Information Elicitation *Sans* Verification

Bo Waggoner, Harvard University
Yiling Chen, Harvard University

The recent advent of human computation — employing groups of non-experts to solve problems — has motivated study of a question in mechanism design: How do we elicit useful information when we are unable to verify reports? Existing methods, such as peer prediction and Bayesian truth serum, require assumptions either on the mechanism's knowledge about the participants or on the information structure of participants for eliciting private information from the participants. Meanwhile, however, there are simple mechanisms in practice such as the ESP Game that seem to require no such assumptions, yet have achieved great empirical success. We attack this paradox from two directions. First, we provide a broad formalization of the problem of information elicitation without verification and show that, without assumptions on designer knowledge or participants' information, there do not exist mechanisms that can truthfully elicit the private information of the participants for this setting. Second, we define and analyze the output agreement class of mechanisms, an extremely broad but simple mechanism in which players are rewarded based on the metric distance between their reports. Output agreement makes no assumptions on designer knowledge or participants' information and thus cannot be "truthful". We resolve the paradox by showing that it is *useful*: It elicits the correct answer according to the *common knowledge* among the players. We conclude with an analysis of the assumptions and results of various popular mechanisms for information elicitation without verification.

Categories and Subject Descriptors: F.0 [**Theory of Computation**]: General; J.4 [**Social and Behavioral Sciences**]: Economics

General Terms: Theory, Economics

Additional Key Words and Phrases: Human computation, information elicitation, mechanism design, output-agreement mechanisms

ACM Reference Format:
Bo Waggoner and Yiling Chen, 2013. Information Elicitation *Sans* Verification. *ACM* 9, 4, Article 39 (June 2013), 18 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

The emerging field of human computation has harnessed the intelligence of an unprecedentedly large population of people for the purpose of solving computational tasks. For example, in the now-classic ESP game [von Ahn and Dabbish 2004], which has collected semantic labels for over one hundred million images[1], the image labeling task is turned into a fun, online game: Two players are simultaneously shown an image and asked to independently type words related to the image; whenever a word is typed by both players, they score some points and move on to the next image.

[1] http://news.bbc.co.uk/2/hi/technology/7395751.stm

Authors' address:
B. Waggoner and Y. Chen, School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, MA.

DOI:http://dx.doi.org/10.1145/0000000.0000000

The ESP game illustrates many of the common challenges in human computation settings. First, the computational primitives, people, are *self-interested*. Players of the ESP game are interested in scoring points and winning the game, and may behave strategically. Second, the participants often do *not* have *preferences for outcomes of the computation.* Whether the ESP game obtains descriptive labels for images does not necessarily affect the utility of its players. Third, the quality of contributions often is *not verifiable*. The answers may be subjective or too costly to be practical to verify, especially at large scales. These qualities characterize many other human computation settings, including crowdsourcing marketplaces such as Amazon Mechanical Turk and CrowdFlower, where requesters commonly "hire" human workers to solve problems that are difficult for computers.

Because such systems involve *self-interested agents*, they naturally motivate a question in mechanism design: What types of information may be elicited in such systems, and how? Yet the problem lacks two key features that are leveraged to align incentives in many mechanism design settings: the case where participants have *preferences over the outcome*, as in auctions or facility location problems; and the case where the designer may *objectively verify* some information related to participants' inputs, as with proper scoring rules and prediction markets.

In this paper, we formalize the *information elicitation without verification* setting to capture mechanism design for problems that do not have these two characteristics. In this problem, a designer is interested in obtaining useful answers to a query from a group of participants who have private information or knowledge about it; however, the designer cannot verify the answers provided. In the past decade, many mechanisms in this category have been proposed, including the peer-prediction method [Miller et al. 2005] and Bayesian truth serum (BTS) [Prelec 2004]. The goal of these approaches is to strictly incentivize participants to report *truthfully*. Here, for a given query and set of information available to the participant, truthfulness specifies a particular "best" answer that fits the mechanism designer's objectives. (Usually, this is just the "signal" or information observed by the participant.) Such approaches reward a participant based on not only her report but also reports of other participants and can, under certain conditions, achieve a truthful equilibrium.

However, these mechanisms have strong limitations. For example, peer prediction requires the designer to know the information structure of all participants, while the equilibrium of BTS depends on restrictions on the information structure of participants[2]; this has inspired extensions [Witkowski and Parkes 2012a,b] that attempt to relax these assumptions. However, despite such efforts and a broad variety of other results in the literature, all proposed mechanisms for information elicitation without verification require some such assumptions for their results to hold.

Curiously, this contrasts with practical experience in which simple, direct methods such as the ESP Game seem to make no such explicit restrictions, yet still elicit useful data. In fact, their ability to extract the "wisdom of the crowd" has been in many cases a remarkable success despite their use of mechanisms with little theoretical justification. This observation seems to be in tension with the theoretical literature, where mechanisms tend to be more complex and require restricted information structures or specialized designer knowledge to achieve positive results.

This paper. Here, we address this conundrum on both fronts. We begin by showing that the assumptions made in the literature are not frivolous: Without restrictions on the mechanism's knowledge about the participants or on the nature of the participants' information, there is *no* truthful mechanism for eliciting information without verifi-

[2] We provide a description and comparison of these methods in Section 6.

cation. This resolves the search for an equally powerful elicitation mechanism with fully relaxed assumptions. We also show that, even with such assumptions, (almost) all mechanisms have bad equilibria in which the designer learns nothing from the player's actions. These impossibility results can be extended to cover all mechanisms in the literature for this setting.

This raises the question of how to reconcile these impossibility results with the success of simple, practical mechanisms such as the ESP Game. We provide an answer by formally defining and analyzing the *output agreement* class of mechanisms, inspired by von Ahn and Dabbish [2008]. These mechanisms ask players to compute and report some query, compare their responses in some way (*e.g.*, a distance metric), and pay players based on how closely their answers "agree". This covers the ESP Game, for example, but also captures a large class of games beyond those originally noted by von Ahn and Dabbish. For example, requesters on Mechanical Turk may pay a participant more if their answer to a query matches the answer of another participant (or of the majority of participants). More complex scenarios are also possible: Ask participants to report probability distributions over a random variable, paying them based on the relative entropy (KL-divergence) of their reports; ask participants to transcribe an audio clip of speech to text, paying them based on the Hamming or Levenshtein distance between responses; and so on.

Output agreement mechanisms do not require the mechanism designer to have any knowledge about participants and make no assumptions on the player information structure. Therefore, by our impossibility results, they are not "truthful", but we show that they are nevertheless *useful*: They elicit the correct answer to a query according to *common knowledge* among participants. This result is quite positive in a human computation setting, where it seems natural to ask questions to which most or all participants would be expected to know the answer.

Output agreement is simple, has been in use in crowdsourcing for a decade, and includes some of the most prominent examples of human computation mechanisms. However, this work is the first to provide a general game-theoretic analysis and formal explanation of its success. One reason for this may be that the result requires generalizing the traditional notion of truthfulness to capture and quantify situations in which agents respond according to some, rather than all, of their knowledge. We formalize this criterion as *specificity*. Mechanism design often considers mechanisms that compute *approximately* optimal solutions, yet the literature focuses on a binary criterion of truthfulness when evaluating agent reports. This work reveals the fundamental limitations imposed on elicitation problems by truthfulness; it also demonstrates the positive results obtainable with a more nuanced approach.

Contributions. This work makes the following conceptual contributions. First, it provides the first formal definition and general analysis of the problem of information elicitation without verification. This includes broad impossibility results whose proofs illuminate structural features of the setting, including the key difficulty of eliciting expert knowledge. Second, it generalizes the criterion of truthfulness of player reports to capture the *specificity* or amount of information according to which a player reports; it then applies this criterion in a space where truthfulness is impossible but revelation of useful information is nevertheless desirable and achievable. To our knowledge, this work is the first in mechanism design to make such a distinction on player reports. Third, it formalizes and analyzes the broad, popular class of output agreement mechanisms. The results both provide novel insights and crystallize empirical intuition by demonstrating that output agreement elicits *common knowledge*.

The work makes two technical contributions. First, we prove broad impossibility theorems. Primarily, we show that no mechanism can elicit truthful responses with strict incentives unless it has knowledge of the information structure or is restricted to particular types of player information. Second, we provide an analysis of the output agreement class of mechanisms. In addition to the primary equilibrium analysis, we examine player inference processes and note extensions such as introducing a trusted player for equilibrium selection and various mechanisms on many players.

The discussion includes an analysis of various mechanisms in the literature; the comparison of assumptions and results provides context to the results and shape to the general setting of information elicitation without verification.

1.1. Related Work

We are broadly concerned with mechanisms for eliciting information. Thus, our work is somewhat related to recent work on information elicitation in the presence of ground truth [Lambert et al. 2008; Lambert and Shoham 2009; Lambert 2011] as well as on the design of human computation mechanisms when the quality of contributions *is* verifiable, *e.g.* using an auction or contest [*e.g.* Jain et al. 2009; Chawla et al. 2012].

