Incentive-Compatible Inference from Subjective Opinions Without Common Belief Systems

Blake Riley
University of Illinois at Urbana-Champaign
briley2@illinois.edu

No Institute Given

Abstract. Peer-prediction mechanisms elicit information about unverifiable or subjective states of the world. Existing mechanisms in the class are designed so participants maximize their expected payments when reporting honestly. However, these mechanisms do not account for participants desiring influence over how reports are used. When participants want the conclusions drawn from reports to reflect their own opinion, the inference procedure must be subjected to incentive-compatibility constraints to ensure honesty.
In this paper, I develop mechanisms without payments for discerning the true answer to a binary question, even in the presence of a false consensus. I first characterize all continuous, neutral, and anonymous mechanisms in this setting that can be implemented in interim-rationalizable strategies. Using this representation, I optimize across the class of mechanisms for accuracy in distinguishing the true state. Because the mechanism does not require knowledge of the distribution of agent types and is neutral between both outcomes, it can serve as a test for bias in the surveyed population.

1 Introduction

Polls are a standard part of modern life: students rate teacher quality, websites such as Rotten Tomatoes or Metacritic aggregate movie reviews, economists weigh in on the effects of stimulus, viewers "like" videos on YouTube, customers rate the helpfulness of tech support in online questionnaires, and so on. While some polls come down to a purely personal preference, many intend to collect subjective judgments about an objective "ground" truth. A question like "Did the American Recovery and Reinvestment Act of 2009 increase US employment by the end of 2010?" is true or false independent of the opinions of economists, but without an answer handed down from the heavens, discovering the truth of the matter depends on subjective judgments about counterfactuals, the validity of various models, and statistical methodology. Unless someone specializes in macroeconomic estimation themselves, the consensus of macroeconomists is the obvious place to turn for answers to this question.

Unfortunately, consensus might fail to reveal the truth if participants are biased. Poll results alone cannot distinguish between a consensus based on solid evidence and one fueled by other motives. Response bias or selection bias in the polling process can also skew results. If a means of discovering the truth in the presence of a false majority was available, it could defend legitimate consensuses from accusations of bias, overturn illegitimate ones, and mitigate concerns about selection bias.

The Bayesian truth serum of Prelec (2004) is one possible solution to this problem. In addition to encouraging honesty in Bayes-Nash equilibrium, the average scores assigned by the mechanism can identify the true answer even when the majority is wrong (Prelec and Seung, 2007). The Bayesian truth serum is one of many peer-prediction mechanisms developed in the last decade for eliciting judgments from strategic agents about questions without external verification. Existing peer-prediction mechanisms encourage honesty through payments, assuming participants have no stake in the conclusions drawn from a poll. However, even if no direct conflict of interest exists, participants often want the final result to reflect their sincerely held personal opinion. To account for desires for influence, the mechanism design problem must model the inference procedure used to assess reports.

Adjusting payments for agent preferences over inferences in an existing mechanism like the Bayesian truth serum is not going to be a simple task though[1]. In contrast to, say, an auction design setting where money has a clear connection to allocation preferences, it's not obvious how poll results trade off against payments. Agents with the same judgment might vary wildly in the amount they'd give up for influence. Besides posing modeling difficulties, adequate payments might be infeasible due to ethical concerns, transactions costs, or budget constraints.

To address these challenges, I investigate peer-prediction mechanisms for evaluating binary propositions without payments, assuming agents maximize the expected support given to their opinion by the mechanism. In order to serve as a test for bias in the population, a mechanism must be neutral between both positions, without built-in assumptions about the likelihoods of opinions. For robustness, mechanisms should not depend on a precise specification of how higher-order beliefs form nor presume agents have common knowledge of the belief-formation processes, reflecting the detail-free approach advocated by Wilson (1987). Implementation in dominant strategies or ex-post equilibrium are common means of avoiding dependence on common belief systems, but these concepts are too strong for this setting. As suggested by the name "peer-prediction," participants will submit predictions of the opinions of others in addition to their own opinion. Complete independence from beliefs about other participants would render the mechanism trivial. Instead, I consider implementation in interim rationalizable strategies, so that honest reporting survives successive elimination of strategies that are never interim best-responses.

I first show that to be implementable in interim-rationalizable strategies, a neutral and anonymous mechanism with real-valued, continuous output must have a specific form, up to the choice of two functions. Using this representation, I then optimize over the class of mechanisms for accuracy in predicting the true state, assuming more agents hold a particular opinion when it is correct and that, on average, agents predict the opinions of others more accurately when their own opinion is correct [2]. The resulting mechanism

[1] Boutilier (2012) provides a partial answer for how to adjust payments for scoring rules, where the true answer is eventually revealed, to compensate an agent for their own interest.

[2] A weaker condition than agents having common knowledge of the true opinion likelihoods and updating on their own opinion from an arbitrary prior.

dominates majority vote in predictive accuracy, producing more accurate predictions for all opinion likelihoods without requiring any additional input from the mechanism operator. I also consider approximately incentive-compatible mechanisms for further improvements in accuracy.

1.1 Related Literature

Peer-Prediction Mechanisms Scoring rules (Brier, 1950; Good, 1952) are a well-known means of eliciting probabilistic beliefs. Scoring rules pay agents conditional on their response and the eventual realization of the truth so that, in expectation, the agent maximizes their payment by giving their true belief. Peer-prediction mechanisms eliminate the need for the truth to be exogenously revealed as with scoring rules. Rather than conditioning payments on the actual outcome, payments depend on how agents' answers compare to one another. Once free of dependence on external verification, the principal can collect information on a substantially wider scope of questions, such as events in the distance future, counterfactual events, or vaguely defined and subjective information. Motivated by the problem of soliciting feedback about product quality online, Miller et al. (2005) develop a mechanism where agents tell the truth in Nash equilibrium, assuming the principal knows the common prior of participants. Prelec (2004)'s "Bayesian truth serum" demonstrates honesty can be generated in equilibrium even when the principal has no knowledge of the common prior or signal likelihoods, although the result holds only for a sufficiently large number of participants that depends on the unknown prior. Witkowski and Parkes (2012) construct a variant of Prelec's mechanism which is incentive compatible for finite participants in the case of binary questions.

Information Elicitation With Expert Preferences The principal-agent literature on elicitation centers around how information transmission is limited by conflicts in preferences. The classic example is Crawford and Sobel (1982), done in the case of a single expert. Groups of imperfectly informed experts have been considered by Austen-Smith (1993); Wolinsky (2002); Battaglini (2004); Gerardi et al. (2009), among others. Of these, Gerardi et al. (2009) has a setting closest in spirit to this paper. The preferences of experts are private information, transfers are not present, and the state of the world is never revealed. The mechanism induces honest reporting by randomly selecting one expert and distorting the decision in their favor if their reported signal agrees with the majority, approximately implementing almost any social choice function, in contrast to this work which optimizes for the most accurate output function.

A series of papers by Glazer and Rubinstein (2001, 2004, 2006) are among the few to derive statistical-optimal mechanisms in persuasion games, maximizing the probability the principal will be convinced of the truth. In their models, agents have fixed and known opinions and claims can be partially verified.

The political economy literature has studied information aggregation in elections and committees. Austen-Smith and Banks (1996) demonstrate the difficulties that can arise

when voters act strategically and condition their vote on being pivotal. Feddersen and Pesendorfer (1997) show how strategic behavior becomes insignificant in sufficiently large electorates. Morgan and Stocken (2008) study information transmission through polling. They find polls which do not account for strategic behavior are biased and have deception margins of error. New estimators are constructed to account for strategic behavior.

