Decision
2017, Vol. 4, No. 1, 25–51

2325-9965/17/$12.00 http://dx.doi.org/10.1037/dec0000052

Inducing Honest Reporting of Private Information in the Presence of Social Projection

Arthur Carvalho
Erasmus University

Stanko Dimitrov and Kate Larson
University of Waterloo

We discuss payment structures that induce honest reporting of private information by risk-neutral agents in settings involving multiple-choice questions. Such payment structures do not rely on the existence of ground-truth answers, but instead they rely on the assumption that agents exhibit social projection. Social projection is a strong form of the well-known psychological phenomenon called the false-consensus effect, where an agent believes that his private answer to a multiple-choice question is the most popular answer. From a theoretical perspective, we first show that when social projection holds true, honest reporting strictly maximizes an agent's expected reward from a payment structure that simply compares agents' reported answers and rewards agreements. Furthermore, we suggest how to induce honest reporting by taking the distance between reported answers into account when social projection is strong, i. e., when an agent believes that his private answer is more likely to be reported by a random peer than all the other answers combined. We also discuss how to derive the above results in terms of proper scoring rules. From an empirical perspective, we investigate the consequences of using a payment structure that rewards agreements in a content-analysis experiment on Amazon Mechanical Turk. We obtain some evidence that, under such a payment structure, agents report more accurate answers than when there are no direct incentives for honest reporting of private answers. Moreover, we find that priming agents by briefly mentioning the theoretical properties of the underlying payment structure results in even more accurate answers.

Keywords: incentive engineering, social projection, false-consensus effect, peer-prediction method, crowdsourcing, Amazon Mechanical Turk

There are many scenarios where a *requester* is interested in eliciting private information from a group of *agents*. For example, when agents are market experts, a company might elicit predictions about consumer demand and material supply to make its production plan. Likewise, when agents are weather forecasters, farmers might elicit weather predictions so as to formulate guidelines for long-range or seasonal agricultural planning in terms of crops that are best suited to the anticipated climatic conditions. *Crowdsourcing* is another domain highly dependent on the elicitation of agents' private information. Crowdsourcing consists of the practice of obtaining relevant information or services from a large group of people (Chiu, Liang, & Turban, 2014a; Geiger & Schader, 2014). Recent technological advances have facilitated the outsourcing of a variety of tasks to "the crowd," for example, the decision support regarding various phases of managerial decision making and problem solving (Chiu et al., 2014b), the design of

This article was published Online First March 14, 2016.

Arthur Carvalho, Rotterdam School of Management, Erasmus University; Stanko Dimitrov, Department of Management Sciences, University of Waterloo; Kate Larson, Cheriton School of Computer Science, University of Waterloo.

We acknowledge Craig Boutilier, Pascal Poupart, Daniel Lizotte, Selcuk Onay, Ariel Procaccia, and Xi Alice Gao for useful discussions. We thank Carol Acton, Katherine Acheson, Stefan Rehm, Susan Gow, and Veronica Austen for providing gold-standard answers for our experiment. We also thank the Natural Sciences and Engineering Research Council of Canada for funding this research.

Correspondence concerning this article should be addressed to Arthur Carvalho, Rotterdam School of Management, Erasmus University, Burgemeester Oudlaan 50, T Building, Room 40-9th floor, 3062 PA Rotterdam, The Netherlands. E-mail: carvalho@rsm.nl

advertisements (Ren, Nickerson, Mason, Sakamoto, & Graber, 2014), the development and testing of software applications, the design of websites, and so forth.

In this article, we are particularly interested in eliciting information from agents regarding the most suitable answer for a multiple-choice question. We argue that several tasks can be phrased in terms of multiple-choice questions, for example, rating, where the rating scale defines the underlying answers; categorization/classification, where answers are the available categories; sentiment analysis, where answers are major emotional states; and so forth. In these settings, each agent has a private answer, and the requester is interested in obtaining that answer. Honest reporting means that an agent reports precisely his private answer.

We note, however, that in the absence of a well-chosen incentive structure, agents are not necessarily honest when reporting their answers. For example, agents might bias their reported answers in a way that makes them look more favorable to the requester or toward what they believe is socially desirable (Antin & Shaw, 2012). In personality tests, which are multiple-choice questions where agents rate the degree to which some statements reflect their own behavior, the social desirability bias might account for as much as 10%–75% of the variance in the reported answers, depending on the underlying context and population (Nederhof, 1985). A crucial question is then how to promote honest reporting of private information.

In the context of elicitation of subjective probabilities, *proper scoring rules* (Winkler & Murphy, 1968) are traditional devices that incentivize honest reporting of beliefs by risk-neutral agents. More specifically, agents maximize their expected scores from a proper scoring rule by honestly reporting their beliefs. Proper scoring rules rely on the assumption that there is an observable future outcome, or a ground truth, which is not always a reasonable assumption. For example, when market analysts provide sales forecasts on a potential new product, there is no guarantee that the product will ever be produced. Hence, the actual number of sales may never be observed.

In this article, we suggest payment structures to induce honest reporting of private answers by risk-neutral agents that do not rely on the existence of ground-truth answers. Instead, those payment structures rely on the assumption that *social projection* holds true. As we elaborate later, social projection is a strong form of the *false-consensus effect* (Ross, Green, & House, 1977). Social projection captures the psychological phenomenon where agents believe that their own private answers are more popular than alternative answers. This cognitive bias leads to the perception of a consensus that does not necessarily exist, that is, a "false consensus."

We first show that when agents exhibit social projection and they cannot communicate to one another, a payment structure that simply compares pairs of reported answers and rewards agreements is enough to induce honest reporting of private answers by risk-neutral agents. Although this first payment structure might work well when the number of possible answers is small, we argue that using this simple payment structure might be troublesome when the number of possible answers is high due to the potential lack of agreements. A potential solution to this issue is to take the distance between reported answers into account when rewarding agents. We show that a payment structure that penalizes disagreements in proportion to the distance between pairs of reported answers induces honest reporting of private answers when social projection is strong, that is, when an agent believes that his private answer not only is the answer most likely to be reported by a random peer but also more likely than all the other answers combined. Our last theoretical contribution is about how to derive the above results in terms of proper scoring rules.

On the empirical side, we report the results of a content-analysis experiment on Amazon Mechanical Turk. In particular, we show how making payments based on pairwise comparisons between reported answers results in slightly more accurate and similar reported answers than when agents only receive fixed payments. This is an interesting and rather surprising result because it relates honest reporting to the accuracy of the reported answers. We also show how a requester can further increase accuracy and similarity by priming the agents through a brief description of the theoretical properties of the payment structure.

Related Work

Two prominent methods to induce honest reporting without relying on the existence of a ground-truth answer have been recently proposed: the *Bayesian truth serum method* (BTS) (Prelec, 2004) and the *peer-prediction method* (Miller, Resnick, & Zeckhauser, 2005).

Similar to the setting we investigate in this article, the BTS method works on a single multiple-choice question with a finite number of answers. Each agent endorses the answer most likely to be correct and predicts the empirical distribution of all available answers. The requester then pays each agent based on the accuracy of his prediction and on whether the reported answer is more common than collectively predicted.

Mathematically, the score received by an agent from the BTS method has two major components. The first one, called *information score*, evaluates the agent's reported answer according to the log-ratio of its actual-to-predicted endorsement frequencies. The second component, called *prediction score*, is a penalty proportional to the relative entropy between the empirical distribution of reported answers and the agent's prediction of that distribution. Under the BTS scoring method, collective honest reporting is a Bayes-Nash equilibrium, that is, the best action that an agent can take (in expectation) given that all his peers are reporting honestly is also to report honestly.

The BTS method has been used to promote honest reporting in many different domains, for example, when sharing rewards among a set of agents (Carvalho & Larson, 2011) and in policy analysis (Weiss, 2009). However, there are three major drawbacks with the BTS method. First, it requires the population of agents to be large. Second, besides reporting their answers, agents must also make predictions about how their peers will report. Although the artificial intelligence community has recently addressed the former issue (Radanovic & Faltings, 2013; Witkowski & Parkes 2012a), the latter issue is still an intrinsic requirement for using the BTS method. Third, it might be complicated for the requester to explain the mathematics behind the BTS to less mathematically inclined agents.

The peer-prediction method (Miller et al., 2005), on the other hand, does not share the drawbacks of the BTS method. In the original setting of the peer-prediction method, a number of agents experience a product and rate its quality. Mathematically, each agent observes a signal of the product's type where, conditional on the product's type, agents' signals are independent and identically distributed. A requester then rewards the agents based on the reported ratings. The peer-prediction method makes use of the stochastic correlation between the signals observed by the agents from the product to achieve a Bayes–Nash equilibrium, where every agent reports honestly. We quickly discuss how to adapt the peer-prediction method to our setting in the next section.

In spirit, the peer-prediction method is similar to models from cultural consensus theory (Romney et al., 1986). In particular, cultural consensus models assume that a group of agents answers questions that might not have objective answers. Thereafter, each agent receives a score based on how common/likely his answers are, which is referred to as the agent's individual (cultural) competence. However, different from the peer-prediction method, agents' scores are not used to induce honest reporting. Instead, they serve as weights when combining agents' answers.

The peer-prediction method further relies on two crucial assumptions, namely that the underlying conditional distribution is common knowledge and agents hold a common prior distribution over the product's type. Witkowski and Parkes (2012b) suggested a way to circumvent these assumptions under the extra assumption that the requester is able to distinguish the time before agents observe the signals from the time after agents observe the signals. More precisely, the model proposed by Witkowski and Parkes (2012b) has four major steps. First, each agent reports his private prior belief about what other agents will observe. Second, each agent observes a binary signal. Third, each agent reports either his signal value ("the shadow mechanism") or his private posterior belief ("the basic private-prior peer-prediction mechanism"). Finally, the requester makes payments based on proper scoring rules so as to induce honest reporting by risk-neutral agents. Besides the assumption of the existence of such a temporal structure, a potential drawback with the payment structures proposed by Witkowski and Parkes (2012b) when applied to our setting is that they can only be used when the underlying

multiple-choice question has two possible answers.

As we discuss in the following section, our model is arguably simpler and more intuitive than peer-prediction models. In short, we are interested in obtaining privately observed signals from agents, where such signals follow a categorical distribution with an unknown parameter. This unknown parameter represents the *population knowledge*; that is, it defines the distribution of agents' private signals.

Moreover, we explicitly assume that the concept of social projection holds true. As we mentioned before, social projection is a cognitive bias where an agent believes that his private answer is the most popular answer among his peers or, in other words, the answer most likely to be reported by a random peer from the population of agents. Thus, social projection serves as an egocentric heuristic for inductive reasoning where agents project themselves onto others. In this way, social projection is a strong form of the well-known psychological phenomenon called the *false-consensus effect* (Ross et al., 1977). We elaborate on such a connection in the Social Projection section.