Prior work in information elicitation without verification includes notably the Peer Prediction method [Miller et al. 2005] (without a common prior [Witkowski and Parkes 2012a]) and the (robust [Witkowski and Parkes 2012b]) Bayesian truth serum [Prelec 2004]; these are most closely related to output agreement, and we give an overview and comparison of these mechanisms in Section 6 and more details in Appendix C. The literature also includes a series of work by Jurca and Faltings; often, the focus is on using *automated mechanism design* for objectives such as budget minimization [Jurca and Faltings 2006], collusion-resistance [Jurca and Faltings 2007a, 2009]; and robustness to outside preferences [Jurca and Faltings 2007b]. Jurca and Faltings [2008] provides an online (dynamic) mechanism that is not individually truthful but yields an accurate aggregate with many agents. Other approaches focus on observations drawn i.i.d. from an unknown distribution in $\mathbb{R}$ [Lambert and Shoham 2008; Goel et al. 2009]; we overview these mechanisms in Appendix C.

Our work examines the theoretical properties of output agreement games. This term was introduced by von Ahn and Dabbish [2008], with a primary example being the ESP Game [von Ahn and Dabbish 2004]. Such games have been investigated experimentally [Weber et al. 2008; Huang and Fu 2012].

But to our knowledge, there has been no theoretical analysis of the general output agreement setting. Jain and Parkes [2008] consider the special case of the ESP Game, but introduce a very ESP-Game-specific model: Players first select an effort level; based on this level, nature randomly samples a list of words from a dictionary to produce a report. This introduces specific restrictions on player strategies (indeed, our results show that they must make some such restriction to achieve strict incentives). In contrast, the output agreement class defined here covers a far broader setting than image labeling (for example, reports are points in an arbitrary metric space). Further, we do not make any such assumptions on player strategies; our focus is on examining the nature of play by unrestricted rational agents.

2. A GENERAL MODEL OF INFORMATION ELICITATION WITHOUT VERIFICATION

Here, we formally define mechanisms for information elicitation without verification. In this setting, there is a set of players, each holding some private information. A mechanism designer interested in some computational task queries each player separately (*i.e.*, without communication between players). The designer selects an outcome of the mechanism and assigns monetary transfers to each agent. Thus the mechanism, when applied to particular players, induces a Bayesian simultaneous-move game.

We adopt the general *states of the world* model, which has been widely used in economics for modeling private information [Aumann 1976; McKelvey and Page 1986;

Nielsen et al. 1990; Ostrovsky 2009]. There is a finite set of possible states of the world Ω, shared by all players. An *event* is a subset of Ω; for example, the event $Q \subseteq \Omega$ could be "it is raining outside" and would consist of every state of the world in which it is raining. Nature selects a true state of the world $\omega^* \in \Omega$; an event Q is said to *occur* if $\omega^* \in Q$. Thus, the true state of the world implicitly specifies all events that occur or do not: whether it is raining, whether Alice speaks French, whether $\mathsf{P} = \mathsf{NP}, \ldots$.

A player's knowledge is specified by a *prior distribution* $\mathcal{P}[\omega]$ on Ω along with a partition Π_i of Ω. A *partition* of a set Ω is a set of nonempty subsets of Ω such that every element of Ω is contained in exactly one subset. For example, both $\{\{\omega_1\}, \{\omega_2, \omega_3\}\}$, and $\{\{\omega_1, \omega_2\}, \{\omega_3\}\}$ are partitions of $\{\omega_1, \omega_2, \omega_3\}$. When the true state of the world is ω^*, each player i learns the element of their partition that contains ω^*, denoted $\Pi_i(\omega^*)$. Informally, i is unsure which state in $\Pi_i(\omega^*)$ is the true state of the world; more precisely, i updates to a *posterior distribution* $\Pr[\omega \mid \Pi_i(\omega^*)] = \Pr[\{\omega\} \cap \Pi_i(\omega^*)] / \Pr[\Pi_i(\omega^*)]$. In line with literature on information elicitation, $\Pi_i(\omega^*)$ will be referred to as i's *signal*. (In mechanism design terms, it is player i's type.)

Throughout, we let the the set of states Ω and the number of players $n \geq 2$ be fixed.

A particular set of n players is modeled by an *information structure* $\mathcal{I} = (\mathcal{P}[\omega], \Pi_1, \ldots, \Pi_n)$, where each Π_i is a partition for player i and all players share the prior $\mathcal{P}[\omega]$. $\mathcal{I}$ is common knowledge; this is the standard Bayesian game setting. We use $\mathtt{I}$ to denote the set of valid information structures on Ω with n players.

A *mechanism* M for information elicitation without verification contains the following components:

- A set A_i of actions for each player i. We generally consider an action $a_i \in A_i$ to be a report or answer to the mechanism's query.
- A set of outcomes O. An outcome $o \in O$ can be any result for the computational task, such as some aggregation of player's reports or a vector consisting of all reports.
- A choice function $f^M : \mathtt{I} \times A_1 \times \cdots \times A_n \to O$ that specifies the outcome of the mechanism on a given information structure and set of player actions.
- For each player i, a reward function $h_i^M : \mathtt{I} \times A_1 \times \cdots \times A_n \to \mathbb{R}$. When the information structure is $\mathcal{I}$ and each player j reports a_j, player i's *utility* is defined to be equal to $h_i^M(\mathcal{I}, a_1, \ldots, a_n)$.

For simplicity, we will let the action and outcome spaces be implicitly specified by the functions and refer to a mechanism as $M = (f^M, h_1^M, \ldots, h_n^M)$.

Given these definitions, we define a *game* for information elicitation without verification to be a pair $G = (M, \mathcal{I})$ consisting of a mechanism and an information structure. Given a mechanism M, we refer to the set $\{G = (M, \mathcal{I}) : \mathcal{I} \in \mathtt{I}\}$ as the games *induced by* M or the different *settings* of M. A mechanism *elicits* a property (under certain conditions) if that property holds for every game induced by that mechanism (subject to those conditions); *e.g.*, "Bayesian truth serum elicits a truthful equilibrium when signals are conditionally independent."

Strategies and equilibria. In a game $G = (M, \mathcal{I})$, a *strategy* for player i is a mapping s_i that specifies, for each element $\Pi_i(\omega)$ of i's partition, a probability distribution on A_i. In state ω^*, i learns element $\Pi_i(\omega^*)$ of his partition and draws an action $a_i \sim s_i(\Pi_i(\omega^*))$. A strategy profile $(s_1, \ldots, s_n)$ is a *Bayes-Nash Equilibrium* (or just *equilibrium*) of the game G if every player's strategy s_i is a *best response* to s_{-i} (the profile with s_i omitted): For every state of the world ω^* and for all i, the probability distribution $s_i(\Pi_i(\omega^*))$ on A_i is an optimal solution to

$$\max_{s_i'(\Pi_i(\omega^*))} \sum_{\omega \in \Pi_i(\omega^*)} \Pr[\omega \mid \Pi_i(\omega^*)] \, \mathbb{E}\left[h_i^M(\mathcal{I}, s_1(\Pi_1(\omega)), \ldots, s_i'(\Pi_i(\omega^*)), \ldots, s_n(\Pi_n(\omega)))\right],$$

where the inner expectation is taken over the actions a_j drawn from each $s_j(\Pi_j(\omega))$, $j \neq i$, and a_i drawn from $s_i'(\Pi_i(\omega^*))$. The strategy profile $(s_1, \dots, s_n)$ is a *strict* equilibrium if every s_i is the unique best response to s_{-i}.

Comparison with general mechanism design setting. The above definitions encode the two requirements for a setting of information elicitation without verification. First, players do not have preferences over outcomes: Their utilities do not depend on the outcome o of the mechanism, but only on their reward. Second, ground truth is not available: The mechanism's choice function f^M and reward functions h_i^M do not depend on the true state of the world, but only on the actions played and (possibly) the information structure. Because f^M does not affect incentives, we will not specify the particular choice of f^M in the rest of the paper; however, it serves a vital purpose as the goal of such mechanisms is to compute or produce a useful output.

Modeling knowledge. Following the seminal work of Aumann [1976], we have modeled players' private information using partitions of the state space. Another common approach to modeling private information is the "signals" model in which nature selects some hidden event and there is a common prior over the joint distribution of players' signals conditional on the event. This model is used in peer prediction, for example. The two models are known to be equivalent in that each can model any scenario described by the other. (For completeness, this is proved in Appendix A.)

However, the partitions-of-the-state-space model allows a more intuitive understanding of players' *common knowledge*. This makes both the impossibility results in Section 3 and the analysis of output agreement mechanisms in Section 5 far simpler and more straightforward. Given partitions $\{\Pi_1, \dots, \Pi_n\}$, the *common-knowledge partition* Π is defined to be the meet of these partitions. The *meet* of a set of partitions of Ω is the finest partition of Ω that is coarser than each individual partition. Partition Ψ is *coarser* than partition Γ (equivalently, Γ is *finer* than Ψ) if each element of Ψ can be written as a union of elements of Γ. (*i.e.*, each element of Ψ is partitioned by a subset of Γ.) In this case, Ψ may also be called a *coarsening* of Γ.

Informally, an element $\Pi(\omega)$ of the common-knowledge partition is a set of states (an event) such that, when the state of the world ω^* is in $\Pi(\omega)$, all players know that this is the case, and all know that the others know that this is the case, and so on *ad infinitum*. For two players with partitions $\{\{\omega_1\}, \{\omega_2, \omega_3\}\}$ and $\{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$ for example, the common knowledge partition Π is $\{\{\omega_1\}, \{\omega_2, \omega_3\}\}$. If the second player has partition $\{\{\omega_1, \omega_2\}, \{\omega_3\}\}$, then Π is just $\{\{\omega_1, \omega_2, \omega_3\}\}$.