2 Design Setting

The state $\omega \in \{A, B\}$ denotes a binary property about the world, such as whether global average temperature will increase by more than four degrees over the next century, which of two potential candidates is most likely to win against the incumbent in an election, or whether Rocky Marciano would beat Muhammad Ali in a boxing match (as depicted in the fictional match *The Super Fight*).

The respondent pool contains n agents. Each individual i has an *opinion* $x_i \in \{a, b\}$ about the state and a *prediction* $p_i \in (0, 1)$ about the percentage of other respondents who hold opinion a. In a slight abuse of notation, let x_i also be an indicator variable that participant i has opinion a where convenient. Let $n_a = \sum_i x_i$ be the number of participants stating opinion a, $n_b = n - n_a$ be the number of participants stating opinion b, and $\bar{x} = n_a/n$ be the proportion of respondents with opinion a. Opinions and predictions are private knowledge.

Opinions are independent conditional on the state, with likelihoods $q_A = \Pr(x_i = a \,|\, \omega = A)$ and $q_B = \Pr(x_i = a \,|\, \omega = B)$. The likelihoods satisfy $q_A > q_B$, so opinions are correlated with the truth. Under this condition, support for a proposition tends to be larger when it's true, although those in support could be in the minority on average regardless of the state. For binary states and opinions, correlation is sufficient to ensure the Bayesian posterior belief in a state increases from the prior belief after observing the corresponding opinion.

Predictions p_i of the opinions of others are independent conditional on opinion x_i. Let F_a and F_b be the distributions of predictions for those with opinions a and b respectively. Agents' beliefs about the opinions of others are not required to be consistent with Bayesian updating on their opinion from an common or random prior. Instead, predictions satisfy a weaker condition that individuals are more accurate predictors of the opinions of others when their own opinion is correct. In other words, knowledge of the state conveys greater meta-knowledge about how many others know the state. The precise sense in which agents have better conditional meta-knowledge depends on the metric or divergence used to measure accuracy:

Definition 1 (Meta-knowledge in d). *Agents have meta-knowledge in d if the prediction distributions F_a, F_b and opinion likelihoods q_A, q_B satisfy*

$$\mathrm{E}[d(q_A, p) \,|\, p \sim F_a] < \mathrm{E}[d(q_A, p) \,|\, p \sim F_b]$$
$$\mathrm{E}[d(q_B, p) \,|\, p \sim F_b] < \mathrm{E}[d(q_B, p) \,|\, p \sim F_a]$$

The choice of metric only matters when agents significantly depart from Bayesian rationality. Suppose each agent has fixed conditional predictions p_a^i and p_b^i when agent i holds opinion a and b, respectively, such that $q_B \leq p_a^i < p_b^i \leq q_A$. This is immediately implied by agents being Bayesian and updating on their own opinion. Then the meta-knowledge assumption holds for all metrics since F_a first-order stochastically dominates F_b and each has support strictly contained in $[q_B, q_A]$. The meta-knowledge condition will only be used when evaluating mechanisms for predictive accuracy, not for guaranteeing incentive-compatibility.

Mechanisms will collect opinions and predictions from participants and output a "test statistic" $S(x, p) \in \mathbb{R}$. Let S be defined so that $S > 0$ indicates evidence in favor of A and $S < 0$ evidence in favor of B. The principal's objective is to find a test statistic which maximizes the ex-ante chance of favoring the correct state in the output of the induced mechanism. If the principal knew the true opinion likelihoods q_A and q_B, the optimal decision rule takes the form of a likelihood ratio test, as expressed by the Neyman-Pearson Lemma. In this case, the predictions are ancillary and yield no additional information about the state. However, without conditioning on the likelihoods, the meta-knowledge assumption can be used to identify the state by comparing the empirical distribution of opinions to the predictions.

Test statistics under consideration will be neutral and anonymous so the mechanism doesn't presume participants are biased in a particular direction:

Definition 2 (Neutrality). *A test statistic $S(x, p)$ is neutral between states if relabeling states A and B only changes the sign of S, i.e. $S(x, p) = -S(1 - x, 1 - p)$ for all x and p.*

Definition 3 (Anonymity). *A test statistic $S(x, p)$ is anonymous if relabeling agents does not change S, i.e. $S(x, p) = S(\sigma(x), \sigma(p))$ for all permutations σ.*

Agent's preferences are linear in S, signed in favor of their own opinion, so agents with opinion a want S to be as high as possible in expectation, while agents who hold opinion b want S is be as low as possible. This correlated private-value model allows flexibility in interpreting opinions as signals or preferences. An expert might desire to be persuasive because she sincerely thinks her opinion is correct or because that decision would benefit her personally. Whatever the precise motive, opinions are held tightly enough that the agents do not revise them upon learning the opinions of others in the pool. This seems like a plausible description of many expert judgments, where exposure to the views of others could have played a large initial role in shaping the agent's opinion, but additional exposure has negligible effect after a basic level of familiarity.

If two reports give the same expected utility for an agent, it will sometimes be convenient to break ties in ensure strict preference rather than indifference. In particular, I assume agents are *partially honest*:

Definition 4 (Partially Honest Preferences). *Agents are partially honesty if they prefer reports with their true opinion x_i when they would otherwise be indifferent.*

This condition seems natural to this setting, with agents wanting to endorse their own opinion, although this is a second-order concern to influence the mechanism output. Lexicographic preferences for honesty can substantially ease implementation requirements (Dutta and Sen, 2012; Holden et al., 2013), but will primarily allow constraints to bind exactly here, rather than carry around an additional small incentive for strict preference, and will only apply to a small number of types.

3 Rationalizably-Implementable Test Statistics

Mechanism design problems entail finding a procedure for collecting messages from agents and aggregating the reports into the desired outcome for each type profile, while respecting the incentives of each participant. A general mechanism $\mathcal{M} = (M, g)$ consists of a space of message profiles M and an outcome function $g : M \to A$, where A is the set of possible outcomes. A mechanism implements S when the outcome of the induced game under some solution concept matches S. Peer-prediction mechanisms have a message space where agents report an opinion and a probability distribution over the opinions of others.

The Bayesian truth serum (BTS) is a leading example of a peer-prediction mechanism, with truth-telling as a Bayes-Nash equilibrium for sufficiently large groups of payment maximizers. With two answers, the average difference in payments to each agent is

$$\begin{aligned} S_{\text{BTS}}(x,p) &= \ln\left(\frac{\hat{x}_{-i}}{\bar{p}_{-i}}\right) - \ln\left(\frac{1-\hat{x}_{-i}}{\overline{1-p}_{-i}}\right) \\ &\quad + \sum_{i=1}^{n}\left(\frac{x_i}{n_a} - \frac{1-x_i}{n_b}\right)\left(\bar{x}_{-i}\ln\left(\frac{p_i}{\hat{x}_{-i}}\right) + (1-\bar{x}_{-i})\ln\left(\frac{1-p_i}{1-\hat{x}_{-i}}\right)\right) \\ \text{s.t.} \quad & \hat{x}_{-i} = \Big(1 + \sum_{j\neq i} x_j\Big)/n, \quad \bar{p}_{-i} = \Big(\prod_{j\neq i} p_j\Big)^{\frac{1}{n-1}}, \\ & \overline{1-p}_{-i} = \Big(\prod_{j\neq i} 1-p_j\Big)^{\frac{1}{n-1}} \end{aligned}$$

which can distinguish the true answer asymptotically, even with in the presence of false consensus (Prelec and Seung, 2007). However, if an agent cares about this quantity directly, honesty is no longer optimal. Notice how an agent can make S_{BTS} arbitrarily low by reporting $x_i = a$ and p_i close to 1 or arbitrarily low by reporting $x_i = b$ and p_i close to zero. If payments are involved, these strategies would require the agent to make an arbitrarily large payment to the mechanism, partially offsetting the desire for influence. Without payments, BTS becomes highly manipulable.

