than the level of s. That is, each question gives the respondent strictly more information about the relative probabilities, and the answers to questions are "closer" to terminal states in this sense.

Proof. First note that if S_n is nonempty for all n, the cardinality of T_n would grow without bound, which is not possible since $T_n \subseteq S$, a finite set. Thus, let $N < \infty$ be the smallest value such that $S_N = \varnothing$. Then, $T_N = T_{N-1}$, and consequently $S_{N+1} = \varnothing$ as well. Therefore, it must be that $T_{N'} = T_N$ for all $N' \geq N$.

We must now show that $T_N = S$. Suppose for contradiction that $\tilde{S} \equiv S \setminus T_N$ is nonempty. Pick a state $\tilde{s}_0 \in S$. It must be that for all $q \in Q(\tilde{s})$ there is some $\tilde{s} \in A(q)$ such that $\tilde{s} \in \tilde{S}$; otherwise, $\tilde{s}_0 \in S_{N+1}$.[18] Pick any such $\tilde{s}_1$. Iteratively pick $\tilde{s}_i$ as the answer to a potential question from $\tilde{s}_{i-1}$ such that $\tilde{s}_i \in \tilde{S}$. Note that the of $\mathcal{L}(\tilde{s}_i) < \mathcal{L}(\tilde{s}_{i-1})$ for all i. Since the levels of all states are bounded below by 0, it must be that *regardless of the particular choices of states in this sequence*, $\mathcal{L}(\tilde{s}_k) = 0$ for some $k < \infty$. But then $\tilde{s}_k$ is terminal. Since terminal states are in T_N, we have a contradiction. It must be that $\tilde{S} = \varnothing$ and $T_N = S$. □

[18] Note also that $Q(\tilde{s})$ is nonempty, since otherwise $\tilde{s}$ would be terminal.