Some formal models have been proposed to model different forms of social projection, for example, Brenner and Bilgin (2011) proposed a social projection model based on support theory (Tversky & Koehler, 1994), whereas Busemeyer and Pothos (2012) discussed how to interpret social projection using a quantum model (Pothos & Busemeyer, 2009). In this article, we propose a flexible social projection model that considers the strength of the projection within a Bayesian learning framework. This modeling choice is desirable because it has been shown that the strength of the projection is context dependent; for example, there is evidence that social projection exists in intrasocial groups and, to a less degree, in intersocial groups (Robbins & Krueger, 2005). Given our social projection model, we show that it is possible to induce honest reporting by simply comparing reported answers and rewarding agreements. Intuitively, the rationale behind this result is that when a risk-neutral agent believes that his private answer is the most popular answer, then honest reporting maximizes the likelihood of a random agreement.

Rewarding agents based on pairwise comparisons has been empirically proven to be an effective payment structure in different domains such as crowdsourcing and games with a purpose (Shaw, Horton, & Chen, 2011; von Ahn & Dabbish, 2008). For example, Huang and Fu (2013) showed that informing the agents that their rewards will be based on how similar their responses are to other agents' responses results in more accurate responses than telling the agents that their rewards will be based on how similar their responses are to gold-standard responses.

Our work adds to the existing body of literature by providing a theoretical justification, and further empirical evidence, on why payment structures based on pairwise comparisons work well in environments such as crowdsourcing. Moreover, our work provides a novel characterization of such payment structures different than, for example, a recent game-theoretic characterization by Waggoner and Chen (2013). Waggoner and Chen argued that rewarding agreements does not elicit honest answers. Instead, agents report the correct answer according to their common knowledge. We obtain a different result because, in contrast to Waggoner and Chen's work, we make assumptions on the nature of agents' information structure so as to model social projection.

The Model

We consider a multiple-choice question with a total of $n \geq 2$ exhaustive and mutually exclusive answers $A_1, \ldots, A_n$. This modeling choice is rather flexible because one can phrase many different tasks in terms of a multiple-choice question; for example, a categorization (labeling) task is a multiple-choice question where categories (labels) are the underlying answers. Similarly, classification tasks, content/sentiment analysis, and rating can be phrased in terms of multiple-choice questions.

We assume that the *population knowledge* is represented by an *unknown* categorical distribution Ω with parameter $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$, where $0 \leq \omega_k \leq 1$ and $\sum_{k=1}^{n} \omega_k = 1$. A possible interpretation of ω_k is that it is the probability that an agent selected at random from the population of agents has A_k as the answer to the multiple-choice question.

Each agent possesses a privately observed *signal* from Ω. We refer to observed signals as *hon-*

est answers. We denote the honest answer of an agent i by $t_i \sim \Omega$, where $t_i \in \{A_1, \ldots, A_n\}$. Honest answers are independent, that is, $P(t_i \mid t_j) = P(t_i)$. As mentioned before, agents are not necessarily honest when reporting their answers, for example, an agent might report an answer he believes is more socially desirable than his honest answer. Therefore, we distinguish between honest answers and *reported answers*. We say that agent i is *reporting honestly* when his reported answer $r_i \in \{A_1, \ldots, A_n\}$ is equal to his honest answer, that is, $r_i = t_i$.

A *requester* is responsible for eliciting the answers and for rewarding the agents. Let s_i be agent i's *reward* after he reports r_i. We discuss how to compute s_i in the following sections. Rewards are somehow coupled with relevant incentives, whether social-psychological, such as praise or visibility, or material rewards through prizes or money. We make five major assumptions in our model:

1. *Autonomy*: Agents cannot influence other agents' answers; that is, they do not know each other's identity, and they are not allowed to communicate with one another during the elicitation process.
2. *Risk neutrality*: Agents behave to maximize their expected rewards.
3. *Prior distributions*: Each agent has a prior distribution over $\boldsymbol{\omega}$, that is, $P(\boldsymbol{\omega})$.
4. *Posterior distributions*: Every agent i updates his prior after observing t_i, that is, $P(\boldsymbol{\omega} \mid t_i)$.
5. *Social projection*: Each agent i's posterior distribution satisfies the following inequality: $\mathbb{E}[\omega_x \mid t_i = A_x] > \mathbb{E}[\omega_y \mid t_i = A_x]$, for all $y \in \{1, \ldots, n\}$ such that $y \neq x$.

The first assumption means that agents work individually on the multiple-choice question. It describes how, for example, crowd workers traditionally solve tasks on the crowdsourcing platform Amazon Mechanical Turk. The second assumption means that agents are self-interested, and no external incentives exist for each agent. In other words, agents only care about their expected rewards, without considering what the requester will do with the reported answers. It is worth mentioning that there is related work that considers the case when agents are also interested in influencing a potential decision of a decision maker (e.g., Boutilier (2012); Chen & Kash (2011); Dimitrov & Sami (2010); Othman & Sandholm (2010)).

The third assumption means that each agent has prior information on the population knowledge. Such an assumption allows any agent to calculate the expected value of ω_k, $\mathbb{E}[\omega_k]$, which is a subjective estimate on how likely it is that an agent selected at random from the population of agents has A_k as the answer to the multiple-choice question. It is worth mentioning that, different from the original peer-prediction model (Miller et al., 2005), our model does not rely on the assumption that agents' prior distributions are the same or on common knowledge assumptions.

The fourth assumption means that each agent i updates his prior information on the population knowledge after observing his signal t_i. Furthermore, the posterior distributions are consistent with Bayesian updating, that is,

$$P(\boldsymbol{\omega} \mid t_i) = \frac{P(t_i \mid \boldsymbol{\omega}) P(\boldsymbol{\omega})}{P(t_i)}.$$

Given the above posterior distribution, $\mathbb{E}[\omega_k \mid t_i]$ is the subjective posterior estimate on how likely it is that an agent selected at random from the population of agents has A_k as the answer to the multiple-choice question. The probability vector $\mathbb{E}[\boldsymbol{\omega} \mid t_i] = (\mathbb{E}[\omega_1 \mid t_i], \mathbb{E}[\omega_2 \mid t_i], \ldots, \mathbb{E}[\omega_n \mid t_i])$ is referred to as agent i's *posterior predictive distribution* because it provides the distribution of others' answers given the observed signal t_i.

The fifth assumption means that every agent believes that his honest answer is the most popular answer among his peers. This is the main idea behind social projection, a subcase of the false-consensus effect (Ross et al., 1977), which is the tendency to expect similarities between oneself and others. We elaborate on social projection in the Social Projection section. We say that agent i's social projection is strong when he believes that his honest answer is more popular than all the other answers combined, that is, $\mathbb{E}[\omega_x \mid t_i = A_x] > \sum_{y \neq x} \mathbb{E}[\omega_y \mid t_i = A_x]$ or, equivalently, $\mathbb{E}[\omega_x \mid t_i = A_x] > 0.5$.

Clearly, social projection and strong social projection are equivalent when $n = 2$, that is, when the underlying multiple-choice question has only two possible answers. Note that we

model social projection by constraining the posterior distribution, which is equivalent to adding constraints to both the prior distribution and likelihood function. Constraining priors and likelihoods, however, would likely result in a more complex model.

It is worth mentioning how our model differs from the original peer-prediction model (Miller et al., 2005). When adapting the peer-prediction model to our setting, the following two conditions must be added: (a) the underlying multiple-choice question has an unknown type, that is, a correct answer; and (b) agents have a common prior distribution over the question's type. An agent's honest answer is then a signal of the question's type, where conditional on the question's type, agents' signals are independent and identically distributed. The conditional distribution of signals given types is common knowledge. We argue that our model is simpler and more intuitive than the original peer-prediction model. First, an extra conditional distribution is no longer necessary when agents reason in terms of the population knowledge, rather than in terms of the question's type. Furthermore, our model does not rely on common prior and common knowledge assumptions. Finally, the extra assumption that agents exhibit social projection is natural and possible to verify in practice, as we later discuss in the article.

Instantiating the Basic Model Using Dirichlet Distributions

For illustration's sake, we discuss in this section how to instantiate our basic model using Dirichlet distributions as prior and posterior distributions. The resulting model is similar to the model proposed by Carvalho and Larson (2012). We emphasize, however, that our basic model is not dependent on any specific distribution. For our purposes, the *Dirichlet distribution* is as a continuous distribution over parameter vectors of a categorical distribution. Because $\boldsymbol{\omega}$ is the unknown parameter of the categorical distribution that models population knowledge, it is natural to consider a Dirichlet distribution as a prior for $\boldsymbol{\omega}$. Given a *hyperparameter* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, which is a vector of positive reals that determines the shape of the Dirichlet distribution, the probability density function of the Dirichlet distribution over $\boldsymbol{\omega}$ is:

$$P(\boldsymbol{\omega} \mid \boldsymbol{\alpha}) = \frac{1}{\beta(\boldsymbol{\alpha})} \prod_{k=1}^{n} \omega_k^{\alpha_k - 1} \tag{1}$$

where

$$\beta(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{n} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{n} \alpha_k\right)}$$

and $\Gamma(\cdot)$ is the Gamma function. For the Dirichlet distribution in (1), the expected value of ω_k is $\mathbb{E}[\omega_k \mid \boldsymbol{\alpha}] = \boldsymbol{\alpha}_k / \sum_{x=1}^{n} \alpha_x$. As mentioned before, our model does not rely on the assumption of common priors. Hence, agents might have different hyperparameters $\boldsymbol{\alpha}$. We say that the Dirichlet distribution is *uninformative* (or symmetric) when all of the elements making up the vector α have the same value; that is, $\alpha_1 = \alpha_2 = \cdots = \alpha_n > 0$. Uninformative priors are used when there is no prior knowledge favoring one probability event over another. In our setting, the assumption of uninformative Dirichlet priors implies that before an agent sees the underlying multiple-choice question and knowing only the number of possible answers n, the agent's response to the query "what is the probability that one of your peers will report the answer A_k?" is $\frac{1}{n}$, for all $k \in \{1, \ldots, n\}$. That is, all the answers are equally likely a priori. Formally,

$$P(A_k \mid \boldsymbol{\alpha}) = \mathbb{E}[\omega_k \mid \boldsymbol{\alpha}] = \frac{\alpha_k}{\sum_{y=1}^{n} \alpha_y} = \frac{1}{n}.$$

An interesting property of the Dirichlet distribution is that it is the *conjugate prior* of the categorical distribution (Bernardo & Smith, 1994); that is, the posterior distribution $P(\boldsymbol{\omega} \mid \boldsymbol{\alpha}, t_i)$ is itself a Dirichlet distribution. This relationship is often used in Bayesian statistics to estimate hidden parameters of categorical distributions. To illustrate this point, suppose that agent i observes the signal $t_i = A_k$, for $k \in \{1, \ldots, n\}$. After applying Bayes's rule, agent i's posterior distribution is: $P(\boldsymbol{\omega} \mid \boldsymbol{\alpha}, t_i = A_k) = P(\boldsymbol{\omega} \mid (\alpha_1, \ldots, \alpha_k + 1, \ldots, \alpha_n))$. Then, agent i's posterior predictive distribution is:

$$\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, t_i = A_k] = \left(\frac{\alpha_1}{1 + \sum_{y=1}^{n} \alpha_y}, \ldots, \frac{\alpha_k + 1}{1 + \sum_{y=1}^{n} \alpha_y}, \ldots, \frac{\alpha_n}{1 + \sum_{y=1}^{n} \alpha_y} \right). \quad (2)$$