Additionally, BTS depends on agents sharing a common prior. While common priors are often singled-out as unrealistic, a possibly more concerning feature is that agents receive a single signal with agreed-upon conditional likelihoods. Realistically, each expert has seen evidence of various levels of strength that he may or not have updated on properly. While I assume agents hold probabilistic beliefs used to calculate expected utilities, the model

should remain agnostic about why agents hold the beliefs they do and whether those are justified. Accommodating non-Bayesian agents points to implementation in rationalizable strategies, allowing agent to hold any conjectures about the types and strategic choices of others.

The robust mechanism design literature also suggests rationalizability as a relevant solution concept. Even if we were confident players have common knowledge of each other's priors, small violations of this assumption have the potential to cause large changes in mechanism outcomes. Oury and Tercieux (2012) identify implementation in Nash equilibrium on some type space and for type spaces "close" to the original as nearly coinciding with implementation in rationalizable strategies.

Since players are making predictions of the opinions of others, strategies should be rationalizable at the interim stage of the mechanism, when players know their own type, but not the types of other players. Players can have any conjectures about the types and actions of other consistent with their own p_i, including correlated conjectures. Interim (correlated) rationalizability characterizes what is possible under common certainty of rationality in incomplete information settings. For a detailed development of this solution concept, see Dekel et al. (2007). I will rely on the following definitions:

Definition 5 (Interim Rationalizability). *Given a mechanism $\mathcal{M} = (M, g)$, strategy m_i is interim rationalizable for agent i of type (x_i, p_i) if m_i survives the iterated deletion of strictly interim dominated strategies, where beliefs about the types and strategy choices of other agents can be correlated. Let the set of all interim rationalizable strategies for player i of type (x_i, p_i) be $B_i^{\mathcal{M}}(x_i, p_i)$.*

Definition 6 (Interim Rationalizable Implementation). *Mechanism $\mathcal{M} = (M, g)$ implements S in interim rationalizable strategies if every profile of interim rationalizable strategies m for type profile (x, p) satisfies $g(m) = S(x, p)$.*

As shown in the following theorem, all neutral, anonymous, and continuous test statistics S implementable in interim rationalizable strategies for given n have a specific functional form, up to functions κ and ξ on the unit interval. In this characterization, the *base score* κ represents the support for A based solely on the proportion of agents endorsing it. The base score is adjusted by differences of *prediction scores* for each agent, signed according to their opinion. The base score will need to have sufficient slope so reports with a false opinion are interim dominated for each agent. Conditioning on each player always wanting to honestly reveal their true opinions, agents will want to give their true prediction as long as their marginal influence is a proper scoring rule for the proportion of a endorsements. The function ξ weights prediction scores, controlling the magnitudes of rewards and punishments for prediction accuracy in each region of the unit interval.

With this representation, the entire class of rationalizably-implementable mechanisms can be optimized over with minimal constraints. The differences in κ should be just large enough to offset players' incentives to misreport, so attention can be restricted to the case when the differences equal f. I will refer to test statistics of this form as *net score statistics*.

Theorem 1. *A continuous, neutral, and anonymous test statistic S for n participants is implementable in interim rationalizable strategies only if S can be represented as*

$$S(x,p) = \kappa(\bar{x}) + \sum_{i:\, x_i=a} \int_0^{p_i} (\bar{x}_{-a} - t)\xi(t)\, dt - \sum_{i:\, x_i=b} \int_0^{1-p_i} (1 - \bar{x}_{-b} - t)\xi(t)\, dt \qquad (1)$$
$$s.t. \quad \bar{x} = n_a/n, \quad \bar{x}_{-a} = (n_a - 1)/(n-1), \quad \bar{x}_{-b} = n_a/(n-1).$$

for Lebesgue-measurable $\xi : [0,1] \to \mathbb{R}_+$ and $\kappa : [0,1] \to \mathbb{R}$ such that

1. *κ is negatively symmetric around $1/2$, i.e. $\kappa(1/2+\epsilon) = -\kappa(1/2-\epsilon)$ for all ϵ and*
2. *there exists $z_1, z_2 \in [0,1]$ such that $\forall m \in \{0, \ldots, \lceil n/2 - 1 \rceil\}$,*

$$\kappa\left(\frac{m+1}{n}\right) - \kappa\left(\frac{m}{n}\right) \geq f\left(\frac{m}{n}\right) \qquad \textit{where}$$

$$f\left(\frac{m}{n}\right) = \max\left\{ -\int_0^{z_1} \left(\frac{m}{n-1} - t\right)\xi(t)\, dt \right.$$
$$\left. - \int_0^{1} \left(\frac{m}{n-1} - t\right)\xi(t)\, dt - \int_0^{z_2} \left(\frac{n-1-m}{n-1} - t\right)\xi(t)\, dt \right\}.$$

This representation is sufficient for full implementation if the bound in condition (2) holds with strict inequality or agents are partially honest.

4 Approximately Incentive-Compatible Test Statistics

While rationalizable implementation dictates summing the prediction scores of each participant, identifying the state according to the meta-knowledge assumption suggests comparing the average prediction quality of each side. Consider the following modification of the net score statistic, which I will refer to as *average score statistics*:

$$S(x,p) = \kappa(\bar{x}) + \frac{1}{n_a} \sum_{i:\, x_i=a} \int_0^{p_i} (\bar{x}_{-a} - t)\xi(t)\, dt - \frac{1}{n_b} \sum_{i:\, x_i=b} \int_0^{1-p_i} (1 - \bar{x}_{-b} - t)\xi(t)\, dt \qquad (2)$$

Rather than taking the total difference of prediction scores, the scores inside each group are averaged together, allowing more direct comparison of the predictive accuracy of each group according to the metaknowledge assumption. Because the effect of an agent's prediction is larger when fewer people agree with her, giving an honest prediction is no longer a best response when others are honest about their opinions. However, honesty can be approximately rationalizable as the number of agents increases, given conditions stated in the following theorem.

The theorem yields a class of nearly unconstrained, approximately incentive-compatible mechanisms that can now be optimized for accuracy over functions ξ, with potential gains

over the optimal net score statistic due to the slight relaxation of incentive constraints. More general forms of κ than stated are sufficient for an average score statistic to be approximately incentive-compatible, but these turn out to be unnecessary in the later optimization. This does not exhaust the class of approximately optimal alternatives, but provides a basic robustness check.