To model information that is elicited, we may associate a mechanism M with a *query* $T = (T_1, \dots, T_n)$, where each T_i is a function mapping a distribution on Ω to a set of probability distributions on A_i. Intuitively, T_i will specify the "correct" or "truthful" report for player i for a given posterior belief. For example, this could be to report the posterior distribution itself, or the expected value of some random variable, or the set of states on which the posterior has positive probability (that is, i's signal). More generally, it might specify multiple valid reports or valid distributions over reports; this covers cases where a player may report either correct answer in case of a "tie", is asked to compute a function to within a certain error bound, or is asked to compute a randomized function, *e.g.* a query on a differentially private database.

3. IMPOSSIBILITY RESULTS

3.1. The prevalence of uninformative equilibria

In information elicitation without verification, the goal is to design mechanisms with "good" equilibria in which information is revealed. However, it has previously been noted informally and observed for individual mechanisms or special cases [Lambert and Shoham 2008; Jurca and Faltings 2005; Della Penna and Reid 2012] that such mechanisms often also have equilibria that are “bad” in some way. The conjecture that this holds more generally may be considered something of a suspected folk theorem in the literature. The following characterization formalizes this intuition in a very broad setting and for very “bad” equilibria: those in which absolutely no information is revealed. The characterization uses a game of complete information induced by “stripping away” the information structure from a Bayesian game.

Definition 3.1. A strategy s_i for player i is *uninformative* if for all ω, ω', $s_i(\Pi_i(\omega)) = s_i(\Pi_i(\omega'))$. An equilibrium $(s_1, \dots, s_n)$ is *uninformative* if s_i is uninformative for all i.

Definition 3.2. For any Bayesian game $G = (M, \mathcal{I})$ for information elicitation without verification, let G' denote the induced simultaneous-move game of complete information where each player i selects and reports an action $a_i \in A_i$ and receives a payoff of $h_i^M(\mathcal{I}, a_1, \dots, a_n)$. A strategy in G' is a probability distribution over actions; a profile of best response strategies is a Nash equilibrium.

THEOREM 3.3. *A game G of information elicitation without verification has an uninformative equilibrium if and only if there exists a Nash equilibrium in G'.*

PROOF. We show a one-to-one correspondence between the two. First, we note that strategy sets in G' are vectors of probability distributions $(p_1, \dots, p_n)$ from which players draw their actions. Second, we note that uninformative strategy sets in G are determined uniquely by a vector of distributions $(p_1, \dots, p_n)$, because for each i and for all $\omega, \omega' \in \Omega$, $s_i(\Pi_i(\omega)) = s_i(\Pi_i(\omega')) = p_i$. Therefore, there is a one-to-one correspondence between strategy sets in G' and uninformative strategy sets in G. But each player i's reward for a realized profile of actions $(a_1, \dots, a_n)$ is identical in G' and in G (by construction of G'). So when each player j draws an action from p_j, drawing actions from to p_i maximizes i's expected utility in G' if and only if it does so in G. This completes the proof. □

COROLLARY 3.4. *The following mechanisms for information elicitation always have uninformative equilibria:*

(*1*) *Those where each A_i is finite.*
(*2*) *Those where each A_i is compact in some metric space and each h_i^M is continuous.*
(*3*) *Those for which (a) an equilibrium always exists and (b) each h_i^M does not depend on the information structure.*

PROOF. (1) and (2) follow because in each case G' satisfies the sufficient conditions for existence of a Nash equilibria in (respectively) Nash [1951] and Glicksberg [1951]. For (3), consider such a mechanism M, *i.e.*, for any profile of reports $(a_1, \dots, a_n)$ and for all $\mathcal{I}, \hat{\mathcal{I}}$, we have that each $h_i^M(\mathcal{I}, a_1, \dots, a_n) = h_i^M(\hat{\mathcal{I}}, a_1, \dots, a_n)$. Let $\mathcal{I}$ have $\Pi_i = \{\Omega\}$ for all i; let $G = (M, \mathcal{I})$. It is immediate that, since G has an equilibrium, G' has an equilibrium. But G' is identical to $\hat{G}'$ for any $\hat{G} = (M, \hat{\mathcal{I}})$, since rewards do not depend on $\mathcal{I}$. □

No mechanism in the literature avoids this issue; an interesting problem for future work might be to identify a useful mechanism and associated setting that do (carefully avoiding the conditions of *e.g.* Corollary 3.4). In the meantime, two common solutions are to endeavor to make “good” equilibria focal in some sense (for instance, the ESP Game randomizes player matchings so that the image is the players' only coordinating device), and to introduce some small amount of objective verification, such as evaluating a small proportion of answers. A method that combines both approaches is to introduce trustworthy players who always play a "good" strategy [Jurca and Faltings 2005]. We briefly discuss this approach for output agreement mechanisms in Section 5.2.

3.2. Assumptions required for truthfulness

We identify two desirable properties of a mechanism for information elicitation without verification. First, it has a "truthful" equilibrium in all of its induced games; restriction to certain types of information structures is not required. Second, the choice function and payoff functions do not depend on the information structure, but only on the actions taken; the designer is not required to know the information structure of the participants.

Peer prediction satisfies the first criterion, but violates the second (requiring full knowledge by the mechanism designer). Many mechanisms in the literature have been constructed with an explicit goal being to relax the second assumption; for instance, Bayesian truth serum. However, all such mechanisms violate the first criterion. For example, they assume (as in BTS) that players' signals are conditionally independent.

Here, we show that such violations were unavoidable: no strictly truthful mechanism can satisfy both criteria.[3] (It is of note that non-strict incentives are achieved by simply asking all players to report truthfully and paying them a constant amount.)

To prove this result, we require a formal definition of "truthful" strategies for general mechanisms.[4] This is captured by a query T. Given T, we say that player i's strategy s_i is *truthful* if, for every $\omega^* \in \Omega$, $s_i(\Pi_i(\omega^*)) \in T_i(\Pr[\omega \mid \Pi_i(\omega^*)])$. An equilibrium $(s_1, \dots, s_n)$ is *truthful* if every s_i is truthful. A query T is considered trivial if, for every setting of M, every player has a probability distribution on actions that is truthful for every signal she receives. We require T to be non-trivial.

A primary goal is to make truthful equilibria *strict*; this ensures that making the truthful report is the unique best response. But when a query allows for multiple truthful reports (or multiple truthful distributions over reports) for a given signal, a designer might be interested in a somewhat weaker notion than strict equilibrium: Any of these truthful reports (or distributions) may give equal utility, *i.e.*, be a best response, as long as non-truthful ones give strictly less. However, our proof shows that even this weaker goal is still not achievable without assumptions. Formally, a truthful equilibrium $(s_1, \dots, s_n)$ is *strongly truthful* if, for each i, if s_i' is a best response to s_{-i}, then s_i' is truthful. Every strict, truthful equilibrium is strongly truthful, and when the query T specifies singleton sets, the criteria are equivalent.

THEOREM 3.5. *Fix T and let $M = (f^M, h_1^M, \dots, h_n^M)$ be a mechanism for information elicitation without verification. If each h_i^M does not depend on the information structure, then M does not elicit a strongly truthful equilibrium.*

PROOF. Formally, suppose that, for any $(a_1, \dots, a_n)$, $h_i^M(\mathcal{I}, a_1, \dots, a_n) = h_i^M(\mathcal{I}', a_1, \dots, a_n)$ for all $\mathcal{I}$ and $\mathcal{I}'$. The approach will be to consider a player whose truthful response depends on which signal she receives, but who is much better informed than her opponents. We will then construct a simple game where her best responses are the same for different signals she receives. This will imply that in some state, she has a non-truthful best response.

Consider, by nontriviality, a $\mathcal{P}[\omega]$, player i, and partition Π_i such that, for any probability distribution p_i on A_i, there is some state in which p_i is not truthful, *i.e.*

[3] An interesting case is the relaxation of the assumption that beliefs are consistent with existence of a common prior; such models are considered by Lambert and Shoham [2008]; Witkowski and Parkes [2012a]. Theorem 3.5 extends to such cases once players' best responses are well-defined.

[4] One could simply consider a report of a player's signal to be truthful and all other reports untruthful. However, a revelation-principle argument would *not* suffice to extend this result to cases where players compute on inputs rather than merely report information, or settings where information is compressed into reports; this motivates our far more general treatment of truthfulness.

$p_i \notin T_i(\Pr[\omega \mid \Pi_i(\omega^*)])$ for some ω^*. Consider a game with prior $\mathcal{P}[\omega]$ in which player i has partition Π_i and all other players j have a trivial partition $\Pi_j = \{\Omega\}$.

Let $(s_1, \ldots, s_n)$ be a truthful equilibrium. Pick a particular state ω^*; in this state, player i plays according to the distribution $p_i^* = s_i(\Pi_i(\omega^*))$. Since $(s_1, \ldots, s_n)$ is an equilibrium, p_i^* maximizes expected utility against s_{-i} in state ω^*. But s_{-i} is constant on all states of the world (since players $2, \ldots, n$ receive the same signal in every state). So construct the strategy s_i' where, for every ω, $s_i'(\Pi_i(\omega)) = p_i^*$. We immediately have that s_i' is also a best response to s_{-i}.