Given the posterior predictive distribution in (2), the assumption that agents exhibit social projection means that $\mathbb{E}[\omega_k \mid \boldsymbol{\alpha}, t_i = A_k] > \mathbb{E}[\omega_y \mid \boldsymbol{\alpha}, t_i = A_k]$, for all $y \in \{1, \ldots, n\}$ such that $y \neq k$, which implies that $\alpha_k + 1 > \alpha_y$ or, alternatively, $\alpha_y - \alpha_k < 1$. Strong social projection then means that $\alpha_k + 1 > \sum_{y \neq k} \alpha_y$, that is, $\sum_{y \neq k} \alpha_y - \alpha_k < 1$. Under the extra assumption of uninformative Dirichlet priors, the elements of the probability vector in (2) become:

$$\mathbb{E}[\omega_x \mid \boldsymbol{\alpha}, t_i = A_k] = \begin{cases} \frac{1}{n} + z & \text{if } x = k, \\ \frac{1}{n} - \frac{z}{n-1} & \text{otherwise} \end{cases}$$

for $0 < z = \frac{n-1}{n + \alpha_k \cdot n^2} < 1$. The above equation illustrates that after solving the multiple-choice question, an agent's response to the query, "what is the most popular answer?" is skewed toward the agent's honest answer. Furthermore, the value of z and, consequently, the elements $\alpha_1, \ldots, \alpha_n$ of the hyperparameter $\boldsymbol{\alpha}$ determine the strength of the social projection, where the higher the value of z, the stronger the social projection effect. We note that this Dirichlet model of social projection was first hinted by Carvalho, Dimitrov, and Larson (2014).

Because of their flexibility and powerful theoretical properties, Dirichlet distributions have been used in many different domains, for example, to model reputation in social networks (Regan, Poupart, & Cohen, 2006), to aid in the detection and recognition of objects in visual scenes (Torralba, Willsky, Sudderth, & Freeman, 2005), in recommendation systems (Chien & George, 1999), to cluster different agents based on their individual differences (Navarro et al., 2006), in natural language processing (Vlachos, Korhonen, & Ghahramani, 2009), and so on.

Social Projection

An important assumption in our model is that agents exhibit social projection. Social projection, as defined in our work, is a strong form of the psychological concept known as the false-consensus effect (Ross et al., 1977). False consensus refers to an egocentric bias that occurs when agents estimate consensus for their own behaviors or beliefs. Specifically, the false-consensus hypothesis asserts that human beings who engage in a certain behavior, or hold a certain belief, estimate that behavior/belief to be more common than the estimations from peers who engage in alternative behaviors or hold different beliefs.

Traditionally, the false-consensus effect is not restricted to cases where agents believe that the majority share their values. In other words, the term false consensus has also been used when there is no consensus, that is, when agents do not necessarily believe that the majority of others share their views but their estimates of the number of agents who share their views tend to exceed the actual number. Our concept of social projection, on the other hand, asserts that an agent believes that his honest answer is the most popular answer among his peers; that is, there is actually a consensus around the agent's honest answer, which might be a false consensus. It is important to note that the term social projection has sometimes been used to describe the mechanism behind the false-consensus effect (e.g., see the work by Marks, Graham, & Hansen, 1992). Our definition of social projection, however, refers to a specific subcase of the false-consensus effect where each agent projects a consensus around his private information.

A pioneer work on false consensus was performed by Katz and Allport (1931), who noticed that the more students admitted that they had cheated on an exam, the more they expected that other students cheated too. Ross et al. (1977), the authors who coined the term false consensus, hypothesized and demonstrated that human beings tend to overestimate the popularity of their own beliefs and preferences. In particular, the results of Studies 1, 3, and 4 in the article by Ross et al. (1977) showed that social projection, as defined in the beginning of this section, holds true when considering the agents' aggregated answers

and estimates, whereas the combined results of Study 2 showed that social projection was present in 5 out of 14 multiple-choice questions.

Since the work by Ross et al. (1977), dozens of studies have documented a systematic relationship between one's perceptions of his own characteristics and his estimates of the percentage of people in the population who share those characteristics in a variety of settings, ranging from questionnaire studies presenting situations, choices, and judgments that are hypothetical to actual conflict situations demanding personally relevant behavioral choices and social judgments (Marks & Miller, 1987; Mullen et al., 1985; Robbins & Krueger, 2005). Moreover, it has been shown that the strength of the false-consensus effect is highly dependent on the underlying agents and task; for example, there is some evidence that elderly people display a higher degree of false consensus than adolescents (Yinon, Mayraz, & Fox, 1994). Furthermore, false consensus is stronger when people make judgments about ingroups than when they make judgments about outgroups (Robbins & Krueger, 2005).

The emergence of the false-consensus effect has been explained under four different perspectives (Marks & Miller, 1987): (a) selective exposure and cognitive availability; (b) salience and focus of attention; (c) motivational processes; and (d) logical information processing.

The perspective of selective exposure and availability suggests that instances of similarity between oneself and others are more readily available from memory than instances of dissimilarity, thereby increasing estimates of consensus for one's preferred opinion. The reason why these instances of similarity are more readily available is due to selective exposure, that is, due to the tendency of human beings to associate with those who are similar rather than dissimilar to themselves. This selective exposure to similar others provides an agent with a biased and restricted sample of information about the population's true diversity of opinion.

The perspective about salience and focus of attention suggests that false consensus arises when an agent focuses attention exclusively on a single position. Consequently, perceived consensus may be augmented because that position is the only one in the agent's immediate consciousness.

The motivational perspective asserts that false consensus has a dissonance-reduction function, where agents unconsciously project their own beliefs onto others to obtain confirmation for their own attitudes and beliefs, that is, to bolster social support, to validate the correctness or appropriateness of a belief, to maintain self-esteem, and so forth.

Finally, the logical information-processing perspective views active reasoning and rational processes as underlying one's estimates about the similarity between oneself and others. Within this perspective, Dawes (1989) reexamined the data obtained by Ross et al. (1977) and argued, from a Bayesian perspective, that subjects were correct in considering their own behavioral choices common in the population. In particular, Dawes (1989) showed that the typical strength of the false-consensus effect was similar to the statistically normative change from prior to posterior probabilities. Moreover, Dawes (1989) discussed that even a sample of size 1, representing an agent's private information as in our model, should have substantial effect on the agent's consensus estimates. Therefore, it is conceivable that agents intuitively understand the logic of statistical induction and perform accordingly.

To summarize, our concept of social projection is strongly grounded in psychology. In particular, social projection is a strong form of the psychological concept known as the false-consensus effect. As the seminal work by Ross et al. (1977) suggests, our concept of social projection happens often and in many different settings when the underlying agents are human beings. There are other studies that also suggest that social projection exists when individuals are faced with multiple-choice questions. For example, the results by Sherman, Presson, Chassin, Corty, and Olshavsky (1983) show that social projection holds true when nonsmokers predict the smoking prevalence among adolescents as well as when smokers predict the smoking prevalence among adults. Our concept of social projection also finds support in the results by Bauman and Geher (2002) involving multiple-choice questions about the acceptance/rejection of death penalty, abortion, legalization of drugs, condom distribution in schools, and lower drinking age.

Inducing Honest Reporting by Making Pairwise Comparisons

If the requester knew a priori agents' honest answers, he could then compare the honest answers to the reported answers and reward agreement. However, because of the subjective nature of our setting, the requester faces a situation where this objective truth is practically unknowable. Our solution to this issue is to induce honest reporting by providing rewards based on pairwise comparisons of reported answers. It is important to note that our results in this section are valid for any prior/posterior distributions as long as social projection holds true, and not only for the instantiation of our model using Dirichlet distributions. Recall that s_i is the reward agent i receives after he reports r_i. We first consider the following payment function:

$$s_i = \tau(r_i, r_j) = \begin{cases} v_{max} & \text{if } r_i = r_j, \\ v_{min} & \text{otherwise} \end{cases} \tag{3}$$

where $v_{max} > v_{min}$ and $j \neq i$ is a randomly selected agent. That is, agent i receives the maximum payment $v_{max} \in \mathbb{R}$ if and only if he reports an answer equal to the answer reported by another agent j randomly selected from the population of agents. We show below that such a payment function induces honest reporting by risk-neural agents in our setting.

> *Proposition 1:* Each agent i strictly maximizes his expected reward from the payment function in (3) if and only if $r_i = t_i$.

Proof. Given the autonomy assumption, agent j cannot influence agent i's reported answer, and vice versa. Further, given the assumption that agents are risk neutral, agent i behaves so as to maximize his expected reward, which is equal to:

$$\mathbb{E}[\tau(r_i, r_j)] = v_{max} \times P(r_j = r_i \,|\, t_i) + \sum_{A_x \neq r_i} v_{min} \times P(r_j = A_x \,|\, t_i).$$

Given that v_{max} and v_{min} are fixed, agent i maximizes the above equation by moving as much probability mass toward v_{max} as possible. Suppose that agent i observes the signal $t_i = A_k$, for some $k \in \{1, \ldots, n\}$. The assumption of social projection then implies that:

$$P(r_j = A_k \,|\, t_i = A_k) = \mathbb{E}[\omega_k \,|\, t_i = A_k] > \mathbb{E}[\omega_x \,|\, t_i = A_k] = P(r_j = A_x \,|\, t_i = A_k),$$

for all $x \in \{1, \ldots, n\}$ such that $x \neq k$. Consequently, agent i strictly maximizes his expected reward if and only if he is reporting honestly, that is, when he reports $r_i = t_i = A_k$. □

The above result means that each agent determines, according to his posterior predictive distribution, the answer most likely to be reported by a random peer. This answer turns out to be the agent's honest answer in the presence of social projection. There are two interesting points regarding the above result. First, because honest answers are independent, Proposition 1 is still valid when the requester makes multiple pairwise comparisons. For example, consider a set of agents N such that $i \notin N$. Then, the payment function $s_i = \sum_{j \in N} \tau(r_i, r_j)$ still induces honest reporting. We argue that there are practical benefits in making multiple pairwise comparisons because it reduces the influence of "unlucky" pairwise comparisons on agents' rewards, in a sense that the requester might reduce the potential variance in the agents' rewards. Second, Radanovic and Faltings (2013) suggested a result similar to Proposition 1 for the specific case when $v_{max} = 1$ and $v_{min} = 0$, where our concept of social projection was then called the "self-dominant assumption."

Taking the Distance Between Reported Answers Into Account

Rewarding agreements based solely on whether two reported answers are equal to each other might work well for small values of n, the total number of answers, but it can be too restrictive and to some degree unfair when n is high and the underlying answers are ordered, that is, when the answer A_k is closer to A_{k+1} than to A_{k+2}. For example, consider the case when the answers are numerical values where $A_1 = 1, A_2 = 2, \ldots, A_n = n$. If $n = 5$ and $r_j = A_5 = 5$, then a reported answer $r_i = A_4 = 4$ seems to be more accurate than a reported answer equal to $r_i = A_1 = 1$. A natural payment function that takes the distance between two reported answers into account is:

$$s_i = \tau(r_i = A_x, r_j = A_y) = v_1 - v_2 \times |x - y|, \quad (4)$$

where v_1 and v_2 are two constants. That is, the payment function in (4) penalizes disagreements based on how distant the reported answers are from each other, which in turn is measured in terms of the absolute difference between the indices of the answers. This payment function is motivated by the ranked probability scoring function, which we describe in the next section. We show below that such a payment function induces honest reporting in the presence of strong social projection.