Theorem 2. *In the n-player revelation game induced by an average score statistic, if agents are partially honest and believe opinions are independent conditional on the state, ξ is non-negative and Lebesgue-measurable, and κ has the form*

$$\begin{aligned}\kappa\left(\frac{n_a}{n}\right) = \operatorname{sign}\left(\frac{n}{2} - n_a\right) \sum_{m=\min\{n_a,n_b\}+1}^{\lceil n/2-1/2\rceil} \frac{1}{m}\int_0^1 \left(\frac{m-1}{n-1} - t\right)\xi(t)\,dt \\ + \operatorname{sign}\left(\frac{n}{2} - n_a\right)\frac{\mathbb{1}(n \text{ odd})}{n+1}\int_0^1 \left(\frac{1}{2} - t\right)\xi(t)\,dt\end{aligned} \tag{3}$$

then the direct mechanism has the following properties:

1. *All interim rationalizable reports contain an agent's true opinion x_i.*
2. *Honest predictions p_i differ from some rationalizable prediction p_i^* by $O\left(n^{-1}\right)$. If there is a unique rationalizable prediction, it is not equal to the honest prediction, understating the proportion of agents with the same opinion.*
3. *All profiles of interim rationalizable reports produce the same outcome, which differs from the honest outcome by $O\left(n^{-1}\right)$.*

5 Simulation Setting

Using the classifications developed in the previous sections, I now consider which test statistics maximize the probability of correctly identifying the true state. Since analytical solutions to this objective are not immediately forthcoming, I evaluate potential test statistics numerically. Two benchmark test statistics are the proportion of a endorsers and the likelihood ratio of the distribution of opinions assuming the true likelihoods were known. Performance of a test statistic under this objective can described by the net percentage of correct classifications made relative to majority vote, normalized by the net percentage of correct decisions the likelihood ratio makes over majority vote. This criterion forms a quality index from 0 to 100 of the percentage of false consensuses identified, where 0 is equivalent to majority vote and 100 is optimal if the likelihoods were known.

In these simulations, each state is equally likely. Opinion likelihoods are drawn uniformly from the unit square, subject to $q_A > q_B$. Based on this specification, majority opinion fails to match the state at least 25% of the time, occurring when $\omega = A$ and $q_A < 0.5$ or when $\omega = B$ and $q_B > 0.5$.

Agent predictions are modeled as normally distributed on a log-odds scale:

$$\begin{aligned} &\ln\left(\frac{p_i}{1-p_i}\right) \sim \text{Normal}(\mu_{x_i}, \sigma^2) \quad \text{s.t.} \\ &\mu = \alpha \ln\left(\frac{q_A}{1-q_A}\right) + (1-\alpha)\ln\left(\frac{q_B}{1-q_B}\right) \\ &\mu_a = \mu + \epsilon, \quad \mu_b = \mu - \epsilon, \quad \alpha, \epsilon \sim \text{Unif}[0,1] \end{aligned} \tag{4}$$

To satisfy the metaknowledge assumption that agents are more accurate in expectation when their opinion matches the state, the centers of the distributions are determined by two uniform variates α and ϵ, which respectively have a rough interpretation as the prior belief that $\omega = A$ and the amount of evidence participants consider their own opinion to be. Note that α and ϵ are constant across agents in a given simulation. Across simulations, I set $\sigma^2 = 1$, which produces dispersed distributions, without bunching around 0 and 1 when transformed into probabilities, as occurs when the variance grows larger. Figure 1 shows typical prediction distributions.

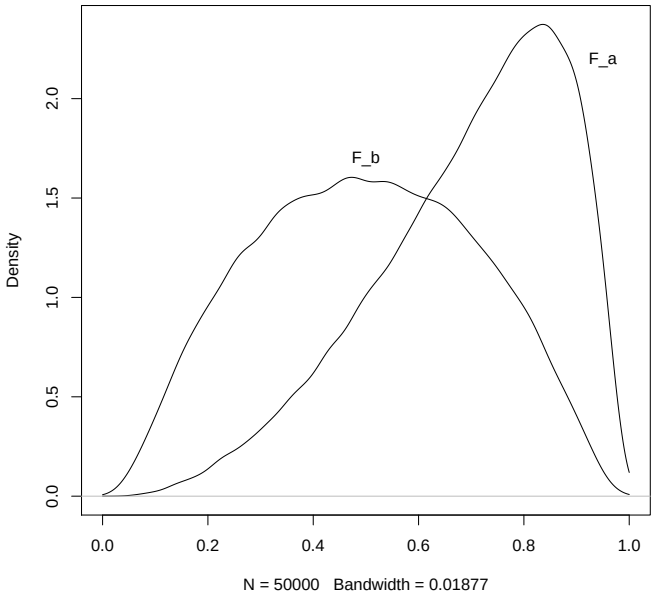

Fig. 1: Simulated prediction distributions for agents with opinions a and b when $q_A = 0.8$, $q_B = 0.4$, $\sigma^2 = 1$, $\alpha = 0.5$, and $\epsilon = 0.5$

6 Optimization over piecewise linear $\xi(t)$

Numerical optimization in this setting entails a search over all non-negative functions on $[0,1]$ as well as z_1 and z_2 for Net Score Statistics. I approach this problem by restricting $\xi(t)$ to be piecewise linear and continuous, with segments at regular intervals. Candidate functions with $h-1$ segments are represented as vectors of length $h+1$, stating the value of the function at $0, 1/h, 2/h, \ldots, (h-1)/h, 1$. Rescaling $\xi(t)$ by a positive constant does not change the objective function, so without loss of generality, the vectors are constrained to $[0,1]^h$.

The optimization is done through controlled random search with local mutation (Kaelo and Ali, 2006; Price, 1983). This global optimization procedure operates like a combination of evolutionary optimization and the Nelder-Mead method. At each step, the worst point from a pool of candidates is compared against a trial point, being replaced when there is improvement. The trial point is generated as the reflection of a simplex formed by the best candidate and h other randomly selected points from the pool. In the local mutation variant, if the trial point fails to improve on the worst candidate of the pool, a second trial point is generated by including the first trial point in a new simplex and reflecting it about the best candidate. This procedure reliably outperformed other considered algorithms such as stochastic hill-climbing, simulated annealing, and Nelder-Mead.

Figures 2 and 3 depict piecewise-linear $\xi(t)$ optimized for $n = 25$ participants for varying h, along with the associated quality index, for net score statistics and average score statistics. The non-monotonicity in quality as h increases is slightly concerning, likely due to the constraint of equidistant knots and the increased difficulty of searching a higher-dimensional space. Both classes can correctly identify about one-third of false majorities, with very modest improvements going from exact to approximate incentive compatibility.

7 Test Statistics from Interval Scoring Rules

Although optimization over piecewise linear functions provides a rough picture of the optimal $\xi(t)$, the resulting test statistics are somewhat complicated. An easily integrable $\xi(t)$ would yield a simpler test statistic, hopefully without much cost to efficiency. The results of the previous section suggest $\xi(t) = \mathbb{1}(k_1 < t < k_2)$ for some $k \in [0,1]^2$ for an average score statistic. Since ξ represents a weighting of predictions, use of this ξ says predictions are increasingly relevant when $p > k_1$, although all predictions $p > k_2$ are treated equally. This yields

$$S(x,p) = \operatorname{sign}(n/2 - n_a)\left(\frac{\mathbb{1}(n \text{ odd})}{n+1} R\left(\frac{n}{2}, k_2\right) + \sum_{j=\min\{n_a,n_b\}+1}^{\lceil n/2-1/2\rceil} \frac{1}{j} R\left(\frac{m-1}{n-1}, k_2\right)\right)$$
$$+ \frac{1}{n_a} \sum_{i:\, x_i = a} R\left(\frac{n_a - 1}{n-1}, p_i\right) - \frac{1}{n_b} \sum_{i:\, x_i = b} R\left(\frac{n_b - 1}{n-1}, 1 - p_i\right)$$