But by nontriviality, there is some state $\omega' \in \Omega$ such that $p_i^* \notin T_i(\Pr[\omega \mid \Pi_i(\omega')])$. Thus, s_i' is not a truthful strategy; hence $(s_1, \ldots, s_n)$ is not a strongly truthful equilibrium. □

It is of note that a query is not part of the definition of a mechanism; rather, the role of a query in the proof is merely to capture some classification of strategies into truthful and non-truthful. We might ask, for a given mechanism M and for a given query T, whether M can truthfully elicit T with strict incentives. Theorem 3.5 shows that, for every such pair (M, T), the answer is in general *no* unless the mechanism's rewards depend on the information structure. The construction of the proof also illustrates the exact difficulty in an information elicitation setting: *expert knowledge*. Player 1's knowledge of events was strictly more specialized than that of the other players; since they could not distinguish certain events, 1 could not be strictly incentivized to report truthfully according to these events.

4. FORMALIZING PLAYER SPECIFICITY

Theorem 3.5 demonstrates the limitations of *truthful* mechanisms in this setting. But in mechanism design, truthfulness is generally only a means to an end: eliciting useful information. Furthermore, truthfulness is a binary property, while information is complex: If an agent predicts "warm" weather, is she being untruthful, or merely imprecise? Here, we approach this problem by generalizing truthfulness to *specificity* of player reports, capturing the following question: What knowledge does a player use in reporting an answer to a query? To our knowledge, this work is the first to consider such an extension to the traditional notion of truthfulness.

Definition 4.1. Fix a mechanism M and query T and let $\hat{\Pi}$ be a partition of Ω. A strategy s_i for player i is $\hat{\Pi}$-*specific* if:

(1) $\hat{\Pi}$ is a coarsening of player i's partition Π_i; and
(2) for every state ω^*, $s_i(\Pi_i(\omega^*)) \in T\left(\Pr\left[\omega \mid \hat{\Pi}(\omega^*)\right]\right)$.

This definition provides nice properties that one might associate with "weak" truthfulness. First, it requires a player report according to an event that *occurs*: $\hat{\Pi}(\omega^*)$ contains ω^*. For example, given that it is August, a player may report according to the event "it is summer", but will not report as though it were spring. Second, it is consistent across each element of $\hat{\Pi}$. For example, if a player reports according to the month being "August or July" when the month is August, then it makes the same report in July. Meanwhile, the granularity of the information is given by $\hat{\Pi}$. For example, we could have month-specific (August versus July) or season-specific (summer versus fall) information.

There are three levels of specificity of primary interest: Π_i-specific or *private-information-specific* or just *truthful*; $\{\Omega\}$-specific, always reporting according to the prior; and Π-specific or *common-knowledge-specific* strategies.

5. OUTPUT AGREEMENT MECHANISMS

Here, we provide a formal definition and game-theoretic analysis of the output agreement class of mechanisms. We focus primarily on the two-player setting, showing that the mechanisms elicit *common-knowledge-specific* responses with strict incentives. We then examine the structure of equilibria in the game and consider player inference processes, followed by mechanisms for many players. We rely on the definitions in Sections 2 and 4.

Definition 5.1. A two-player *output agreement mechanism* M is a mechanism for eliciting information without verification, defined as follows. The mechanism designer announces a report space $A = A_1 = A_2$ and an associated query T where $T_1 = T_2$ (we will abuse notation by just writing T rather than T_i). The designer selects a distance metric d on the space A and a monotonically decreasing reward function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}$. Each player i makes a report $a_i \in A$ and is paid $h_i^M(\mathcal{I}, a_1, a_2) = h(d(a_1, a_2))$.

This definition assumes that the reward function is *symmetric* in that both players receive the same payoff and that it depends only on the distance between reports and not on the identity of the reports. These restrictions are natural and desirable in many cases; symmetry provides fairness and simplicity while independence of location reduces incentives to bias toward particular reports. However, it may be of interest to examine other settings in future work.

We focus on *singleton* queries: those that, given a posterior distribution, specify exactly one correct response. (Formally, for any posterior $p(\omega)$, $T(p(\omega)) = \{q\}$ where q puts full support on some particular $a \in A$.) These queries capture many natural cases, especially where there is structure to the report space as imposed by a distance metric. In $\mathbb{R}^n$, for instance, the Euclidean distance is a natural choice; in audio transcription, a natural structure might be given by the Levenshtein distance. We then discuss extensions.

THEOREM 5.2. *For any singleton query T, there is an output agreement mechanism eliciting a strict equilibrium that is common-knowledge-specific for T.*

PROOF. For each player i, let s_i be a Π-specific strategy with respect to T; that is, $s_i(\Pi_i(\omega^*)) \in T(\Pr[\omega \mid \Pi(\omega^*)])$. Since Π is the common knowledge partition, we have that in every state ω^*, $s_1(\Pi_1(\omega^*)) = s_2(\Pi_2(\omega^*))$. Furthermore, by assumption, $T(\Pr[\omega \mid \Pi(\omega^*)])$ puts full support on some report a.

By definition, any reward function is monotonically decreasing as a function of distance between reports; let us select a reward function that is strictly decreasing. In any state, both players' strategies put full support on the same report; thus, each player does strictly worse by drawing from any other distribution. Thus (s_1, s_2) is a strict equilibrium. □

We may generalize to cases where there are many possible truthful reports; in some cases, however, weaker incentives are obtained. We illustrate the generalization using the ESP Game as an example. The report space consists of lists of image tags; players are rewarded some number of points if their lists have any tags in common, or no points if their lists do not. The query of the ESP Game would specify, for given beliefs, *any* list containing a "relevant" tag for the image according to those beliefs. There exist equilibria that are Π-specific, but with weak incentives; for example, both players reporting matching lists with some relevant and some irrelevant tags. However, there also exists an equilibrium with incentives that are strict in the sense of *strong truthfulness* as defined in Section 3.2: Each player's strategy is to report a list of tags that are all relevant according to common knowledge. Then either player would receive the same number of points by switching to another list as long as it includes some overlapping, relevant tags; but each is strictly worse off to deviate to an "untruthful" list with no relevant tags. (An interesting question is whether such a mechanism is more effective than a stricter reward rule that strictly incentivizes reporting the single "best" list.)

5.1. Structure of equilibria and player inference

Here, we examine player inference in output agreement games, with the following motivation. Mechanisms for elicitation without verification generally present players with a query specifying the correct response according to given beliefs. In the literature, it is standard to consider cases where this query is made focal by the mechanism in some way and examining the properties of the equilibria that follow. However, in output agreement, this standard approach gives rise to an unusual effect: the resulting sets of truthful strategies are not necessarily equilibria. If player 1 is presented with a given query, she might initially consider a Π_1-specific strategy. But she knows that player 2 should play a best response, which in general is not necessarily Π_2-specific; and then she (player 1) should switch to a best response to that strategy, and so on. We refer to this the process of computing a *sequence of best-response strategies* as *player inference*. It is of note that this process does not consist of players taking or observing actions; rather, it is the hypothetical process of a rational agent computing the optimal strategy to play.[5] We do not necessarily claim that such best-response inference is descriptive of how such games are actually played or prescriptive of how players ought to play. However, analyzing output agreement in the standard information-elicitation-without-verification setting poses the question of how rational players might respond to a given, focal query. An example player inference process is given in Figure 5.1; it gives an example where players are asked to report the most likely realization of a random variable, which may be either ★ or △. We revisit the example in Proposition 5.5.

Ideally, such an inference process, beginning with a private-information-specific report, would converge to the equilibrium of the common-knowledge-specific report, and it would do so regardless of which player "begins" the process. (It always converges to *some* ε-equilibrium for any ε; this is proved in Appendix B.) Here, we show that such convergence does occur when eliciting the mean of a random variable. However, this turns out not to be the case for eliciting the median or the mode, as the example in Figure 5.1 might have already suggested.

PROPOSITION 5.3. *Let t be a random variable taking values in $\mathbb{R}^n$. There is an output agreement mechanism for eliciting the mean of t such that any sequence of best-response strategies, beginning with a Π_i-specific strategy, converges to a Π-specific equilibrium.*

In our context, a *random variable* taking values in some space X is a mapping $t : \Omega \to X$ where $t[\omega]$ specifies the value of t when the state of the world is ω. Thus the query for eliciting the mean in $\mathbb{R}^n$ is $T(p(\omega)) = \mathbb{E}_{\omega \sim p}\, t[\omega]$.

PROOF. We select d to be the Euclidean distance and h to be any affine transformation of $-x^2$. This choice ensures that a player's best response is to report her expected value of her opponent's report. More formally, it is straightforward to verify that the

[5] Two related concepts that arise in other settings are best-response dynamics and fictitious play, in which players play against each other repeatedly and best-respond to (respectively) the opponent's previous action or empirical distribution of past actions. However, these concepts are primarily applied to understand behavior of *non-rational* agents in *repeated games* of *complete information*. On the other hand, we are interested in inference of rational players, and it may be unnatural in a Bayesian game to think of a repeated setting where nature draws a new state of the world from the same prior, or of asking players to report responses to the same query repeatedly.

$\underline{\omega_1}$	$\underline{\omega_2}$	$\underline{\omega_3}$
$\mathcal{P}\left[\omega_1\right] = 0.40$	$\mathcal{P}\left[\omega_2\right] = 0.35$	$\mathcal{P}\left[\omega_3\right] = 0.25$
$t = \bigstar$	$t = \triangle$	$t = \triangle$

(a) The states of the world, common prior, and values of t.