Proposition 2: Under strong social projection, each agent i strictly maximizes his expected reward from the payment function in (4) if and only if $r_i = t_i$.

Proof. Without loss of generality, assume that $t_i = A_k$. Moreover, because v_1 and v_2 are only scaling agent i's reward, we can assume in our proof that $v_1 = 0$ and $v_2 = 1$. For ease of notation, let agent i's posterior predictive distribution be $(p_1, p_2, \ldots, p_n) = (\mathbb{E}[\omega_1 \mid t_i], \mathbb{E}[\omega_2 \mid t_i], \ldots, \mathbb{E}[\omega_n \mid t_i])$. Then, the expected reward of agent i when he reports honestly is:

$$\begin{aligned} &-p_1 \times (k-1) - p_2 \times (k-2) - \ \ldots \\ &- p_{k-1} \times 1 - p_k \times 0 - p_{k+1} \times 1 - \ \ldots \\ &- p_n \times (n-k). \end{aligned} \quad (5)$$

We analyze two different cases to determine the optimal report r_i under the payment function in (4): first, when agent i shifts his reported answer to the right of his honest answer, that is, when $r_i = A_{k+x}$, for $k + x \leq n$ and $x > 0$; and second, when agent i shifts his reported answer to the left of his honest answer, that is, when $r_i = A_{k-x}$, for $k - x \geq 1$ and $x > 0$. We show that in both cases, agent i would receive a higher expected reward by reporting his honest answer, that is, when $r_i = t_i = A_k$.

Case 1: $r_i = A_{k+x}$

When agent i reports $r_i = A_{k+x}$, his expected reward is:

$$\begin{aligned} &-p_1 \times (k+x-1) - p_2 \times (k+x-2) - \ \cdots \\ &-p_{k-1} \times (x+1) - p_k \times x - p_{k+1} \times (x-1) \\ &- \ \cdots - p_{k+x-1} \times 1 - p_{k+x} \times 0 - p_{k+x+1} \times 1 \\ &- \ \cdots - p_n \times (n-k-x). \end{aligned} \quad (6)$$

We note that the difference between (5) and (6) is at least $x \times (\sum_{y=1}^{k-1} p_y + p_k - \sum_{y=k+1}^{n} p_y)$. Under the assumption that agents exhibit strong social projection, $p_k > \sum_{y \neq k} p_y$, which implies that $x \times (\sum_{y=1}^{k-1} p_y + p_k - \sum_{y=k+1}^{n} p_y) > 0$. Consequently, honest reporting is more profitable in expectation than shifting the reported answer to the right of the honest answer.

Case 2: $r_i = A_{k-x}$

When agent i reports $r_i = A_{k-x}$, his expected reward is:

$$\begin{aligned} &-p_1 \times (k-x-1) - p_2 \times (k-x-2) - \ \cdots \\ &-p_{k-x-1} \times 1 - p_{k-x} \times 0 - p_{k-x+1} \times 1 - \ \cdots \\ &-p_{k-1} \times (x-1) - p_k \times x - p_{k+1} \times (x+1) \\ &- \ \cdots - p_n \times (n-k+x). \end{aligned} \quad (7)$$

We note that the difference between (5) and (7) is at least $x \times (-\sum_{y=1}^{k-1} p_y + p_k + \sum_{y=k+1}^{n} p_y)$. Under the assumption that strong social projection holds true, $p_k > \sum_{y \neq k} p_y$, which implies that $x \times (-\sum_{y=1}^{k-1} p_y + p_k + \sum_{y=k+1}^{n} p_y) > 0$. Consequently, honest reporting is more profitable in expectation than shifting the reported answer to the left of the honest answer. □

We note that if strong social projection does not hold true, the payment function in (4) does not necessarily induce honest reporting in our setting. For example, consider the scenario where the underlying multiple-choice question has three possible answers ($n = 3$). Moreover, agent i's honest answer is $t_i = A_1$ and his posterior predictive distribution is $\mathbb{E}[\boldsymbol{\omega} \mid t_i = A_1] = (0.4, 0.3, 0.3)$. This setting satisfies social projection but not strong social projection. If agent i reports honestly, that is, $r_i = t_i = A_1$, he then obtains an expected reward equal to $v_1 - v_2 \times (0.4 \times 0 + 0.3 \times 1 + 0.3 \times 2) = v_1 - 0.9 \times v_2$. On the other hand, if agent i misreports by reporting $r_i = A_2$, his expected reward becomes $v_1 - v_2 \times (0.4 \times 1 + 0.3 \times 0 + 0.3 \times 1) = v_1 - 0.7 \times v_2$, hence being more profitable in expectation than honest reporting. We also note that, similarly to Proposition 1, Proposition 2 is still valid when the requester makes multiple pairwise comparisons.

Deriving Previous Results in Terms of Proper Scoring Rules

In this section, we discuss how to derive Proposition 1 and 2 in terms of proper scoring rules (Winkler & Murphy, 1968). We argue that such a derivation is of theoretical value to show the robustness of our results. We note, however, that readers less interested in theoretical results can skip this section without loss of continuity. To derive our previous results in terms of proper scoring rules, we make stronger assumptions than the assumptions in our basic model. In particular, we assume that agents have common uninformative Dirichlet priors (see the section Instantiating the Basic Model Using Dirichlet Distributions); that is, all the agents have the same hyperparameter $\boldsymbol{\alpha}$, and all the elements of $\boldsymbol{\alpha}$ are the same. Furthermore, we assume that this is common knowledge. Together, these extra assumptions imply that belief updating can be expressed as an updating of the parameters of the prior distribution, and the requester can estimate agents' posterior predictive distributions based solely on their reported answers, a point that is explored by our payment structure.

Proper Scoring Rules

Consider a set of exhaustive and mutually exclusive outcomes $\{\theta_1, \ldots, \theta_n\}$, and a probability vector $\mathbf{q} = (q_1, \ldots, q_n)$, where q_k is the probability value associated with the occurrence of outcome θ_k. A *scoring rule* $R(\mathbf{q}, \theta_e)$ is a function that provides a score for the assessment $\mathbf{q}$ on observing the outcome θ_e, for $e \in \{1, \ldots, n\}$. A scoring rule is called *strictly proper* when an agent receives his maximum expected score if and only if his stated assessment $\mathbf{q}$ corresponds to his true assessment $\mathbf{p} = (p_1, \ldots, p_n)$ (Winkler & Murphy, 1968). The expected score of $\mathbf{q}$ at $\mathbf{p}$ for a real-valued scoring rule $R(\mathbf{q}, \theta_e)$ is:

$$\mathbb{E}_{\mathbf{p}}[R(\mathbf{q}, \theta_e)] = \sum_{e=1}^{n} p_e \, R(\mathbf{q}, \theta_e).$$

Proper scoring rules have been used to promote honest reporting in a variety of domains, for example, to incentivize agents to accurately estimate their own efforts to accomplish a task (Bacon et al., 2012), in financial markets set to aggregate agents' subjective probabilities (Hanson, 2003), and so forth. Some of the best known strictly proper scoring rules, together with their scoring ranges, are:

$$\text{logarithmic: } R(\mathbf{q}, \theta_e) = \log \; q_e \quad (-\infty, 0]$$

$$\text{quadratic: } R(\mathbf{q}, \theta_e) = 2q_e - \sum_{k=1}^{n} q_k^2 \quad [-1, 1]$$

$$\text{spherical: } R(\mathbf{q}, \theta_e) = \frac{q_e}{\sqrt{\sum_{k=1}^{n} q_k^2}} \quad [0, 1]$$

The above strictly proper scoring rules are all *symmetric,* in a sense that $R((q_1, \ldots, q_n), \theta_e) = R((q_{\pi_1}, \ldots, q_{\pi_n}), \theta_{\pi_e})$, for all probability vectors $\mathbf{q} = (q_1, \ldots, q_n)$, for all permutations π on n elements, and for all outcomes $\theta_1, \ldots, \theta_n$. We say that a scoring rule is *bounded* if $R(\mathbf{q}, \theta_e) \in \mathbb{R}$, for all probability vectors $\mathbf{q}$ and $e \in \{1, \ldots, n\}$. For example, the logarithmic scoring rule is not bounded because it might return $\log 0 = -\infty$, whenever the probability vector $\mathbf{q}$ contains a probability value equal to zero. On the other hand, both the quadratic and the spherical scoring rules are bounded.

Proposition 3 shows a well-known property of strictly proper scoring rules, namely that they are still strictly proper under positive affine transformations (Gneiting & Raftery, 2007). That is, $\arg\max_{\mathbf{q}} \mathbb{E}_{\mathbf{p}}[\gamma R(\mathbf{q}, \theta_e) + \lambda] = \arg\max_{\mathbf{q}} \mathbb{E}_{\mathbf{p}}[R(\mathbf{q}, \theta_e)] = \mathbf{q}$, for a strictly proper scoring rule R, $\gamma > 0$, and $\lambda \in \mathbb{R}$. As a consequence, the range of a bounded proper scoring rule can be easily changed.

> *Proposition 3:* If $R(\mathbf{q}, \theta_e)$ is a strictly proper scoring rule, then a positive affine transformation of R, that is, $\gamma R(\mathbf{q}, \theta_e) + \lambda$, for $\gamma > 0$ and $\lambda \in \mathbb{R}$, is also strictly proper.

Computing Agents' Rewards Using Proper Scoring Rules

The first step toward computing an agent i's reward using proper scoring rules is to estimate his posterior predictive distribution $\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, t_i]$ based on his reported answer r_i. Let $\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i] = (\mathbb{E}[\omega_1 \mid \boldsymbol{\alpha}, r_i], \ldots, \mathbb{E}[\omega_n \mid \boldsymbol{\alpha}, r_i])$ be such an estimation. The assumptions of common uninformative Dirichlet priors and common knowledge allow the requester to simulate agent i's belief updating process, that is:

$$\mathbb{E}[\omega_k \mid \boldsymbol{\alpha}, r_i] = \begin{cases} \dfrac{\alpha_k + 1}{1 + \sum_{x=1}^{n} \alpha_x} & \text{if } r_i = A_k, \\ \dfrac{\alpha_k}{1 + \sum_{x=1}^{n} \alpha_x} & \text{otherwise.} \end{cases} \quad (8)$$

Recall that the elements of agent i's true posterior predictive distribution in (2) are defined as:

$$\mathbb{E}[\omega_k \mid \boldsymbol{\alpha}, t_i] = \begin{cases} \dfrac{\alpha_k + 1}{1 + \sum_{x=1}^{n} \alpha_x} & \text{if } t_i = A_k, \\ \dfrac{\alpha_k}{1 + \sum_{x=1}^{n} \alpha_x} & \text{otherwise.} \end{cases}$$

Clearly, $\mathbb{E}[\omega_k \mid \boldsymbol{\alpha}, r_i] = \mathbb{E}[\omega_k \mid \boldsymbol{\alpha}, t_i]$ if and only if agent i is reporting honestly, that is, when he reports $r_i = t_i$. The reward of agent i is then determined as follows:

$$s_i = \gamma R(\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i], r_j) + \lambda \quad (9)$$

where γ and λ are constants, for $\gamma > 0$ and $\lambda \in \mathbb{R}$, $j \neq i$ is a randomly selected agent, and R is a strictly proper scoring rule. Scoring rules require an observable outcome, or a reality, to score an assessment. Intuitively, the scoring method in (9) considers the answer reported by a random agent j as the observed outcome, and rewards agent i's estimated posterior predictive distribution in (8) as an assessment of the observed outcome. In the following proposition, we show that the payment structure in (9) induces honest reporting by risk-neutral agents when social projection holds true.