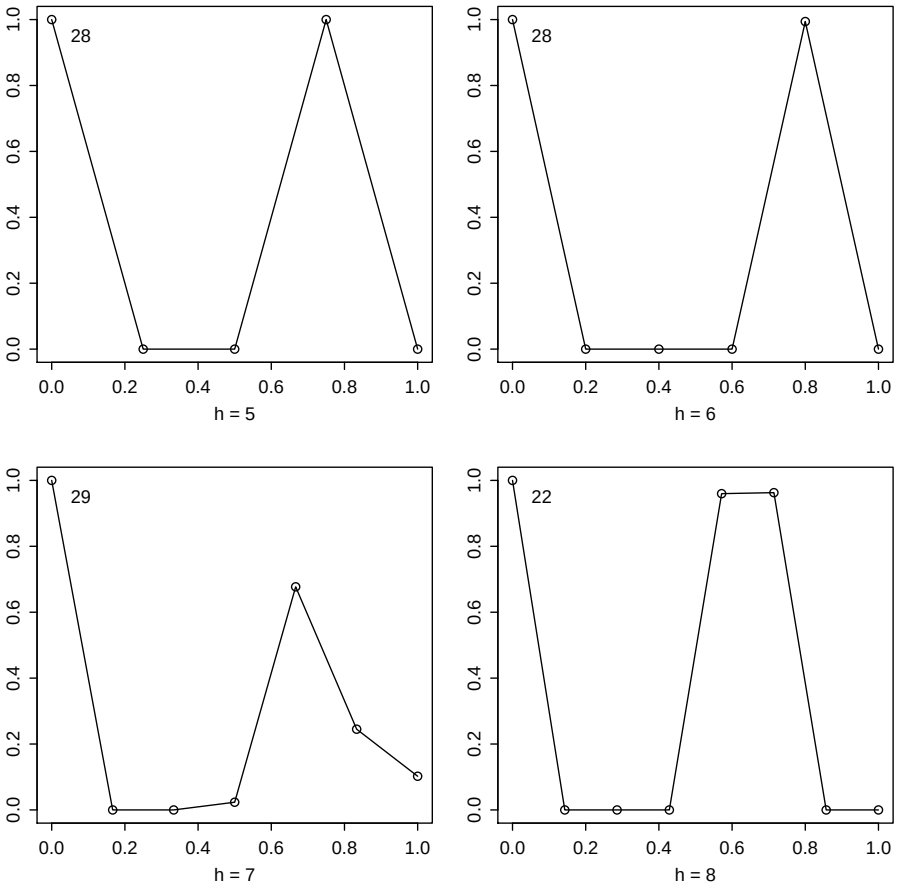

Fig. 2: Piecewise-linear $\xi(t)$ for net score statistics produced by numerical optimization over 1000 steps, evaluated at 10^4 simulations with for 25 participants, with the resulting quality index shown in the upper left.

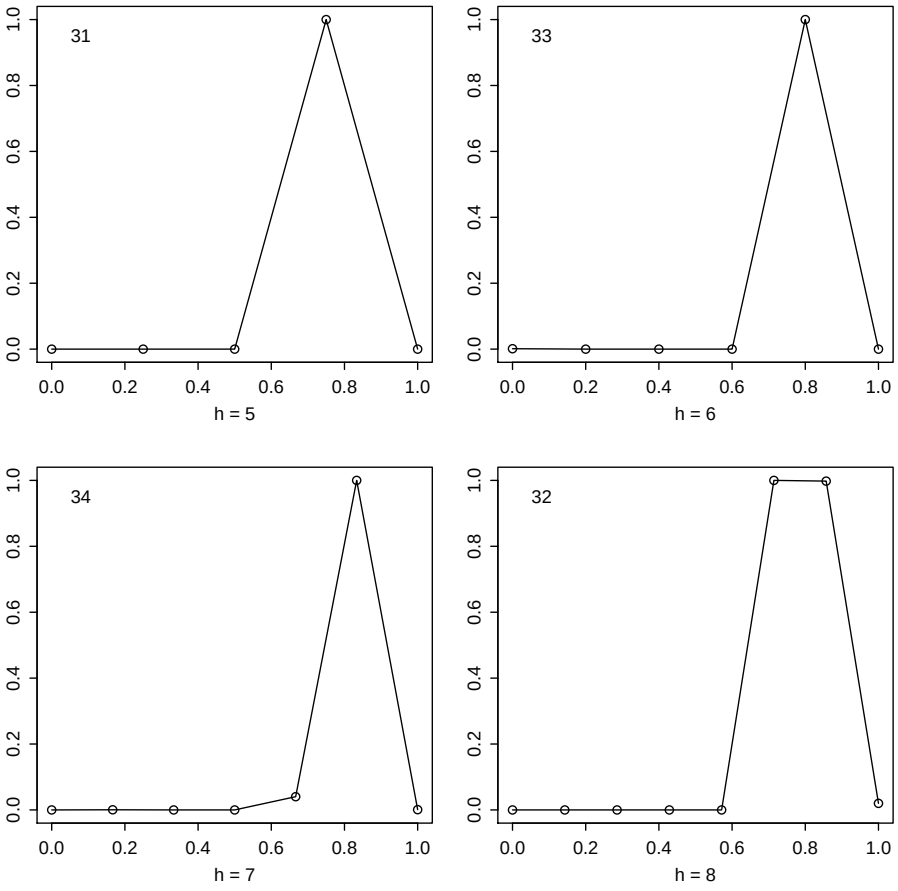

Fig. 3: Piecewise-linear $\xi(t)$ for average score statistics produced by numerical optimization over 1000 steps, evaluated at 10^4 simulations with for 25 participants, with the resulting quality index shown in the upper left.

based on individual prediction scores

$$R(p,q) = \mathbb{1}(p > k_1) \int_{k_1}^{\min\{p,k_2\}} q - t \, dt$$
$$= \mathbb{1}(k_1 < p < k_2) \left(q(p-k_1) - \frac{p^2 - k_1^2}{2} \right) + \mathbb{1}(k_2 \leq p) \left(q(k_2 - k_1) - \frac{k_2^2 - k_1^2}{2} \right).$$

For $k_1 \backsimeq 0.7$ and $k_2 \backsimeq 0.83$, these rules actually outperform the linear spline rules found in the previous section, identifying approximately 35% of false majorities. Figure 4 shows the proportion of correct decisions made by a an interval test statistic when $k_1 = 0.7$ and $k_2 = 0.83$ compared to the likelihood ratio and the Bayesian truth serum. The Bayesian truth serum clearly performs better than the interval test statistic, although the Bayesian truth serum is not guaranteed to solicit honest opinions in this setting, even when transfers can be made.

Figure 5 shows the performance the interval test statistic, the net score statistic, and the Bayesian truth serum at varying levels of bias. In all shown cases, those endorsing A are most likely in the majority even in the B state, so a false consensus exists about 50% of the time. This show the difference between the interval score statistic and the Bayesian truth serum occurs primarily at high levels of bias. With a large proportion of a supporters, the base score of the interval test statistic dominates the difference in prediction scores, reducing it down to majority vote. However, as long as the degree of bias is moderate, the interval test statistic is comparable in predictive power to the Bayesian truth serum.

8 Conclusion

While these numerical results are not definite, they are suggestive of what is possible with such weak assumptions. Test statistics based on indicator scoring rules improve on majority vote, particularly at low levels of bias, without presuming knowledge of the likelihoods or participant bias. Participants are not required to be probabilistically sophisticated, much less share a common prior.

Future directions include comparing this paper with experimental and analytical results. Possible extensions include generalizing from binary to finite states or allowing agent preferences to have varying intensity.

9 Computational Details

This document was typeset in LaTeXvia the **knitr** package (Xie, 2012) in R 2.15.2 on Windows 7. The **nloptr** package (Johnson, 2012) implemented the controlled random search algorithm used in Section 6.