$\underline{\{\omega_1\}}$	$\underline{\{\omega_2, \omega_3\}}$
$\Pr\left[\omega_1\right] = 1.0$	$\Pr\left[\omega_2\right] = 0.58, \Pr\left[\omega_3\right] = 0.42$
mode = ★	mode = △

(b) Player 1's partition Π_1 and posterior beliefs and truthful report in each element of the partition.

$\underline{\{\omega_1, \omega_2\}}$	$\underline{\{\omega_3\}}$
$\Pr\left[\omega_1\right] = 0.53, \Pr\left[\omega_2\right] = 0.47$	$\Pr\left[\omega_3\right] = 1.0$
mode = ★	mode = △

(c) Player 2's partition Π_2 and, for each element of the partition, posterior beliefs and truthful report. Also note that truthful reporting is player 2's best response if 1 is also reporting truthfully.

$\underline{\{\omega_1\}}$	$\underline{\{\omega_2, \omega_3\}}$
$\Pr\left[\omega_1\right] = 1.0$	$\Pr\left[\omega_2\right] = 0.58, \Pr\left[\omega_3\right] = 0.42$
response = ★	response = ★

(d) Player 1's partition, posterior beliefs, and best response when player 2 is reporting truthfully. Player 2's best response to this strategy will be to always report ★, and we will be in equilibrium.

Fig. 1. Truthful strategies and a sequence of best response strategies in an output agreement game. There is a random variable t that can take either the value ★ or the value △ depending on the state of the world. Two players, whose information is given by (b) and (c), are asked to report the most likely value of t and are paid 1 if they agree and 0 if they disagree. To compute a best response given a signal, a player simply determines, according to her posterior beliefs, which value – ★ or △ – is more likely to be reported by her opponent. In this case, if we start with a truthful strategy from either player and iteratively compute best response strategies, we converge to an equilibrium where both players always report ★, even though △ is the common-knowledge most likely value.

unique maximizer of $\mathbb{E}_\omega\, h(d(a, t[\omega]))$ is $a = \mathbb{E}_\omega\, t[\omega]$, where the expectation is taken according to the same distribution in both cases.

We first note that a Π_i-specific report of the mean puts full support on a single response; likewise, by the above, all best responses put full support on a single response (since each is an expected value). Therefore, when considering a sequence of best response strategies beginning with a Π_i-specific one, we need only consider such strategies.

Now we will view a player's strategy s_i as a random variable F_i where $F_i[\omega]$ is the report given full support by $s_i(\Pi_i(\omega))$. Consider the sequence $t, F_i^{(1)}, F_j^{(2)}, F_i^{(3)}, F_j^{(4)}, \ldots$; this is in correspondence with a sequence of best response strategies where $F_i^{(1)}$ is Π_i-specific and each random variable in the sequence consists of expectations of the previous variable according to various posterior beliefs. Formally, in each state ω^*, $F_j^{(k)}[\omega^*] = \sum_\omega \Pr\left[\omega \mid \Pi_j(\omega^*)\right] F_i^{(k-1)}[\omega]$, and the same holds with i and j reversed. This construction allows us to use the following nice result of Samet:

PROPOSITION 5.4 (PROPOSITION $2'$ AND THEOREM $1'$ OF SAMET [1998]). *Let t be a random variable taking values in $\mathbb{R}$ and consider the sequence $t, F_i^{(1)}, F_j^{(2)}, \ldots$ of iterated expected values restricted to states of a fixed element Q of the common-knowledge partition Π. If and only if player beliefs are consistent with the existence of a common prior, then this sequence converges on states in Q, and its value in each state $\omega^* \in Q$ is the same; moreover, this value is $\sum_\omega \Pr\left[\omega \mid Q\right] t[\omega]$.*

This gives that, when $t \in \mathbb{R}$, the sequence of iterated expected values converges to the common-knowledge expected value. To use this result, consider any fixed $Q \in \Pi$. We note that the expected value of a random variable in $\mathbb{R}^n$ is an n-tuple whose k-th entry is the expected value of the k-th entry of the random variable. Therefore, for each

$k = 1, \ldots, n$, a sequence of best responses $F_i^{(1)}, F_j^{(2)}, \ldots$ involves alternately computing, for each ω^*, the expected value of the previous strategy's k-th entry. Therefore, by Samet, the k-th entry of the best response converges to the expected value of the kth entry of t according to the common knowledge posterior when $\omega^* \in Q$.

Because this holds for all entries k of t, this implies that in every state ω^*, the player strategies converge to reporting the expected value of t according to the common knowledge element $\Pi(\omega^*)$. Finally, by Theorem 5.2, we have that reporting the common-knowledge mean actually is an equilibrium. So the inference process converges to the equilibrium where players report the common-knowledge mean. □

This result is encouraging because many natural tasks may be modeled as reporting the mean of some random variable. These could include straightforward numerical queries such as estimating the number of cells in a microscope image; geographical tasks such as estimating the facility location that would minimize average commute time for a large population; or numerical prediction tasks for long-term events (where waiting to reward agents until ground truth becomes available may be undesirable).

However, this nice convergence result does not extend to two of the other most natural properties: median and mode. In fact, this holds more broadly than in $\mathbb{R}^n$; we consider (non-constant) random variables taking values in an arbitrary metric space. By *median* of t, we mean a value in the range of t that minimizes the expected distance to $t[\omega]$. By *mode*, we mean a value in the range of t with highest total probability.

PROPOSITION 5.5. *When $|\Omega| \geq 3$, no output agreement mechanism for eliciting the median or mode of a random variable in an arbitrary metric space ensures for all settings that a sequence of best-response strategies, beginning with a Π_i-specific strategy for either player i, converges to a Π-specific equilibrium.*

PROOF. We first demonstrate that a necessary condition for a sequence of best-response strategies to converge to a Π-specific equilibrium would be that the composition of reward function h and distance metric d be a *strictly proper scoring rule*[6] for the given property (median or mode). We then show that no mechanism with this property is successful by constructing a counterexample for any nontrivial report space.

Consider the case where player 2 is completely informed: the partition Π_2 consists of singleton sets. In every state of the world, player 2 learns the exact value of t. Let player 1 have some strictly coarser partition. Now consider a Π_2-specific strategy of player 2; player 1 has some best response s_1. Then player 2's best response will be to exactly mimic this strategy; player 2 can set $s_2(\{\omega^*\}) = s_1(\Pi_1(\omega^*))$ for every state ω^*. Both players receive the maximum reward in every state, so this is an equilibrium.

This equilibrium is common-knowledge-specific whenever player 1's best response to a Π_2-specific strategy is Π_1-specific (since $\Pi_1 = \Pi$ in this case). But for either median or mode, a Π_2-specific strategy is to simply report the value of the random variable in the observed state. So player 1 faces a scoring rule $h(d(s_1(\Pi_1(\omega^*)), t[\omega^*]))$ in state ω^*. Player 1 has a strict incentive to best-respond truthfully according to Π_1 if and only if h is a strictly proper scoring rule for the appropriate property.

Thus, to elicit either the median or mode, h composed with d must be a strictly proper scoring rule; that is, a best response must be to report the median (respectively, mode) of an opponent's strategy. Now construct a counterexample of a random variable whose range consists of two distinct values. Without loss of generality, suppose there are only three states of the world (otherwise, split them into three groups arbitrarily and

[6] In this context, a *scoring rule* $S(a, a^*)$ takes a report $a \in A$ and a realization $a^* \in A$ of a random variable and returns a payoff; it is *strictly proper* for a property if, for fixed beliefs, reporting the value of the property according to those beliefs uniquely maximizes expected score according to those beliefs.

treat each group as a state). Then we construct the information structure in Figure 5.1, where ★ and $\triangle$ represent the two values in the range (and with an arbitrary, but fixed, distance between them). In this case, the median and mode necessarily coincide, and a best response (as argued above) must be the mode of the opponent's strategy. As demonstrated in Figure 5.1, the common-knowledge most likely value is $\triangle$; however, the inference process always reaches an equilibrium in which ★ is always reported. □

The result is quite negative: In state ω_3, both players are certain that the true realization of the random variable is $\triangle$, yet both report ★ due to their uncertainty about the other's report. Furthermore, this may be generalized to an arbitrarily bad example where the true realization is $\triangle$ $1 - \varepsilon$ of the time, and both players know that the realization is $\triangle$ almost $1 - \varepsilon$ of the time, yet the inference process converges to always reporting ★.

5.2. Mechanisms for Many Players

The primary motivation for considering mechanisms on many players, rather than deterministically dividing into many two-player games, is to utilize the common method of introducing a "trustworthy" agent who plays a "good" equilibrium (in this case, Π-specific). This is only useful if done with small probability: The trusted agent can participate in one in every thousand games, for example, with the game randomly selected; this would allow a mechanism designer to elicit information about one thousand different random variables at once with a single trusted agent.