> *Proposition 4:* Each agent i strictly maximizes his expected reward from the payment function in (9) if and only if $r_i = t_i$.

Proof. For ease of notation, let $\mathbf{p} = \mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, t_i]$ and $\mathbf{q} = \mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i]$.

(If part) Since R is a strictly proper scoring rule, from Proposition 3 we have that:

$$\arg\max_{\mathbf{q}} \mathbb{E}_{\mathbf{p}}[\gamma R(\mathbf{q}, r_j) + \lambda] = \mathbf{p}.$$

If $r_i = t_i$, then by construction $\mathbf{q} = \mathbf{p}$, that is, the estimated posterior predictive distribution in (8) is equal to the true posterior predictive distribution in (2). Consequently, honest reporting maximizes agents' expected rewards.

(Only if part) Using a similar argument, given that R is a strictly proper scoring rule, from Proposition 3 we have that:

$$\arg\max_{\mathbf{q}} \mathbb{E}_{\mathbf{p}}[\gamma R(\mathbf{q}, r_j) + \lambda] = \mathbf{p}.$$

By construction, $\mathbf{q} = \mathbf{p}$ if and only if $r_i = t_i$ (see Equations 2 and 8). Thus, agents' expected rewards are maximized only when agents are reporting honestly. □

Intuitively, an agent's true posterior predictive distribution in (2) is equal to his estimated posterior predictive distribution in (8) when the agent is reporting honestly. Consequently, the expected score resulting from a strictly proper scoring rule is strictly maximized when the expectation is taken with respect to the agent's true posterior predictive distribution. We show in the following subsections that natural interpretations of the scoring method in (9) arise depending on the underlying strictly proper scoring rule. In particular, we discuss two different interpretations: (a) when R is a symmetric and bounded strictly proper scoring rule; and (b) when R is the strictly proper scoring rule sensitive to distance called *ranked probability scoring rule*.

Rewarding Agreements

Because each agent's prior distribution is an uninformative Dirichlet prior, the elements of the agent's true and estimated posterior predictive distributions can take on only two possible values (see Equations 2 and 8 for $\alpha_1 = \cdots = \alpha_n$). Consequently, if R is a symmetric scoring rule, then the term $R(\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i], r_j)$ in (9) can take on only two possible values because a permutation of similar probability values in $\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i]$ does not change the score from a symmetric scoring rule. When R is also strictly proper, then $R(\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i], r_j) = \delta_{max}$, when $r_i = r_j$, and $R(\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i], r_j) = \delta_{min}$, when $r_i \neq r_j$, where $\delta_{max} > \delta_{min}$. Thus, the payment function in (9) can be written as:

$$s_i = \gamma R(\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i], r_j) + \lambda = \begin{cases} \gamma\delta_{max} + \lambda & \text{if } r_i = r_j, \\ \gamma\delta_{min} + \lambda & \text{otherwise.} \end{cases}$$

When R is also bounded, the requester can set $\gamma = \frac{v_{max} - v_{min}}{\delta_{max} - \delta_{min}}$ and $\lambda = \frac{v_{min} \times \delta_{max} - v_{max} \times \delta_{min}}{\delta_{max} - \delta_{min}}$, for $v_{max} > v_{min}$, and agent i's reward becomes:

$$s_i = \tau(r_i, r_j) = \begin{cases} v_{max} & \text{if } r_i = r_j, \\ v_{min} & \text{otherwise} \end{cases}.$$

This is precisely what constitutes the payment function in (3). Hence, we obtain an intuitive interpretation of the scoring method in (9): Whenever two reported answers are equal to each other, the underlying agents receive a reward of v_{max}, otherwise they receive a reward of v_{min}.

Strictly Proper Scoring Rules Sensitive to Distance

We now show how to derive the payment function in (4) by using a strictly proper scoring rule in (9) that is *sensitive to distance*. Using the notation of the Proper Scoring Rules section, recall that $\mathbf{q} = (q_1, \ldots, q_n)$ is a probability vector over a set of outcomes $\theta_1, \ldots, \theta_n$. Assuming that the outcomes are ordered, we denote the cumulative probabilities by capital letter: $Q_k = \sum_{j \le k} q_j$. We first define the notion of distance between two probability vectors as proposed by Staël von Holstein (1970). We say that a probability vector $\mathbf{q}'$ is more distant from the jth outcome than a probability vector $\mathbf{q} \neq \mathbf{q}'$ if:

$$\begin{aligned} Q'_k &\ge Q_k, \text{ for } k = 1, \ldots, j-1 \\ Q'_k &\le Q_k, \text{ for } k = j, \ldots, n \end{aligned}.$$

Intuitively, the above definition means that $\mathbf{q}$ can be obtained from set $\mathbf{q}'$ by successively moving probability mass toward the jth outcome from other outcomes (Staël von Holstein, 1970). A scoring rule R is said to be sensitive to distance if $R(\mathbf{q}, \theta_j) > R(\mathbf{q}', \theta_j)$, whenever $\mathbf{q}'$ is more distant from $\mathbf{q}$ for all j. Epstein (1969) introduced the *ranked probability score* (RPS), a strictly proper scoring rule that is sensitive to distance. Using the formulation of Epstein's result proposed by Murphy (1970), we have for a probability vector $\mathbf{q}$ and an observed outcome θ_j, for $j \in \{1, \ldots, n\}$:

$$RPS(\mathbf{q}, \theta_j) = -\sum_{k=1}^{j-1} Q_k^2 - \sum_{k=j}^{n} (1 - Q_k)^2. \quad (10)$$

The above formulation of RPS is defined in terms of summations of quadratic scores. Jose, Nau, and Winkler (2009) extended the above formulation to other forms of scores, such as logarithmic scores. When using RPS as the strictly proper scoring rule in (9), agents receive rewards based on how close their reported answers are to the answers taken as observed outcomes. We show next that such a closeness measure has a very natural interpretation when agents' priors are uninformative Dirichlet priors with $\alpha_1 = \alpha_2 = \ldots = \alpha_n = \epsilon$, for an arbitrarily small constant $\epsilon > 0$. In this scenario, each element of the posterior predictive distribution in (2) is:

$$\mathbb{E}[\omega_k \mid \boldsymbol{\alpha}, t_i] = \begin{cases} \dfrac{\varepsilon + 1}{1 + \sum_{x=1}^{n} \varepsilon} & \text{if } t_i = A_k, \\ \dfrac{\varepsilon}{1 + \sum_{x=1}^{n} \varepsilon} & \text{otherwise.} \end{cases}$$

For a sufficiently small ϵ, the above values are approximately:

$$\mathbb{E}[\omega_k \mid \boldsymbol{\alpha}, t_i] \approx \begin{cases} 1 & \text{if } t_i = A_k, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, each element of the estimated posterior predictive distribution in (8) is approximately:

$$\mathbb{E}[\omega_k \mid \boldsymbol{\alpha}, r_i] \approx \begin{cases} 1 & \text{if } r_i = A_k, \\ 0 & \text{otherwise.} \end{cases}$$

Under the above circumstances, when R in (9) is the ranked probability scoring rule, the term $R(\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i], r_j)$ becomes:

$$RPS(\mathbb{E}[\boldsymbol{\omega} \mid \boldsymbol{\alpha}, r_i], r_j) \approx -\sum_{k=1}^{j-1} \left(\sum_{x \le k} \mathbb{E}[\omega_x \mid \boldsymbol{\alpha}, r_i] \right)^2 - \sum_{k=j}^{n} \left(1 - \sum_{x \le k} \mathbb{E}[\omega_x \mid \boldsymbol{\alpha}, r_i] \right)^2 = -|i - j|.$$

That is, a score from R is approximately equal to the negative of the absolute difference between the indices of two reported answers. Consequently, the scoring method in (9) can be written as:

$$s_i = \tau(r_i = A_x, r_j = A_y) \approx \gamma - \lambda \times |x - y|$$

which is precisely what constitutes the payment function in (4). Hence, the scoring method in (9) has a very natural interpretation in this setting: the requester penalizes agents in proportion to how far their reported answers are from each other.

Experiments

In this section, we describe a content-analysis experiment designed to test the effectiveness of payment structures based on pairwise comparisons between agents' reports. In the following subsections, we discuss Amazon Mechanical Turk, the platform used in our experiments, the experimental design, and the experimental results.

Amazon Mechanical Turk

Amazon Mechanical Turk[1] (AMT) is an online labor market originally developed for human computation tasks, that is, tasks that are relatively easy for human beings, but nonetheless challenging or even currently impossible for computers, for example, human debate-side classification (Walker et al., 2012), sentiment analysis (da Silva et al., 2014; Mohammad, 2012), audio transcription, filtering adult content, extracting data from images, and so forth. Some studies have shown that AMT can effectively collect valid data in these settings (Marge et al., 2010; Snow, O'Connor, Jurafsky, & Ng, 2008).

More recently, AMT has also been used as a platform for conducting behavioral experiments (Mason & Suri, 2012). One of the advantages that it offers to researchers is the access to a large, diverse, and stable pool of people willing to participate in the experiments for relatively low pay, thus simplifying the recruitment process and allowing a faster iteration between developing theory and executing experiments. Furthermore, AMT provides an easy-to-use built-in mechanism to pay workers that greatly reduces the difficulties of compensating individuals for their participation in the experiments and a built-in reputation system that helps requesters distinguish between good and bad workers and, consequently, to ensure data quality. Numerous studies have shown that results of behavioral studies conducted on AMT are comparable to results obtained in other online domains and in offline settings (Buhrmester, Kwang, & Gosling, 2011; Horton, Rand, & Zeckhauser, 2011), thus, providing evidence that AMT is a valid means of collecting behavioral data.