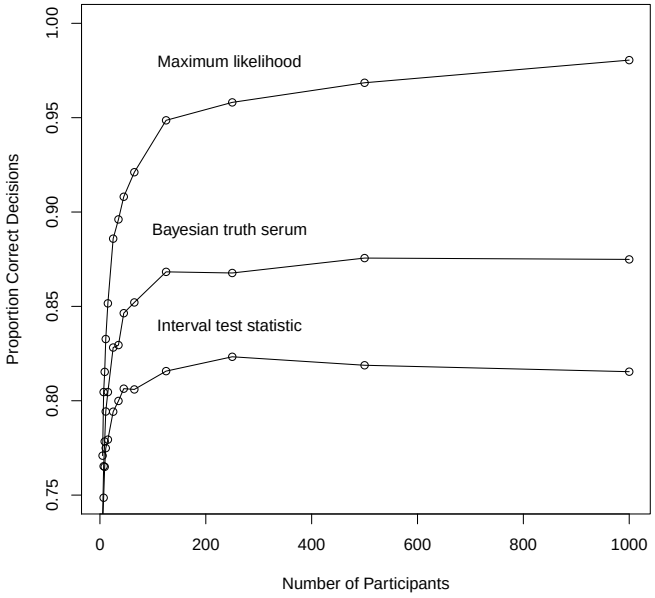

Fig. 4: Proportion of correct decisions made by a interval test statistic when $k = (0.7, 0.83)$ vs maximum likelihood and the Bayesian truth serum across 10^4 simulations for each pool size.

Appendix

Proof (). **Necessity of Proposition 1.** This characterization follows from interim incentive compatibility, which is itself necessary by well-known arguments underlying the revelation principle. Here, since agents' utilities are simply S signed in favor of their opinion, the relevant version of incentive compatibility is

$$\begin{aligned}\sum_{x_{-i},p_{-i}} \pi(x_{-i},p_{-i}) S((a,x_{-i}),(p_i,p_{-i})) &\geq \sum_{x_{-i},p_{-i}} \pi(x_{-i},p_{-i}) S((x_i',x_{-i}),(p_i',p_{-i})) \\ &\geq \sum_{x_{-i},p_{-i}} \pi(x_{-i},p_{-i}) S((b,x_{-i}),(p_i,p_{-i}))\end{aligned} \tag{5}$$

for all x_i', p_i, p_i', and beliefs π such that

$$\mathrm{E}_\pi[\bar{x}_{-i}] = \sum_{x_{-i},p_{-i}} \pi(x_{-i},p_{-i}) \frac{\#(x_j = a \,|\, j \in -i)}{n-1} = p_i$$

to be consistent with prediction p_i.

To establish incentive compatibility, suppose mechanism $\mathcal{M} = (M, g)$ implements S in interim-rationalizable strategies. Let $B^{\mathcal{M}}(x_i, p_i)$ be the interim-rationalizable strategies for

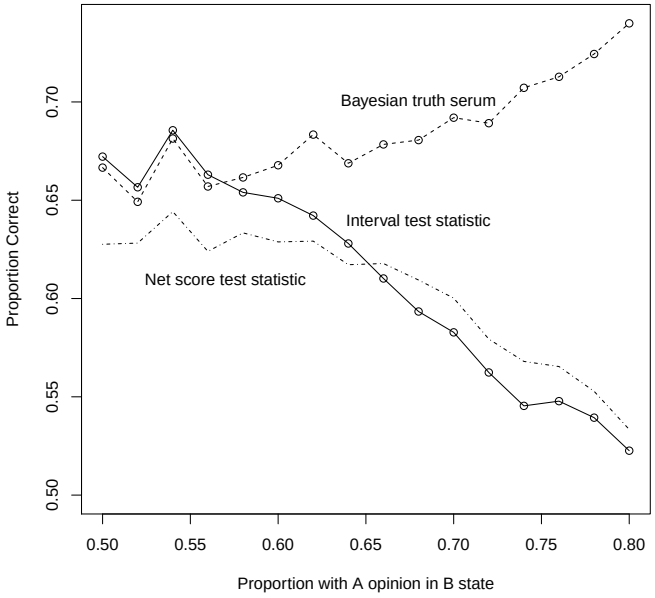

Fig. 5: Proportion of correct decisions for an interval test statistic with $k = (0.7, 0.83)$, the optimal net score statistic with a linear-spline ξ for $h = 5$, and the Bayesian truth serum across in simulations for fixed q_B and $q_A = q_B + 0.1$.

type (x_i, p_i). Given any $m_i \in B^{\mathcal{M}}(a, p_i)$ and $m_i' \in B^{\mathcal{M}}(x_i', p_i')$, we must have

$$\begin{aligned}\sum_{x_{-i},p_{-i}} \pi(x_{-i}, p_{-i})\, S((a, x_{-i}), (p_i, p_{-i})) &= \sum_{x_{-i},p_{-i}} \pi(x_{-i}, p_{-i}) \sum_{m_{-i}} \phi(m_{-i} \,|\, x_{-i}, p_{-i})\, g(m_i, m_{-i}) \\ &\geq \sum_{x_{-i},p_{-i}} \pi(x_{-i}, p_{-i}) \sum_{m_{-i}} \phi(m_{-i} \,|\, x_{-i}, p_{-i})\, g(m_i', m_{-i}) \\ &= \sum_{x_{-i},p_{-i}} \pi(x_{-i}, p_{-i})\, S((x_i', x_{-i}), (p_i', p_{-i}))\end{aligned}$$

for m_i to be a best response when agent i is type (a, p_i) with beliefs π and ϕ such that

$$\begin{gathered}\mathrm{E}_\pi[\bar{x}_{-i}] = p_i \quad \text{and} \\ \phi(m_{-i} \,|\, x_{-i}, p_{-i}) > 0 \implies m_{-i} \in \prod_{j \in -i} B^{\mathcal{M}}(x_j, p_j).\end{gathered}$$

This follows similarly for types (x_i', p_i') and (b, p_i), yielding line (5).

Suppose agent i believes p_{-i} is fixed conditional on x_{-i}, reducing beliefs over the types of others to $\pi(x_{-i})$. Incentive compatibility implies

$$\sum_{x_{-i}} \pi(x_{-i})\, S((a, x_{-i}), (p_i, p_{-i})) \geq \sum_{x_{-i}} \pi(x_{-i})\, S((a, x_{-i}), (p_i', p_{-i}))$$

for all p_i, p_i', p_{-i}, and π such that $\mathrm{E}_\pi[\bar{x}_{-i}] = p_i$, so that agent i does not want to misreport her prediction p_i. Notice that this condition says S is a proper scoring rule for the mean of x_{-i} from the perspective of agent i, holding $x_i = a$ fixed. By the Schervish (1989) and Lambert (2011) representations of continuous scoring rules, S must be representable from the perspective of agent i as

$$S((a, x_{-i}), p) = \kappa_i(x, p_{-i}) + \int_0^{p_i} (\bar{x}_{-i} - t)\xi_i(t, p_{-i})\, dt$$

for some Lebesgue-measurable $\xi : [0,1] \times [0,1]^{(n-1)} \to \mathbb{R}_+$. This representation prescribes the specific way that p_i and the proportion $\bar{x}_{-i}$ must interact, up to a weighting by ξ. For S to be neutral between A and B, we must have

$$S((b, x_{-i}), p) = \kappa_i(x, p_{-i}) - \int_0^{1-p_i} (1 - \bar{x}_{-i} - t)\xi_i(t, 1 - p_{-i})\, dt$$

so $S(x, p) = -S(1 - y, 1 - q)$. With this form for each agent , it follows by anonymity that

$$S(x, p) = \kappa(\bar{x}) + \sum_{i:\, x_i = a} \int_0^{p_i} (\bar{x}_{-i} - t)\xi(t)\, dt - \sum_{i:\, x_i = b} \int_0^{1-p_i} (1 - \bar{x}_{-i} - t)\xi(t)\, dt,$$

since $\bar{x}$ contains all information preserved under permutations of x and ξ can't depend on the identity of the agent. Although ξ_i could have depended on the predictions of other agents to be a proper scoring rule for agent i, those predictions can only appear in their respective integrals to be proper for the remaining agents. The negative-symmetry of κ around $1/2$ then follows from neutrality.