The mechanism first collects reports, then uses some random procedure to pick a reference report a_j for each player i; i's payoff is $h(d(a_i, a_j))$. Possible random procedures include dividing players into pairs; ordering players into chains or cycles with each compared against the next; and selecting for each player a reference report uniformly at random. Each method may also choose to leave a player unmatched (as with the end of a chain) with a certain probability and pay him a fixed amount; this may reduce the budget required, for example. The following is straightforward:

THEOREM 5.6. *For each of the above comparison procedures, for any singleton query T, there is mechanism eliciting a common-knowledge-specific strict equilibrium (where common knowledge is of every player in the game). Furthermore, there is a variant of each for eliciting the mean of a random variable such that any sequence of best response strategies, for any fixed ordering of the players, to any Π_i-specific strategy converges to a Π-specific equilibrium.*

6. COMPARISON OF MECHANISMS FOR ELICITING INFORMATION

Here, we give context to our results by providing a comparison of several mechanisms for information elicitation without verification. We describe the mechanisms at a high level and defer further details to Appendix C. The mechanisms use a "signals" model of private information: There are different possible events $e \in E$ of interest[7] and a prior $\mathcal{P}[e]$ on e; nature selects an event e and each player i receives some signal t_i; the joint distribution of signals conditional on e is given by a prior $\mathcal{P}[t_1, \ldots, t_n \mid e]$.

6.1. Peer-Prediction (PP) [Miller et al. 2005]

There are two players sharing a common prior $\mathcal{P}[e]$, $\mathcal{P}[t_1, t_2 \mid e]$ that is known to the mechanism designer. It is assumed that signals are *stochastically relevant*: $\Pr[t_j \mid t_i]$ is not the same for all values of t_i. The mechanism uses its knowledge of the priors and the reported signal of each agent to compute a posterior "prediction" for the other agent's signal; this is scored against the signal actually reported.

[7] The literature uses the terminology "states of the world", but these are not equivalent to the states of the world in this paper's model. The correspondences between the two models is given in Appendix A.

The designer learns a large amount of new and useful information: By truthfully eliciting every signal (or element of the partition) of each agent, she learns a finely-partitioned event space given by the intersection of the agents' reported events. However, the theoretical assumptions are quite strong: The mechanism designer must have full knowledge of the common prior distribution as well as the players' partition structures. Furthermore, the signal structure as known by the mechanism designer and reported by the agents must capture agents' beliefs completely. For example, consider rating a restaurant from one to five. Two agents who observe a restaurant's quality as "four stars" may have completely different posterior distributions over others' ratings: The rater may be particularly picky, for instance, or particularly fond of Chinese food. The mechanism's computed posterior must take this into account to maintain incentives; this may make implementation in practice difficult.

Peer Prediction Without a Common Prior (PPwoCP) [Witkowski and Parkes 2012a]. In this variant, the PP setting is restricted to just two possible signals, low and high; each agent has a *private* prior distribution over events e and signals, not necessarily known to the other or to the mechanism, but satisfying that signals are generated independently conditional on e. It is required that agents first report a prior probability that some other agent receives a high signal, *then* receive their signal, update to a posterior probability that some other agent receives a high signal, and report it. This requirement and that of binary, conditionally independent signals limit PPwoCP's use cases, but it is straightforward to implement.

6.2. Bayesian Truth Serum (BTS) [Prelec 2004]

There is a "sufficiently large" or countably infinite population of agents who share a common prior. The information structure is assumed to be *impersonally informative*: Signals are generated i.i.d. conditional on e, and no two signals map to the same posterior.[8] Each agent i makes two reports: his signal t_i, and his posterior probability distribution p_i predicting the empirical distribution of responses.

The mechanism designer need not know the information structure. However, there is a small catch. In PP, the mechanism designer knows the information structure and thus can perfectly interpret reported signals to reason about events or random variables. In BTS, the designer does not necessarily have such information and never learns how signals map to states of the world or other values of interest. Of course, in cases such as surveys, the distribution of signals may be the sole desired information.

Robust Bayesian Truth Serum (RBTS) [Witkowski and Parkes 2012b]. This modification to BTS relaxes the requirement of a large population, functioning for $n \geq 3$ agents, but restricts to binary signals. The common prior must satisfy the impersonally informative condition. It elicits limited information: The operator essentially learns the number or proportion of "high" signals in the population. However, implementation is easy. Unlike PP, it does not require the mechanism designer to write down and compute with a common prior; unlike PPwoCP, it does not require temporal separation of signals; and unlike BTS, it does not require a large population of players.

6.3. Output Agreement Mechanisms (OA)

Output agreement mechanisms achieve weaker honesty results than the other methods in that they elicit only common-knowledge-specific, rather than private-information-specific, responses. This is the primary drawback of the method. However, it obtains several advantages in return.

[8] The phrase "impersonally informative" refers to the fact that any two agents receiving the same signal will have identical posterior beliefs, regardless of their identities.

Table I. Characteristics of Mechanisms

	PP	PPwoCP	BTS	RBTS	OA
common prior must exist	yes	no	yes	yes	yes
other assumptions on information structure	stochastic relevance	conditional independence	impersonally informative	impersonally informative	none
designer must know information structure	yes	no	no	no	no
# of players	≥ 2	≥ 2	$\approx \infty$	≥ 3	≥ 2
# of signals	any	2	any	2	any
type of player reports	signal	prior and posterior predictions of signals	signal and prediction of signals	signal and prediction of signals	arbitrary
designer learns?	everything	distribution of signals	distribution of signals	distribution of signals	common knowledge

Properties of various mechanisms for eliciting information without access to ground truth: peer prediction, PP without a common prior, Bayesian truth serum, robust BTS, and output agreement.

First, OA allows for very relaxed assumptions on the knowledge of the mechanism designer. It is the only method which makes no assumptions either on designer's knowledge of the information structure or the information structure itself.

Second, OA does not make restrictions on agents' information structures beyond consistency with existence of a common prior. All other methods except PP impose strict restrictions on the signal structure; meanwhile, PP imposes a *communication complexity* restriction: As mentioned, signals in PP must capture the entirety of agent beliefs, so we are forced to limit the complexity of agent beliefs in the model to the number of bits we are able to communicate in a reasonable amount of time. Thus, for example, producing a restaurant rating may actually be a task better-suited to OA, where it might be interpreted as "expected quality of experience" and reports may be a single number, than to PP, where for incentive compatibility it *must* be interpreted as "report all relevant information about your experience there".

Third, OA allows for an arbitrary report space and allows the designer to specify a query of interest directly rather than eliciting signals and distributions over signals. This is achieved by Lambert and Shoham [2008]; Goel et al. [2009], but only for special cases of conditionally independent random processes in $\mathbb{R}$.

Of course, OA is limited to cases in which common knowledge is the desideratum. However, such cases may be quite common in *e.g.* markets for crowdsourcing, where private information may be difficult to distinguish from noise and common-knowledge consensus serves as validation. It is also of note that common knowledge is a property of the particular group playing the game; selecting a group of experts to play an output agreement game could result in high-quality responses if the answers are common knowledge among experts in that field.

7. CONCLUSIONS

A broad theme and common goal of mechanism design is the elicitation of accurate information. Much recent work focuses on understanding what types of information can be elicited, how, and under what circumstances. In particular, human computation has motivated the proposal of a large number of mechanisms for eliciting information when ground truth is unverifiable.

Formalizing this setting as a sub-field of mechanism design yields new insights and broad impossibility results on the capabilities of mechanisms in this space. First, such mechanisms (almost) always have uninformative equilibria. Second, a mechanism cannot incentivize agents to report truthfully according to private information unless it either assumes that the mechanism designer has some knowledge of the players' information structures or else restricts the class of information structures. The applications of these results are most immediate in the area of human computation; additionally, the implications may extend to other mechanism design settings such as differential privacy, where one wishes to solicit computations on a privately held dataset but may not be able to verify reported outputs.

These results highlight a discrepancy between theory and practice: Theoretical approaches tend to be complex and make restrictive assumptions, yet simple, useful mechanisms are observed in practice. Resolution of this paradox requires a new approach to truthfulness in mechanism design. By generalizing truthfulness to the criterion of *specificity* of reports, we are able to show that the output agreement class of mechanisms, an extremely broad class that assumes neither designer knowledge nor restricted information structures, successfully elicits *common knowledge*. This result may be particularly positive in common human computation settings, where such an answer is exactly what is desired by the mechanism designer.

REFERENCES

AUMANN, R. J. 1976. Agreeing to disagree. *Annals of Statistics 4,* 6, 1236–1239.

CHAWLA, S., HARTLINE, J., AND SIVAN, B. 2012. Optimal crowdsourcing contests. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '12. SIAM, 856–868.

DELLA PENNA, N. AND REID, M. 2012. Crowd & prejudice: An impossibility theorem for crowd labelling without a gold standard. In *Collective Intelligence*. CI '12.

GLICKSBERG, I. 1951. *A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points*. Defense Technical Information Center.

GOEL, S., REEVES, D., AND PENNOCK, D. 2009. Collective revelation: A mechanism for self-verified, weighted, and truthful predictions. In *Proceedings of the 10th ACM conference on Electronic commerce*. EC '09. ACM, 265–274.

HUANG, S. AND FU, W. 2012. Systematic analysis of output agreement games: Effects of gaming environment, social interaction, and feedback. In *Proceedings of HCOMP 2012: The Fourth Workshop on Human Computation*.

JAIN, S., CHEN, Y., AND PARKES, D. 2009. Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM conference on Electronic commerce*. EC '09. ACM, 129–138.

JAIN, S. AND PARKES, D. 2008. A game-theoretic analysis of games with a purpose. In *Internet and Network Economics*. WINE '08. 342–350.

JURCA, R. AND FALTINGS, B. 2005. Enforcing truthful strategies in incentive compatible reputation mechanisms. In *Internet and Network Economics*. WINE '05. Springer, 268–277.

JURCA, R. AND FALTINGS, B. 2006. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce*. EC '06. ACM, 190–199.