Experimental Design

We designed a content-analysis task on AMT that required workers, henceforth referred to as agents, to review three short texts under three different criteria: grammar, clarity, and relevance. The first two texts were extracts from published poems but with some original words intentionally replaced by misspelled words. The third text contained random words presented in a semistructured way. All the details regarding the texts are included in the Appendix. For each text, we presented three multiple-choice questions to the agents, each one having three possible answers ordered in decreasing negativity order:

- *Grammar:* Does the text contain misspellings, syntax errors, etc.?
  - —A lot of grammar mistakes
  - —A few grammar mistakes
  - —No grammar mistakes
- *Clarity:* Does the text, as a whole, make any sense?
  - —The text does not make sense
  - —The text makes some sense
  - —The text makes perfect sense
- *Relevance:* Could the text be part of a poem related to love?
  - —The text cannot be part of a love poem
  - —The text might be part of a love poem
  - —The text is definitely part of a love poem

We intentionally designed subjective answers so as to emphasize the subjective nature of content-analysis tasks. We translated each reported answer into a numeral value inside the set $\{0, 1, 2\}$. The most negative answer received the score 0, the middle answer received the score 1, and the most positive answer received the score 2. Thus, each agent reported a vector of 9 numerical values (3 texts $\times$ 3 criteria). No

[1] See http://www.mturk.com/.

additional measures or conditions were collected beyond what we report in this section.

We recruited 150 agents on AMT, all of them residing in the United States and older than 18 years old, from the general pool of workers, as opposed to the pool of workers with Masters qualification. We asked each agent to accomplish the task in at most 20 minute. We split the agents into three groups of equal size. Carvalho et al. (2015) suggested that there is no significant benefit in terms of expected accuracy when hiring more than seven workers on AMT. To be on the safe side, we overestimated such a number by using a predetermined sample size of 50 workers. After accomplishing the task, every agent in every group received a payment of \$0.20. Ipeirotis (2010) showed that more than 90% of the tasks on AMT have a baseline payment less than \$0.10, and 70% of the tasks have a baseline payment less than \$0.05. Thus, our baseline payment was much higher than the majority of payments from other tasks posted on Amazon Mechanical Turk.

We randomly assigned each agent to one of the three groups. Agents in two of the groups, the treatment groups, could earn an additional bonus of up to \$0.10. We informed the agents in the first treatment group, henceforth referred to as the *bonus group* (BG), that their bonuses would be proportional to the number of answers similar to their reported answers. Agents in the second treatment group, called the *bonus and information group* (BIG), received similar information, but they also received a short summary of some theoretical results presented in this article:

> A group of researchers from the University of Waterloo (Canada) formally showed that the best strategy to maximize your expected bonus in this setting is by being honest, i.e., by considering each question thoroughly and deciding the best answers according to your personal opinion.

It is often the case in empirical works that the authors do not fully explain to the agents the mathematics behind the payment structure used to induce honest reporting. Instead, some authors resort to describe in plain text the properties behind the underlying payment structure. For example, Weaver and Prelec (2013) said the following in their application of the Bayesian Truth Serum method:

> We explained that "BTS scoring" assigns scores to survey answers in a way that rewards honesty and that although we cannot know true private opinions or beliefs, and questions about such matters have no objectively correct answer, respondents nevertheless score higher on average by telling the truth. To lend credence to these claims, we reported that the method was recently invented by an MIT professor, and published in the academic journal Science.

The main reason behind using the BIG group was to study whether such a priming effect would have any impact on the accuracy of agents' reported answers beyond the use of pairwise comparisons alone. It is important to note that for the BIG group, we are measuring the effectiveness of the payment structure in conjunction with the truthfulness statement and not the effectiveness of the truthfulness statement alone. One cannot take our experimental results as a measure of the effectiveness of the truthfulness statement alone because the truthfulness statement would be confounded with the payment structure. Last, members of the third group, henceforth called the *control group* (CG), neither received extra explanations nor bonuses. The underlying payment structure was thus a *payment per completed task,* the canonical payment structure on AMT. We used CG members' reported answers as the control condition.

We computed bonuses by rewarding agreements. Recall that each agent reported nine answers (3 texts $\times$ 3 criteria). For each answer reported by an agent i, we calculated the total number of agreements (# agreements) between agent i's reported answer and the answers reported by members of agent i's group. Consequently, for each reported answer, there could be at most 49 similar reported answers because each group had 50 members. We then used the formula $\frac{10}{9} \times \frac{\#\text{agreements}}{49}$ to calculate the reward for an individual answer. Such a payment structure can be seen as a positive affine transformation of the scoring method in (3) and, consequently, it induces honest reporting due to Proposition 1. Given that each agent reported nine answers, if the answers reported by all members of a group were the same, then all group members would receive the maximum bonus of \$0.10.

It is fair to mention one important issue regarding our experimental design, which is the possibility of the payment amount to be confounded

with the payment structure. In particular, under the current experimental design, a member of either the BG group or the BIG group receives no less money than a member of the CG group. We could have tackled this problem in two different ways. First, we could have used a different payment structure in the control condition. We argue that such a solution is not satisfactory because a fixed payment (i.e., payment per completed task) is a canonical payment structure in crowdsourcing, for example, it is the standard payment structure in crowdsourcing platforms such as AMT and CrowdFlower.[2]

The second alternative would be to provide an extra bonus to members of the CG group based on the expected bonus received by members of BG/BIG groups. This alternative is not satisfactory as well for many reasons. First, it is hard to assess a priori the expected bonus received by members of BG/BIG groups. Second, the expected bonus received by members of the BG group might not be equal to the expected bonus received by members of the BIG group. Finally, since the extra bonus paid to members of the CG group would be based on expected values, there would still be individual differences in the amount paid by different payment structures, thus not entirely solving the initial confounding problem. Although not free of problems, we see our experimental design as the best approach in that all group members receive a similar baseline payment, and members of the treatment groups receive an extra bonus for the sake of honest reporting. Furthermore, as reported by Buhrmester et al. (2011), the level of compensation on AMT does not have much influence on the quality of workers' outputs. As a consequence, we think our experimental results when comparing Group CG to other groups are still of interest and valid. We emphasize, however, that the above confounding problem is not an issue when comparing the Groups BG and BIG, because the only difference between these two treatments is the statement about honest reporting received by members of the Group BIG.

Objectives and Gold-Standard Answers

As discussed in the Social Projection section, the literature on the false-consensus effect has shown that social projection is often a valid assumption. Moreover, the assumption of risk-neutral behavior is theoretically valid when the stakes are small (Arrow, 1971), as in our experiment. Hence, instead of investigating the validity of the assumptions behind our model, the main purpose of our experiment was to investigate the practical benefits of using a payment structure based on pairwise comparisons. We did so by comparing reported answers from different treatment groups to *gold-standard answers*. We further explain in the upcoming Applicability section why crowdsourcing of microtasks is an ideal setting to apply our payment structures.

We were able to derive gold-standard answers for each multiple-choice question because we knew the source and original content of each text a priori, that is, before we conducted the content analysis experiment. To avoid confirmation bias,[3] we asked five professors and tutors from the English and Literature Department at the University of Waterloo, Waterloo, Ontario, Canada, to provide their answers for each multiple-choice question. We set the gold-standard answer for each question as the median of the answers reported by the professors and tutors. Coincidentally, each median value was also the mode of the underlying answers. The Appendix contains distributions of the answers reported by the agents and professors/tutors as well as the respective gold-standard answers.

Hypotheses

Our first research question was whether providing rewards through pairwise comparisons makes the reported answers more accurate, that is, closer to the gold-standard answers. Our hypothesis was:

> *Hypothesis 1:* The average error of Group BIG is less than the average error of Group BG, which in turn is less than the average error of Group CG.

In other words, the resulting answers would be on average more accurate when agents received bonuses based on pairwise comparisons. Moreover, the extra explanation regarding the theory behind the payment structure would provide more credibility to it, thus making the reported answers even more accurate. That is,

[2] See http://www.crowdflower.com/.

[3] The tendency to interpret information in a way that confirms one's preconceptions (Plous, 1993).

framing has a positive influence on the accuracy of the reported answers. Regarding the resulting bonuses, because honest reporting maximizes agents' expected rewards in our model (Proposition 1), our second hypothesis was:

> *Hypothesis 2:* The average bonus received by members of Group BIG is greater than the average bonus received by members of Group BG, which in turn is greater than the average bonus received by members of Group CG.

To test whether or not Hypothesis 2 was true, we used the bonus the members of Group CG would have received had they received any bonus. It is noteworthy that we measured Hypothesis 1 by comparing how close the reported answers were to the gold-standard answers, whereas we measured Hypothesis 2 by making pairwise comparisons between reported answers: the higher the number of agreements, the greater the resulting bonus.

Another metric we used to compare groups' performance was the task completion time. The amount of time agents spent on the content-analysis task can be seen as a proxy for the effort they exerted to complete the task. Regarding this metric, we expected agents who received bonuses to be more cautious when completing their tasks. Moreover, the extra explanation regarding the theory behind the payment structure would provide more credibility to it, thus making the members of Group BIG work longer on the task. Hence, our third hypothesis was:

> *Hypothesis 3:* The average task completion time of Group BIG is greater than the average task completion time of Group BG, which in turn is greater than the average task completion time of Group CG.

Experimental Results

In the following subsections, we describe our experimental results and analyze our hypotheses.

Error on individual criteria. In our first analysis, we defined the error of each agent's reported answer as the absolute difference between his answer and the corresponding gold-standard answer. Thus, the outcome measure for each multiple-choice question was an integer with a value between zero and two, and the closer this value was to zero, the better the resulting accuracy. Table 1 shows the average error for each group.

Table 1
The Average of the Absolute Difference Between the Reported Answers and the Corresponding Gold-Standard Answers for Each Group

				p values		
Variable	BG	BIG	CG	BIG-BG	BIG-CG	BG-CG
Text 1						
Grammar	0.50 (0.51)	**0.32** (0.47)	0.44 (0.50)	.10*	.17	.73
Clarity	0.82 (0.66)	**0.62** (0.60)	0.86 (0.73)	.10*	.10*	.41
Relevance	0.22 (0.51)	**0.20** (0.45)	0.30 (0.58)	.48	.34	.34
Text 2						
Grammar	0.44 (0.50)	**0.36** (0.48)	0.38 (0.49)	.63	.63	.73
Clarity	0.50 (0.65)	**0.38** (0.60)	0.54 (0.61)	.23	.20	.33
Relevance	**0.44** (0.50)	0.64 (0.48)	0.66 (0.48)	.98	.63	.04**
Text 3						
Grammar	**0.76** (0.85)	0.78 (0.86)	1.02 (0.84)	.54	.12	.12
Clarity	0.14 (0.40)	**0.00** (0.00)	0.16 (0.37)	.009**	.005**	.12
Relevance	0.12 (0.44)	**0.10** (0.36)	0.20 (0.49)	.49	.18	.18

Note. For each criterion, the lowest average is highlighted in boldface type. Standard deviations are in parentheses. One-tailed *p* values resulting from rank-sum tests are shown in the last three columns. Given the notation A-B, the null hypothesis is that the outcome measures resulting from Groups A and B are equivalent, and the alternative hypothesis is that the outcome measure resulting from Group A is less than the outcome measure resulting from Group B. The *p* values are adjusted for multiple comparisons using the method by Benjamini and Yekutieli (2001). BG = bonus group; BIG = bonus and information group; CG = control group.
* $p \leq .1$. ** $p \leq .05$.