Incentive compatibility also implies[3] S is higher in expectation when agent i reports her true type (a, p_i) than when reporting (b, p_i'), i.e.

$$\begin{aligned}
&\sum_{n_a=0}^{n-1} \pi(n_a) \Bigg(\kappa\left(\frac{n_a + 1}{n}\right) + \int_0^{p_i} \left(\frac{n_a}{n-1} - t\right) \xi(t)\, dt \\
&\qquad + \sum_{j:\, x_j = a} \int_0^{p_j} \left(\frac{n_a}{n-1} - t\right) \xi(t)\, dt - \sum_{j:\, x_j = b} \int_0^{1-p_j} \left(\frac{n - 2 - n_a}{n-1} - t\right) \xi(t)\, dt \Bigg) \\
&\geq \sum_{n_a=0}^{n-1} \pi(n_a) \Bigg(\kappa\left(\frac{n_a}{n}\right) - \int_0^{1-p'} \left(\frac{n - 1 - n_a}{n-1} - t\right) \xi(t)\, dt \\
&\qquad + \sum_{j:\, x_j = a} \int_0^{p_j} \left(\frac{n_a - 1}{n-1} - t\right) \xi(t)\, dt - \sum_{j:\, x_j = b} \int_0^{1-p_j} \left(\frac{n - 1 - n_a}{n-1} - t\right) \xi_i(t)\, dt \Bigg) \qquad (6) \\
\iff &\sum_{n_a=0}^{n-1} \pi(n_a) \left(\kappa\left(\frac{n_a + 1}{n}\right) - \kappa\left(\frac{n_a}{n}\right) + \sum_{j:\, x_j = a} \int_0^{p_j} \frac{1}{n-1} \xi(t)\, dt + \sum_{j:\, x_j = b} \int_0^{1-p_j} \frac{1}{n-1} \xi(t)\, dt \right) \\
&\geq -\int_0^{p_i} (p_i - t)\, \xi(t)\, dt - \int_0^{1-p'} (1 - p_i - t)\, \xi(t)\, dt \qquad (7)
\end{aligned}$$

[3] Again in the special case of a degenerate distribution on p_{-i} conditional on x_{-i}.

for all $p_i, p_i', p_j(x_{-i})$, and beliefs π[4] such that $\mathrm{E}_\pi[n_a/(n-1)] = p_i$. Taking $p_j = 0$ if $x_j = a$ and $p_j = 1$ if $x_j = b$ implies

$$\sum_{n_a=0}^{n-1} \pi(n_a) \left(\kappa\left(\frac{n_a+1}{n}\right) - \kappa\left(\frac{n_a}{n}\right) \right) \geq -\int_0^{p_i} (p_i - t)\,\xi(t)\,dt - \int_0^{1-p'} (1 - p_i - t)\,\xi(t)\,dt,$$

i.e. the expectation of κ's first differences in n_a must be greater than a function of the mean of the distribution. This is equivalent to the differences at any given n_a being bounded away from the right-hand side by some convex function. Since the right-hand side is quasi-convex in p' (non-increasing at $p' < p_i$ and non-decreasing at $p' > p_i$), the inequality is satisfied for all p' if and only if it holds for $p' \in \{0, 1\}$.

Following a similar argument for agents with opinion b yields the differences in κ being bounded away by a convex function from two more quantities (for $p'' = 0$ and $p'' = 1$) at each p_i. Since the four quantities are concave in n_a, each can be bounded above by a supporting line at points in $[0, 1]$. The constraints for agents of different opinion types are mirrored around $1/2$, so it suffices to choose two supporting points z_1 and z_2. The pointwise maximum of the lines f is then a convex function bounding each relevant quantity, completing the proof of necessity.

Proof (). **Sufficiency of Proposition 1.** The sufficiency of this representation follows from iterated deletion of interim dominated strategies in the direct mechanism. Consider an agent of type (a, p_i) who conjectures the average proportion of reported opinions is $\hat{p}_i$. By the conditions on κ, a report of $(a, \hat{p}_i)$ weakly dominates all reports (b, p'). A comparison of lines (6) and (7) above shows the agent will strictly prefer $(a, \hat{p}_i)$ to (b, p') as long as the agent thinks there is some chance that p_j and $1 - p_j$ (when $x_j = a$ and $x_j = b$, respectively) are outside a neighborhood of zero where $\xi(t)$ is uniformly zero. Otherwise, a strict bound on the differences in κ or partial honesty is necessary to guarantee strict dominance of all (b, p'). An analogous argument for agents of type (b, p_i) rules out all (a, p'). Since each agent strictly prefers submitting their true opinion, it follows that each agent weakly prefers submitting their true prediction of the opinions of other agents since S is a proper scoring rule for each agent. Weak dominance on this step is sufficient because indifference occurs only when S is constant, with ξ uniformly zero in some interval containing those reports.

Proof (). **Proof of Proposition 2.** Since the mechanism is neutral between A and B, all properties can be analyzed from the perspective of an agent with opinion $x_i = a$ without loss of generality.

1. *All interim rationalizable reports contain an agent's true opinion x_i.*
   Setting up the incentive-compatibility constraint similarly to the necessity proof of Theorem 1 yields

$$\kappa\left(\frac{m+1}{n}\right) - \kappa\left(\frac{m}{n}\right) = \max\left\{ -\frac{1}{m+1}\int_0^1 \left(\frac{m-1}{n-1} - t\right)\xi(t)\,dt \right.$$
$$\left. -\frac{1}{n-m}\int_0^1 \left(\frac{n-m-1}{n-1} - t\right)\xi(t)\,dt \right\}$$

[4] Without loss of generality, treated as a distribution on $n_a = \sum_{j \in -i} x_j$ rather than on x_{-i} directly. Although agents can hold asymmetric beliefs about their peers, this information is irrelevant since the mechanism is anonymous.

as a sufficient condition for reports with false opinions to be weakly dominated. The κ given on line 3 simply adds up these successive differences, with adjustments to be negatively symmetric. Partial honesty guarantees reports with false opinions are strictly interim dominated and hence not rationalizable.