JURCA, R. AND FALTINGS, B. 2007a. Collusion-resistant, incentive-compatible feedback payments. In *Proceedings of the 8th ACM conference on Electronic commerce*. EC '07. ACM, 200–209.

JURCA, R. AND FALTINGS, B. 2007b. Robust incentive-compatible feedback payments. In *Agent-Mediated Electronic Commerce*, M. Fasli and O. Shehory, Eds. Vol. LNAI

4452. Springer-Verlag, Berlin Heidelberg, 204–218.

JURCA, R. AND FALTINGS, B. 2008. Incentives for expressing opinions in online polls. In *Proceedings of the 9th ACM Conference on Electronic Commerce*. EC '08. ACM, 119–128.

JURCA, R. AND FALTINGS, B. 2009. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research 34,* 1, 209.

LAMBERT, N. 2011. Elicitation and evaluation of statistical forecasts. Manuscript.

LAMBERT, N., PENNOCK, D., AND SHOHAM, Y. 2008. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*. EC '08. ACM, 129–138.

LAMBERT, N. AND SHOHAM, Y. 2008. Truthful surveys. *Internet and Network Economics*, 154–165.

LAMBERT, N. AND SHOHAM, Y. 2009. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the Tenth ACM Conference on Electronic Commerce*. EC '09. 109–118.

MCKELVEY, R. D. AND PAGE, T. 1986. Common knowledge, consensus, and aggregate information. *Econometrica 54,* 1, 109–127.

MILLER, N., RESNICK, P., AND ZECKHAUSER, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science 51,* 9, 1359–1373.

MONDERER, D. AND SHAPLEY, L. 1996. Potential games. *Games and economic behavior 14,* 1, 124–143.

NASH, J. 1951. Non-cooperative games. *Annals of mathematics 54,* 2, 286–295.

NIELSEN, L. T., BRANDENBURGER, A., GEANAKOPLOS, J., MCKELVEY, R., AND PAGE, T. 1990. Common knowledge of an aggregate of expectations. *Econometrica 58,* 5, 1235–1238.

OSTROVSKY, M. 2009. Information aggregation in dynamic markets with strategic traders. In *Proceedings of the 10th ACM conference on Electronic commerce*. EC '09. ACM, 253–254.

PRELEC, D. 2004. A bayesian truth serum for subjective data. *Science 306,* 5695, 462–466.

ROSENTHAL, R. 1973. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory 2,* 1, 65–67.

SAMET, D. 1998. Iterated expectations and common priors. *Games and economic Behavior 24,* 1-2, 131–141.

VON AHN, L. AND DABBISH, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on human factors in computing systems*. CHI '04. ACM, 319–326.

VON AHN, L. AND DABBISH, L. 2008. Designing games with a purpose. *Communications of the ACM 51,* 8, 58–67.

WEBER, I., ROBERTSON, S., AND VOJNOVIC, M. 2008. Rethinking the esp game. In *Proceedings of the 27th International Conference on Human factors in Computing Systems*. CHI '08 Series, vol. 9. 3937–3942.

WITKOWSKI, J. AND PARKES, D. 2012a. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. EC '12. ACM, 964–981.

WITKOWSKI, J. AND PARKES, D. 2012b. A robust bayesian truth serum for small populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. AAAI '12.

Received ??; revised ??; accepted ??

Online Appendix to: Information Elicitation *Sans* Verification

Bo Waggoner, Harvard University
Yiling Chen, Harvard University

A. EQUIVALENCE OF "STATES OF THE WORLD" AND "SIGNALS" MODELS

For completeness, we show that two alternative models of information structure are equivalent by showing that each can model the other. The result is straightforward, but the correspondences are perhaps not immediately intuitive or trivial.

We assume a finite number of n agents.

The *states of the world* model is as follows. There is a finite set of states Ω and a common prior $\mathcal{P}\left[\omega\right]$ on them. Each agent i has a partition Π_i of Ω; when the true state of the world is ω^*, agent i learns that it is in the set of states $\Pi_i(\omega^*)$, the element of his partition containing ω^*. His posterior distribution is given by the prior restricted to this subset, $\Pr\left[\omega \mid \Pi_i(\omega^*)\right]$, which is equal to $\Pr\left[\{\omega\} \cap \Pi_i(\omega^*)\right] / \Pr\left[\Pi_i(\omega^*)\right]$.

The *signals* model is as follows. There is a set E of possible events of interest[9] and a common prior $\mathcal{P}\left[e\right]$. Each agent i has a set T_i of possible signals. When the true state of the world is e^*, the n agents receive the respective signals $(t_1, \ldots, t_n) \in T_1 \times \cdots \times T_n$ with probability given by $\Pr\left[t_1, \ldots, t_n \mid e\right]$. Each agent i updates to a posterior distribution on the state and others' signals according to the following computation derived from Bayes' rule: $\Pr\left[e, \{t_{j \neq i}\} \mid t_i\right] = \Pr\left[t_i, \{t_{j\neq i}\} \mid e\right] \Pr\left[e\right] / \Pr\left[t_i\right]$, with $\Pr\left[t_i\right]$ given by $\sum_{e, \{t_{j \neq i}\}} \Pr\left[t_i, \{t_{j \neq i}\} \mid e\right] \mathcal{P}\left[e\right]$.

PROPOSITION A.1. *The models are equivalent.*

PROOF. ($\subseteq$) Let a "states" model $(\Omega, \mathcal{P}\left[\omega\right], \{\Pi_i\})$ be given. Let $E = \Omega$, $\mathcal{P}\left[e\right] = \mathcal{P}\left[\omega\right]$, and $T_i = \Pi_i$ with each signal corresponding to an element of the partition. It follows that $\Pr\left[t_1, \ldots, t_n \mid e\right] = 1$ if the state e is in the intersection of the sets t_i and 0 otherwise. We must show that, when agent i receives signal t_i, his posterior is given by $\Pr\left[e, \{t_{j\neq i}\} \mid t_i\right]$. Our mapping gives us immediately that he has a posterior $\Pr\left[e \mid t_i\right]$ on states e; the rest follows because the state e completely determines the signals $t_1, \ldots, t_n$.

($\supseteq$) Let a "signals" model $(E, \mathcal{P}\left[e\right], \{T_i\}, \Pr\left[t_1, \ldots, t_n \mid e\right])$ be given. We let $\Omega = E \times T_1 \times \cdots \times T_n$. When $\omega = (e, t_1, \ldots, t_n)$, we let $\mathcal{P}\left[\omega\right] = \Pr\left[t_1, \ldots, t_n \mid e\right] \mathcal{P}\left[e\right]$. We let each element of player i's partition correspond to a signal $t_i \in T_i$, so that when $\omega^* = (e^*, t_1, \ldots, t_n)$, $\Pi_i(\omega^*) = \{(e, \hat{t}_1, \ldots, \hat{t}_n) : \hat{t}_i = t_i\}$. It only remains to show that an agent's posterior in state ω^* is given by $\Pr\left[\omega \mid \Pi_i(\omega^*)\right]$. We are given that the agent computes a posterior $\Pr\left[e, \{t_{j\neq i}\} \mid t_i\right]$; but since e, $\{t_{j \neq i}\}$, and t_i determine ω, and $t_i = \Pi_i(\omega^*)$, we are done. □

B. CONVERGENCE OF PLAYER INFERENCE IN OUTPUT AGREEMENT

Here, we show that that any sequence of best responses in an output agreement game converge to a ε-equilibrium for every ε. A strategy profile $(s_1, \ldots, s_n)$ is an *ε-equilibrium* if, for each player i, the expected utility from playing any strategy s_i' is no more than ε greater than the expected utility for playing s_i when i's opponents play s_{-i}.

[9] In the literature, E is commonly referred to as the "states of the world"; however, they are not equivalent to the states of the world $\omega \in \Omega$ and so we use the term "event" to avoid confusion.

DOI: http://dx.doi.org/10.1145/0000000.0000000

PROPOSITION B.1. *Every sequence of best response strategies in an output agreement game converges to a ε-equilibrium for every ε.*

PROOF. Let a two-player output agreement game $G = (M, \mathcal{I})$ be given. We use an argument in the style of *potential games* [Rosenthal 1973; Monderer and Shapley 1996]. Fixing $\Pi_1, \Pi_2, \mathcal{P}[\omega]$, let $F(s_1, s_2) = \sum_\omega \mathcal{P}[\omega]\, h(d(s_1(\Pi_1(\omega)), s_2(\Pi_2(\omega))))$; this is the expected utility of both players when playing strategies s_1 and s_2. Then F is bounded above by the maximal value of h (*i.e.* $h(t,t)$ for a constant t), and $F(s_1', s_2) \geq F(s_1, s_2)$ if s_1' is a best response to s_2, for all s_1, s_2, s_1' (and analogously for s_2'). Therefore any sequence of best response strategies may be put in correspondence with a monotonically increasing sequence of values for F; since F is bounded above, this sequence converges to some value c.

Thus, for any ε, the values of F eventually converge to a value of $F(s_1, s_2) > c - \varepsilon$; at this point, the improvement each player makes from a best response strategy is bounded by $c - F(s_1, s_2) < \varepsilon$. □

C. MECHANISMS FOR INFORMATION ELICITATION

Here, we overview various mechanisms for information elicitation without verification in further detail, giving the methods of scoring and intuition behind the proofs of incentive compatibility.