Looking at the average errors in isolation, we see that Group BIG is the most accurate group in seven out of nine criteria, and Group BG is the most accurate group in the remaining two criteria (Relevance in Text 2 and Grammar in Text 3). Moreover, Group CG, the control condition that involved no incentives beyond the baseline compensation offered for completing the task, never outperforms Group BIG, and it outperforms Group BG in only two occasions (Grammar in Text 1 and Text 2).

When comparing the Groups BIG and BG, the former is significantly more accurate than the latter in three occasions, whereas BG is significantly more accurate than BIG in one occasion (Relevance in Text 2). When comparing the Groups BIG and CG, the former is significantly more accurate than the latter in two occasions, whereas CG never outperforms BIG. Finally, when comparing the Groups BG and CG, the former is significantly more accurate than the latter in one occasion, whereas CG never significantly outperforms BG.

Given these results, we have some weak evidence that Hypothesis 1 is true for individual criteria, that is, it seems that the resulting answers are on average more accurate when providing rewards based on pairwise comparisons, and the extra explanation regarding the theory behind the payment structure seems to provide more credibility to it, thus reducing the error of the reported answers.

Aggregate error. We also computed the aggregate error for each text as well as for the whole task. In the former case, the outcome measure was the sum of the absolute differences between each reported answer for a given text and the corresponding gold-standard answer. For example, given (0, 1, 2) as the reported answers for Text 1 and (1, 2, 2) as the corresponding gold-standard answers, the outcome measure for Text 1 would be $|0 - 1| + |1 - 2| + |2 - 2| = 2$. For the whole task, we summed the absolute differences across all criteria and texts. Table 2 shows the average aggregate error for each group.

Looking at the average errors, one can see that members of Group CG always report, on average, less accurate answers than members of Group BG and Group BIG. This result is statistically significant for Text 3 and for the overall task. Thus, the experimental results suggest that providing rewards through pairwise comparisons between reported answers produces a weak but slightly significant improvement in quality over the control condition.

Finally, providing an extra explanation about the theory behind the payment structure seems to improve the average quality of the answers because, on average, the answers from Group BIG are at least as accurate as the answers from Group BG. This result, however, is not statistically significant. Therefore, we have some weak evidence that Hypothesis 1 is true on the aggregate level.

Bonus. We show the average bonus per group in the first row of Table 3. From it, we

Table 2
The Average of the Sum of the Absolute Differences Between the Reported Answers and the Corresponding Gold-Standard Answers for Each Group

				p values		
Variable	BG	BIG	CG	BIG-BG	BIG-CG	BG-CG
Text 1	1.54 (1.13)	**1.14** (1.03)	1.60 (1.41)	.13	.13	.59
Text 2	**1.38** (1.07)	**1.38** (0.97)	1.58 (0.99)	.55	.24	.24
Text 3	1.02 (1.19)	**0.88** (0.92)	1.38 (1.19)	.39	.06*	.08*
Overall	3.94 (2.24)	**3.40** (1.69)	4.56 (2.13)	.11	.006**	.1*

Note. For each text and for the whole task, the lowest average is highlighted in boldface type. Standard deviations are in parentheses. One-tailed *p* values resulting from rank-sum tests are given in the last three columns. Given the notation A-B, the null hypothesis is that the outcome measures resulting from Groups A and B are equivalent, and the alternative hypothesis is that the outcome measure resulting from Group A is less than the outcome measure resulting from Group B. The *p* values are adjusted for multiple comparisons using the method by Benjamini and Yekutieli (2001). BG = bonus group; BIG = bonus and information group; CG = control group.
* $p \leq .1$. ** $p \leq .05$.

Table 3
Average Bonus (in Cents) and Completion Time (in Seconds) Per Group

Variable	BG	BIG	CG	p values		
				BIG-BG	BIG-CG	BG-CG
Bonus	.0533 (.009)	**.0584** (.007)	.0495 (.008)	$<.0005^{**}$	$<.0005^{**}$	$.0026^{**}$
Time	178.66 (87.45)	**215.90** (127.75)	196.36 (149.08)	$.04^{**}$	$.04^{**}$	.42

Note. The highest average values are highlighted in boldface type. Standard deviations are in parentheses. One-tailed p values resulting from rank-sum tests are given in the last three columns. Given the notation A-B, the null hypothesis is that the outcome measures resulting from Groups A and B are equivalent, and the alternative hypothesis is that the outcome measure resulting from Group A is greater than the outcome measure resulting from Group B. The p values are adjusted for multiple comparisons using the method by Benjamini and Yekutieli (2001). BG = bonus group; BIG = bonus and information group; CG = control group.
** $p \leq .05$.

conclude that Hypothesis 2 is true, that is, the average bonus received by members of Group BIG is greater than the average bonus received by members of Group BG, which in turn is greater than the average bonus hypothetically received by members of Group CG. All these results are statistically significant with (adjusted) $p \leq .05$. In other words, providing rewards based on pairwise comparisons between reported answers and informing agents about the theory behind the payment structure do indeed increase the number of reported answers that are similar.

It is interesting that there is a strong negative correlation between bonuses and the aggregate error for the whole task shown in the fourth row of Table 2, even though the former is computed by making pairwise comparisons between reported answers, whereas the latter is computed by comparing reported answers with gold-standard answers. The Pearson correlation coefficients for BG, BIG, and CG are, respectively, $-.73$, $-.79$, and $-.72$. This result implies that there exists a strong positive correlation between honest reporting and accuracy in our content-analysis experiment.

Completion time. We show the average completion time per group in the second row of Table 3. We start by noting that Hypothesis 3 is not true. Surprisingly, the average time spent on the task by members of Group BG is statistically equivalent to the average time spent by members of Group CG since the null hypothesis cannot be rejected. The average completion time by members of Group BIG is the highest one among the three groups, and this result is statistically significant.

A possible explanation for the above result is the potential presence of outliers, as illustrated in Figure 1, distorting the average values. We note, however, that even after removing outliers, that is, the highest completion time per group, the above result is still qualitatively the same. Another plausible explanation of the above result is that agents work on the content-analysis task more seriously by taking more time to complete it when they receive a brief explanation regarding some theoretical properties of the proposed payment function, whereas the same agents could be quickly guessing how their peers would report their answers when the extra explanation about the theoretical properties is not provided. This perspective highlights the importance of such a priming effect when inducing honest reporting of private information.

It is noteworthy that even though the average values might suggest that spending more time on the task results in higher bonuses and lower aggregate errors, we do not find any significant correlation between these variables at an individual level.

Discussion

We discuss in this section some practical issues regarding our model and results. In particular, we highlight the importance of inducing honest reporting before aggregating agents' answers, how to measure whether (strong) social projection holds true in practice, settings where the assumptions behind our model are likely to hold true, and how to increase the strength of social projection.

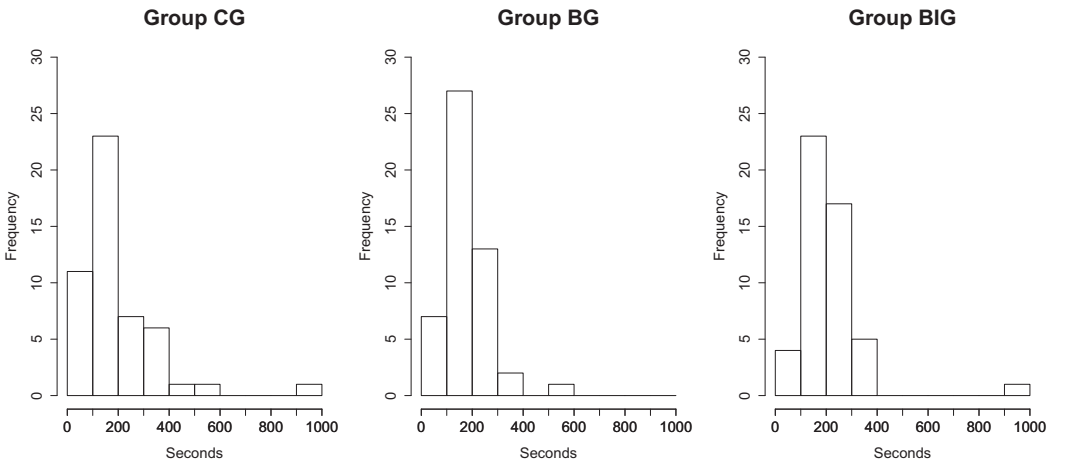

Figure 1. Histograms of task completion time per group. CG = control group; BG = bonus group; BIG = bonus and information group.

Aggregating Reported Answers

After agents report their answers and receive their rewards, there is still the question of how the requester will aggregate the reported answers. We explain in this section why the requester can rely on the empirical distribution of the reported answers when he induces honest reporting.

Let h_k be the number of times that the requester received the answer A_k, for $k \in \{1, \ldots, n\}$. The empirical distribution of the reported answers is then:

$$\left(\frac{h_1}{\sum_{x=1}^{n} h_x}, \frac{h_2}{\sum_{x=1}^{n} h_x}, \ldots, \frac{h_n}{\sum_{x=1}^{n} h_x} \right).$$

Because of the law of large numbers, the empirical distribution of the reported answers converges to $\boldsymbol{\omega}$, the probability vector that represents the population knowledge, as the number of honestly reported answers increases. This result serves as a word of caution on indiscriminately using the empirical distribution of the reported answers as the distribution of agents' observed signals. Such a perspective is unlikely to hold true when agents do not report their private information honestly.

Measuring the Strength of the Social Projection Effect

The payment structures we discussed in this article rely on the assumption that social projection holds true. We discuss in this subsection how to validate such an assumption in practice. A straightforward way to check for the presence of social projection is by eliciting agents' posterior predictive distributions. After an agent reports both his answer and his posterior predictive distribution, the requester can then verify whether the agent's reported answer is also the most likely answer according to his posterior predictive distribution. The requester can also verify whether strong social projection holds true by measuring if the answer reported by the agent is more likely according to his posterior predictive distribution than all the other answers combined.

We note that the requester must induce honest reporting of posterior predictive distributions because they are also agents' private information. To do so, the requester can provide an extra reward by applying a proper scoring rule to an agent's reported posterior predictive distribution. For example, let $\mathbf{q}$ be agent i's reported posterior predictive distribution. Agent i's extra reward is then $R(\mathbf{q}, r_j)$, where r_j is the answer reported by an agent $j \neq i$ randomly selected from the population of agents. Because R is a proper scoring rule, agent i strictly maximizes his expected extra reward by reporting his true posterior predictive distribution, that is, by reporting $\mathbf{q} = \mathbb{E}[\boldsymbol{\omega} \mid t_i]$.

It is interesting to note that when eliciting both answers and predictions, our model becomes similar to the Bayesian Truth Serum

(BTS) model (Prelec, 2004). Unlike BTS, however, we do not use the predictions to determine agents' rewards for their answers. Predictions are only used to verify whether the assumption of social projection is valid. In spirit, eliciting predictions to check whether social projection holds true is equivalent to eliciting an agent's utility function to check whether risk-neutral behavior is a valid assumption. If there is evidence that social projection holds true, for example, based on previous elicitation tasks for similar multiple-choice questions and population of agents, then the requester no longer needs to elicit predictions.