2. *Honest predictions p_i differ from some rationalizable prediction p_i^* by $O\left(n^{-1}\right)$. If there is a unique rationalizable prediction, it is not equal to the honest prediction, understating the proportion of agents with the same opinion.*
Since agents always report their true opinion, rationalizable reports potentially differ only in the prediction. Conditioning on everyone reporting their opinion honestly, an agent with $x_i = a$ chooses his reported prediction to maximize

$$\mathrm{E}\left[\frac{1}{n_a}\int_0^p \left(\frac{n_a - 1}{n-1} - t\right)\xi(t)\,dt \;\middle|\; n_a - 1 \sim \mathrm{Bin}(n-1, p_i)\right],$$

the only term of the outcome the agent's prediction affects.
Although possibly non-differentiable due to discontinuities in ξ, the expression is quasi-concave and continuous in p on a compact domain, and hence has a well-defined, connected set of maximizers. Continuity is straightforward. To establish quasi-concavity, note that the expression's sub- and super-derivatives are bounded by the limit points of the derivative where it exists by continuity. Applying Leibniz's rule to the expression where valid yields

$$\frac{1}{n-1}\left(1 - \mathrm{E}\left[n_a^{-1}\right]\right)\xi(p) - p\;\mathrm{E}\left[n_a^{-1}\right]\xi(p),$$

which is non-negative below and non-positive above the quantity

$$p_i^* = \frac{1}{n-1}\left(\mathrm{E}\left[n_a^{-1}\right]^{-1} - 1\right),$$

so the expression is monotonic to the left and right of p_i^*, and hence quasi-concave. When ξ is positive in a neighborhood of p_i^*, this is the unique maximum. Otherwise, the set of maximizers is the largest connected subset of the closure of $\xi^{-1}(0)$ that contains p_i^*, since p_i^* is a maximizer, but the expression is flat over this region.
The above argument establishes p_i^* as an interim rationalizable prediction, depending on the inverse moment of agent's expectation of n_a. By Jensen's inequality,

$$p_i^* = \frac{1}{n-1}\left(\mathrm{E}\left[n_a^{-1}\right]^{-1} - 1\right) < \frac{1}{n-1}(\mathrm{E}[n_a] - 1) = \frac{1}{n-1}(p_i(n-1) + 1 - 1) = p_i,$$

so the agent understates the proportion of others he expects to share the a opinion. In fact, p_i^* can be computed exactly as

$$p_i^* = \frac{p_i}{1 - (1-p_i)^n}\frac{n}{n-1} - \frac{1}{n-1}$$

using the fact (Garcia and Palacios, 2001) that

$$\mathrm{E}\left[\frac{1}{1+X} \;\middle|\; X \sim \mathrm{Bin}(n,p)\right] = \frac{1 - (1-p)^{n+1}}{p(n+1)}.$$

Then, the distance between honesty and some rationalizable prediction is bounded above by

$$p_i - p_i^* = p_i \left(1 - \frac{1}{1-(1-p_i)^n} \frac{n}{n-1}\right) + \frac{1}{n-1} > 0,$$

which converges to zero at rate $O\big(n^{-1}\big)$.

3. *All profiles of interim rationalizable reports produce the same outcome, which converges to the honest outcome at rate* $O\big(n^{-1}\big)$.
   Two distinct reports are interim rationalizable only if the predictions are both in some neighborhood where ξ is zero. Then both reports give the same ex-post utility since the integral affected by the agent's prediction is unchanging.
   Agents' utilities are simply the final outcome signed in their favor and agent's predictions don't interact, so the overall outcome is identical for all profiles of interim rationalizable reports.
   Since each agent's honest prediction becomes arbitrarily close to a rationalizable prediction and the outcome is a continuous combination of the predictions, the rationalizable outcome also converges to the honest outcome as n increases. The total discrepancy between the rationalizable and honest outcomes is

$$\frac{1}{n_a} \sum_{i:x_i=a} \int_{p_i^*}^{p_i} \left(\frac{n_a - 1}{n-1} - t\right) \xi(t)\, dt - \frac{1}{n_b} \sum_{i:x_i=b} \int_{1-p_i}^{1-p_i^*} \left(\frac{n_b - 1}{n-1} - t\right) \xi(t)\, dt.$$

   For asymptotically constant n_a/n, each integral term is of order $O\big(n^{-1}\big)$, so the difference of the averages is also $O\big(n^{-1}\big)$.

Bibliography

Austen-Smith, D. 1993. Interested experts and policy advice: Multiple referrals under open rule. *Games and Economic Behavior 5,* 1, 3–43.

Austen-Smith, D. and Banks, J. 1996. Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review 90,* 1, 34–45.

Battaglini, M. 2004. Policy advice with imperfectly informed experts. *Advances in Theoretical Economics 4,* 1.

Boutilier, C. 2012. Eliciting forecasts from self-interested experts: Scoring rules for decision makers. In *Proc. of the 11th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS-12)*. 737–744.

Brier, G. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78,* 1, 1–3.

Crawford, V. P. and Sobel, J. 1982. Strategic information transmission. *Econometrica 50,* 6, 1431–1451.

Dekel, E., Fudenberg, D., and Morris, S. 2007. Interim correlated rationalizability. *Theoretical Economics 2,* 1, 15–40.

Dutta, B. and Sen, A. 2012. Nash implementation with partially honest individuals. *Games and Economic Behavior 74,* 1, 154 – 169.

Feddersen, T. and Pesendorfer, W. 1997. Voting behavior and information aggregation in elections with private information. *Econometrica 65,* 5, 1029–1058.

Garcia, N. L. and Palacios, J. L. 2001. On inverse moments of nonnegative random variables. *Statistics and Probability Letters 53*, 235–239.

Gerardi, D., McLean, R., and Postlewaite, A. 2009. Aggregation of expert opinions. *Games and Economic Behavior 65,* 2, 339–371.

Glazer, J. and Rubinstein, A. 2001. Debates and decisions: On a rationale of argumentation rules. *Games and Economic Behavior 36,* 2, 158–173.

Glazer, J. and Rubinstein, A. 2004. On optimal rules of persuasion. *Econometrica 72,* 6, 1715–1736.

Glazer, J. and Rubinstein, A. 2006. A study in the pragmatics of persuasion: A game theoretical approach. *Theoretical Economics 1,* 4, 395–410.

Good, I. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological) 14,* 1, 107–114.

Holden, R., Kartik, N., and Tercieux, O. 2013. Simple mechanisms and preferences for honesty.

Johnson, S. G. 2012. *The NLopt nonlinear-optimization package, version 2.3.*

Kaelo, P. and Ali, M. 2006. Some variants of the controlled random search algorithm for global optimization. *Journal of optimization theory and applications 130,* 2, 253–264.

Lambert, N. S. 2011. Elicitation and evaluation of statistical forecasts.

Miller, N., Resnick, P., and Zeckhauser, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science 51,* 9, 1359–1373.

Morgan, J. and Stocken, P. C. 2008. Information aggregation in polls. *American Economic Review 93,* 3, 864–896.

Oury, M. and Tercieux, O. 2012. Continuous implementation. *Econometrica 80,* 4, 1605–1637.

Prelec, D. 2004. A Bayesian truth serum for subjective data. *Science 306,* October 15, 462–466.

Prelec, D. and Seung, H. S. 2007. An algorithm that finds truth even if most people are wrong.

Price, W. 1983. Global optimization by controlled random search. *Journal of Optimization Theory and Applications 40,* 3, 333–348.

Schervish, M. J. 1989. A general method for comparing probability assessors. *The Annals of Statistics 17,* 4, 1856–1879.

Wilson, R. 1987. Game-theoretic analyses of trading processes. In *Advances in Economic Theory: Fifth World Congress*, T. Bewley, Ed. Cambridge University Press, Cambridge, Chapter 2, 33–77.

Witkowski, J. and Parkes, D. C. 2012. A robust Bayesian truth serum for small populations. In *Proc. of the 26th AAAI Conf. on Artificial Intelligence (AAAI 2012).*

Wolinsky, A. 2002. Eliciting information from multiple experts. *Games and Economic Behavior 41,* 1, 141–160.

Xie, Y. 2012. *knitr: A general-purpose package for dynamic report generation in R.*