C.1. Peer-Prediction (PP) [Miller et al. 2005]

There are two players. Nature selects the state of the world from a common prior $\mathcal{P}[e]$, and each player i receives a signal t_i and reports it. The joint set of signals is drawn from a common prior $\mathcal{P}[t_1, \ldots, t_n \mid e]$. Both priors are known to all players and the mechanism designer. It is assumed that signals are *stochastically relevant*: $\Pr[t_j \mid t_i]$ is not the same for all values of t_i.

Each player i reports his observed signal t_i. The mechanism designer, using her knowledge of the common priors and the reported signal of player 1, computes player 1's posterior distribution $\Pr[t_2 \mid t_1]$ over player 2's signal (and vice versa). The designer then uses a *strictly proper scoring rule*[10] to score this posterior against the signal actually reported by player 2 (and vice versa). The intuition behind truthfulness is straightforward: Since the mechanism designer knows the information structure, the posterior she computes for player 1 is correct whenever player 1 reports truthfully; and because of the strictly proper scoring rule, player 1 in general does strictly best when the posterior computed matches 1's beliefs. We may view this as a revelation-principle-type approach, but with a two-for-one: The mechanism learns both the agent's signal and posterior from a single report; both pieces of information are necessary to run the mechanism. (Along these lines, asking agents to report posteriors instead of signals will only work in cases where the signal can be uniquely inferred.)

C.2. Peer Prediction Without a Common Prior (PPwoCP) [Witkowski and Parkes 2012a]

Here, the PP setting is restricted to just two possible signals, low (0) and high (1). Each player has a private prior distribution over states of the world and signals conditional on the state; beliefs need not be consistent with the existence of a common prior. Beliefs for each player i are assumed to be *admissible*: $|E| \geq 2$; $\mathcal{P}[e] > 0 (\forall e \in E)$; $\Pr[t_i = 1 \mid e] \neq \Pr[t_i = 1 \mid e']$ if $e \neq e'$; $0 < \Pr[t_i = 1 \mid e] < 1 (\forall e)$. Admissibility is a relatively weak requirement, particularly in the two-signal setting. It is also assumed

[10] A strictly proper scoring rule $S(p,t)$ taking a probability distribution p and outcome t satisfies that the unique maximizer of $\mathbb{E}_{t\sim p}\, S(\hat{p}, t)$ is $\hat{p} = p$.

that players believe signals to be generated independently conditional on the state of the world; that is, player i believes that $\Pr[t_1 = 1 \mid e] = \Pr[t_2 = 1 \mid e]$.

Each player first reports a prior belief that the other player will receive a high signal, then receives his signal, updating to and reporting a posterior belief that some other player received a high signal. The mechanism infers his signal from the change in the posterior: If the believed probability of a high signal increases, it is assumed that he received a high signal. The mechanism then scores both of his reports (the prior and posterior) against the other player's inferred signal using a strictly proper scoring rule. The mechanism is truthful because in both scenarios – before and after receiving his signal – a player wishes to report his current beliefs truthfully (due to the use of a strictly proper scoring rule). Further, because he believes signals are generated independently conditional on e, whenever he receives a high signal, he believes it more likely that his opponent will receive a high signal, so his posterior probability increases and the mechanism does indeed infer his signal correctly.

C.3. Bayesian Truth Serum (BTS) [Prelec 2004]

There is a "sufficiently large" or countably infinite population of players. Agents share a common prior $\mathcal{P}[e]$ on states of the world $e \in E$. There is a set of m signals. The information structure is assumed to be *impersonally informative*: For a given signal t, every player who receives t has an identical posterior distribution $\Pr[e \mid t]$. (This is equivalent to conditionally independent signals: $\mathcal{P}[t_i \mid e] = \mathcal{P}[t_j \mid e]$.) It is also assumed that no two signals map to the same posterior. The mechanism designer is not assumed to have any knowledge over the information structure. Each player i makes two reports: his signal t_i, and his posterior probability distribution p_i predicting the empirical distribution of responses.

To score players, first compute for each signal t the empirical frequency $\bar{t}$ and the geometric average prediction $\bar{p}(t)$. All players reporting the signal t then receive an "information" score $\log(\bar{t}/\bar{p}(t))$. This score rewards signals which are "surprisingly common" – *i.e.*, reported by a relatively high fraction of respondents but collectively predicted to be rare. The intuition behind truthfulness is that an player receiving signal t should expect that, on average, the rest of the population will underestimate the how many people observe t. Each player then receives a "prediction" score given by the relative entropy (KL-divergence) between the empirical signal frequencies $\bar{t}$ and the player's prediction p_i. We may view this as simply applying the logarithmic scoring rule, a strictly proper scoring rule, for prediction p_i over each of the signals reported by the other players; truthfulness follows immediately.

C.4. Robust Bayesian Truth Serum (RBTS) [Witkowski and Parkes 2012b]

This modification to BTS relaxes the requirement of a large population, but restricts to binary signals. There are $n \geq 3$ players sharing a common prior $\mathcal{P}[e]$ on the states of the world E. There are two possible signals, 1 (high) and 0 (low), and the players share an admissible (see PPwoCP) common prior $\mathcal{P}[t \mid e]$ of the probability of any player receiving a signal given a state. This implies the impersonally informative condition.

Each player reports a signal t_i and a predicted probability p_i that some other player receives the high signal 1. The player's prediction is scored using a proper scoring rule on his prediction p_i against the reported signal t_k of some other player k.

His information score is computed using the *shadowing method*. The idea is to assign i a prediction, then use a proper scoring rule on that prediction. To assign i a prediction, take the prediction report p_j of some other player j and "shadow" it in the direction of i's report: take $p_j + \delta$ if i received a high signal, or $p_j - \delta$ if i received a low signal (where $\delta > 0$ is any reasonably small constant). We then score the resulting number against the report of (say) player k using the quadratic scoring rule: $-(t_k - p)^2$ for prediction p. This gives the rest of i's payoff.

As intuition for truthfulness, assume that j reports honestly; then p_j is the result of a Bayesian update based on j's signal. Now suppose j were to learn i's signal; then j would do a second Bayesian update to some posterior p_j^*. This would be the "ideal" prediction report given both player's information; so i would like her assigned "prediction" to be as close as possible to this value. (This is always true in particular because we use the quadratic scoring rule.) But if i's signal is high, then $p_j^* > p_j$, and if it is low, then $p_j^* < p_j$ (this follows from the conditional independence of signals). So when i has a high signal, she would rather p_j be shadowed up than down, and vice versa.

C.5. Collective Revelation [Goel et al. 2009]

As with the previous problem, there is a random process generating values i.i.d. from an unknown distribution. Here, however, each agent may observe multiple generated values (signals). It is assumed that the distribution has a particular form; the paper provides a mechanism for the Bernoulli distribution and notes that the normal, Poisson, and exponential distributions also have associated mechanisms. It is assumed that there is a common prior on the distribution of the *parameter* of the underlying distribution. The mechanism asks players to make two reports: Their expected value of the random variable, and a *hypothetical* expected value of the random variable supposing they were to observe a certain number of successes in a certain number of additional trials. The mechanism then exploits a bijection between these pairs of reports and posterior beliefs to infer each player's true beliefs. It is able to use these beliefs to generate a prediction for other reports, as in peer-prediction.

C.6. Truthful Surveys [Lambert and Shoham 2008]

In this work, a mechanism designer is interested in a random process that is generating values i.i.d. according to some unknown distribution. A large population of players each observes a value generated by this distribution. Players may have arbitrary beliefs over the distribution from which values generated as long as they believe them to be i.i.d.

First, consider a scenario where each player draws from a different, but publicly known distribution. In this case, the mechanism designer can use the cumulative distribution function of each player's distribution to convert reports into a uniform random value in $[0, 1]$. So now, without loss of generality, suppose players each report a value drawn from the interval $[0, 1]$. The authors provide a reward rule such that player rewards are zero-sum, and the unique equilibrium is for each player to draw randomly from this interval. The reward rule enforces that, when being scored against player j, player i would prefer to either have a converted report that is 0.5 to 1 greater than j's converted report, or else 0 to 0.5 less than j's converted report.[11] Thus, it is quite intuitive that both players drawing *uniformly* from $[0, 1]$ is the unique equilibrium: If any player puts more weight on a particular part of the interval, the other would like to shift to a report that tends to undercut it, or (if it is low enough) to a report that tends to be much higher than it. This will give positive expected utility, so the first player will have negative expected utility. But by the same argument, now the other player is playing a shifted strategy and so the first player has some shifted best response; and so on. Thus, the only equilibrium will be for both to draw uniformly at random. This intuition extends to when players are scored against *all* other players.

[11] The authors explain the intuition with a lazy hiker analogy in which the direction of the ranges are reversed, but this does not affect incentives.

Now consider the case where the designer does not know the players' private beliefs. However, when all players report random values, she can use these values to compute an unbiased statistical estimator of the empirical cumulative distribution function; she then uses the mechanism above, with the same distribution function for every player, to score reports. Reporting truthfully is an equilibrium: Even when players have different beliefs over the signal generation process, each believes that the empirical distribution will match her own beliefs, so each believes that she will be scored according to the mechanism described in the full-information case.

However, it is not a strict equilibrium to report one's signal truthfully: When all other players are drawing reports randomly, any fixed report is a best response. Weak incentives may be justified with a qualitative argument: The only equilibria of the game are when all players are drawing values from the same random process, so although individual deviations are possible, selection of a different equilibrium would require all players drawing from some other focal random process.