Applicability

It is fair to mention that there are cases where the assumptions behind our model are not valid. For example, due to the autonomy assumption, our model is not applicable to scenarios involving brainstorming or group discussions. In such settings, payments based on agreements do not work because agents can easily collude and agree on a fixed answer to maximize their rewards.

The second assumption, namely risk-neutral behavior, is more likely to hold true when the underlying payments are small. Otherwise, it is likely that the payment structure will cause distortions in the way agents report their private information. For example, when reporting beliefs, agents might underestimate (overestimate) high (small) probability values under traditional proper scoring rules (Carvalho, 2015; Offerman, Sonnemans, Van De Kuilen, & Wakker, 2009). An interesting open question is how to adapt the payment structures discussed in this article by considering agents' attitudes toward risk and uncertainty.

Arguably, the most crucial assumption in our model is that social projection holds true. We draw from the social comparison literature to foresee when such an assumption is most likely to be valid. First, human beings are more likely to project themselves onto others when they answer subjective questions, as opposed to objective questions (Marks & Miller, 1987). One reason for this is that subjective questions involve opinions, whereas objective questions involve skills. Human beings tend to think that their opinions are common, but that they are unique in terms of abilities/skills (Marks, 1984). When a task is highly dependent on agents' skills, the opposite of social projection/false consensus might happen, that is, agents might think that they are unique, a phenomenon called the false uniqueness effect (Campbell, 1986). An interesting research question is about how to develop payment structures to induce honest reporting under the false uniqueness effect.

When considering the above discussion, a scenario that seems to satisfy the assumptions in our model is crowdsourcing involving microtasks, that is, tasks that are relatively easy to complete, that offer low payments, and which often involve human judgment. For example, the tasks posted on AMT are microtasks but referred to as human intelligence tasks because they are complex for current computers to solve but nonetheless are easy for human beings. Many of these tasks involve opinions, as opposed to highly specialized skills, for example, sentiment-analysis tasks, content-analysis tasks, tasks involving ratings, classification/categorization, and so on.

Social Projection and Truthfulness: Guidelines for Requesters

Under our model, the larger the strength of the social projection effect, the more assurance the requester has that a risk-neutral agent is reporting honestly. A question that then arises is how to design multiple-choice questions in such a way to increase the strength of the projection effect.

Mullen et al. (1985) gave some answers to the above question when investigating the strength of the false-consensus effect. Mullen et al.s' analysis showed that the number of questions agents had to answer and the sequence of measurement of answers and predictions significantly predicted the magnitude of the false-consensus effect. In particular, false consensus is stronger when agents answer few multiple-choice questions and estimate consensus before reporting their answers.

When combining our results with the results by Mullen et al. (1985) and Robbins and Krueger (2005), whose analysis showed that false consensus is stronger when agents make judgments about ingroups than when they make judgments about outgroups, we obtain the following guidelines for the requester to reliably elicit agents' private information. Ideally, all the points below should be followed to avoid erroneous conclusions:

1. *Ask few multiple-choice questions*: As suggested by Mullen et al. (1985), agents project more when answering less questions. One reason for this is that when answering many questions, agents might eventually realize their overprojection and start to examine each new question more carefully.
2. *Elicit demographic/personal information*: The requester can use this information to group agents based on similarity, that is, to create ingroups.
3. *State that the payment structure is based on pairwise comparisons between answers reported by similar agents*: According to Robbins and Krueger (2005), social projection is stronger when agents make judgments about similar others.
4. *Briefly describe some theoretical results regarding the payment structure*: As our results show, priming agents by briefly mentioning the theoretical properties of the underlying payment structure results in more accurate and similar reported answers.
5. *Make payments based on pairwise comparisons between reports by similar agents*: As Proposition 1 and 2 show, pairwise comparisons induce honest reporting by risk-neutral agents in our setting.

Conclusion

We suggested payment structures to induce honest reporting of private information by risk-neutral agents in multiple-choice questions. The proposed methods are based on pairwise comparisons between agents' reported answers and, thus, they do not rely on the existence of a ground truth. Instead, the proposed payment structures rely on the psychological phenomenon referred to as social projection, a variant of the well-known false-consensus effect.

The first payment structure we discussed in this article simply compares agents' reported answers and rewards agreements. We showed how to derive such a method in terms of bounded and symmetric proper scoring rules. The second payment structure takes the distance between agents' reported answers into account by penalizing disagreements proportionally to the absolute difference between the indices of the reported answers. We showed how to derive such a payment method in terms of the ranked probability scoring rule.

We tested the effectiveness of payment structures based on pairwise comparisons on a content-analysis experiment using Amazon Mechanical Turk. Our empirical results showed that providing rewards through pairwise comparisons between reported answers results in slightly more accurate answers than when agents have no direct incentives for expressing their honest answers. Moreover, agents tended to agree more with each other when they received such rewards. Finally, we showed that priming agents by briefly mentioning the theoretical properties of the underlying payment structure results in even more accurate and similar reported answers. There are two interesting open questions that deserve further investigation. First, we used bonus in our experiments when testing the effectiveness of our payment structure. It would be interesting to investigate the extent to which the resulting behavior is driven by the extra bonus or by the payment structure per se. Second, it is important to further investigate the extent to which priming drives honest reporting, that is, under which circumstances can one induce honest reporting by using an arbitrarily complex or even a random payment structure and priming?

In our experiments, we took for granted the existence of the social-projection phenomenon. However, we suggested in the Applicability section some conditions for using our payment structures, and we argued that crowdsourcing of microtasks seems to satisfy all the assumptions behind our model, including social projection. Clearly, an interesting research direction is to better understand for which types of tasks and subgroup of workers our modeling assumptions are more appropriate.

Given the encouraging results obtained in our content-analysis experiment, another interesting open question is whether payment structures based on pairwise comparisons would perform as well in other domains. One particularly exciting domain for investigation is the peer-review process as used in massive open online courses because such peer-review tasks can often be modeled as multiple-choice questions. Another question worth contemplating is whether incentives other than from the received rewards play a role in inducing honest reporting. For example, Boons, Stam, and Barkema (2015) suggested that pride and respect also drive workers' engagement in crowdsourcing.

In our setting, one can also conjecture that altruism may play an important role. Specifically, in our content-analysis experiment, agents' reported answers not only affect their own rewards but also the rewards of their peers. In other words, if agents do not put enough effort into reporting high-quality answers, not only might they receive low rewards but also other answers evaluated based on those erroneous answers might also receive low rewards. Thus, an interesting future work is to investigate whether agents have an altruistic motive to put more effort into the underlying task to maximize the potential rewards of their peers.

Regarding social projection, the literature on social comparison has shown that social projection serves as an egocentric heuristic for inductive reasoning (Robbins & Krueger, 2005). However, it is still unclear exactly which social and psychological factors play the largest role in the strength and prevalence of social projection. Understanding this facet is of great value when eliciting agents' private information since this might help a requester to induce social projection and, consequently, honest reporting under payment structures based on pairwise comparisons. Another relevant question regards the most suitable prior/posterior distributions to model the social-projection phenomenon. Dirichlet distributions provide an elegant way to instantiate our general model. However, the effectiveness of such a modeling choice is still an open question. In our work, we modeled social projection by constraining the posterior distributions. This is equivalent to adding constraints to both the prior distributions and likelihood functions. Understanding these prior/likelihood constraints might provide some guidance on how to develop stronger models.

References

Antin, J., & Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2925–2934). Austin, TX: ACM.

Arrow, K. J. (1971). *Essays in the theory of risk-bearing*. Amsterdam, the Netherlands: North-Holland.

Bacon, D. F., Chen, Y., Kash, I., Parkes, D. C., Rao, M., & Sridharan, M. (2012). Predicting your own effort. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* (pp. 695–702). Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems.

Bauman, K. P., & Geher, G. (2002). We think you agree: The detrimental impact of the false consensus effect on behavior. *Current Psychology, 21,* 293–318.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29,* 1165–1188.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York, NY: Wiley.

Boons, M., Stam, D., & Barkema, H. G. (2015). Feelings of pride and respect as drivers of ongoing member activity on crowdsourcing platforms. *Journal of Management Studies, 52,* 717–741.

Boutilier, C. (2012). Eliciting forecasts from self-interested experts: Scoring rules for decision makers. *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, 2,* 737–744.

Brenner, L., & Bilgin, B. (2011). Preference, projection, and packing: Support theory models of judgments of others preferences. *Organizational Behavior and Human Decision Processes, 115,* 121–132.

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5.

Busemeyer, J. R., & Pothos, E. M. (2012). Social projection and a quantum approach for behavior in prisoner's dilemma. *Psychological Inquiry, 23,* 28–34.

Campbell, J. D. (1986). Similarity and uniqueness: The effects of attribute type, relevance, and individual differences in self-esteem and depression. *Journal of Personality and Social Psychology, 50,* 281–294.

Carvalho, A., & Larson, K. (2011, May 2–6). A truth serum for sharing rewards. In K. Tumer, P. Yolum, L. Sonenberg, & P. Stone (Eds.), *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS, 2011)* (pp. 635–642). Taipei, Taiwan: International Foundation for Autonomous Agents and Multiagent Systems.

Carvalho, A., & Larson, K. (2012). Sharing rewards among strangers based on peer evaluations. *Decision Analysis, 9,* 253–273.

Carvalho, A., Dimitrov, S., & Larson, K. (2014). The output-agreement method induces honest behavior in the presence of social projection. *ACM SIGecom Exchanges, 13,* 77–81.

Carvalho, A., Dimitrov, S., & Larson, K. (2015). A study on the influence of the number of MTurkers on the quality of the aggregate output. In N. Bulling (Ed.), *Multi-agent systems, lecture notes in*

computer science (Vol. 8953, pp. 285–300). New York, NY: Springer.

Carvalho, A. (2015). Tailored proper scoring rules elicit decision weights. *Judgment and Decision Making, 10,* 86–96.

Chen, Y., & Kash, I. A. (2011, May 2–6). Information elicitation for decision making. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (pp. 175–182). Taipei, Taiwan: International Foundation for Autonomous Agents and Multiagent Systems.

Chien, Y. H., & George, E. I. (1999). A Bayesian model for collaborative filtering. *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*.

Chiu, C. M., Liang, T.-P., & Turban, E. (2014a). Introduction to the special issue on "Crowdsourcing and Social Networks Analysis." *Decision Support Systems, 65,* 1–2.

Chiu, C. M., Liang, T.-P., & Turban, E. (2014b). What can crowdsourcing do for decision support? *Decision Support Systems, 65,* 40–49.

da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems, 66,* 170–179.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology, 25,* 1–17.

Dimitrov, S., & Sami, R. (2010, June). Composition of markets with conflicting incentives. In *Proceedings of the 11th ACM EC10 Conference* (pp. 53–62). Cambridge, MA: ACM.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology, 8,* 985–987.

Geiger, D., & Schader, M. (2014). Personalized task recommendation in crowdsourcing information systems: Current state of the art. *Decision Support Systems, 65,* 3–16.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102,* 359–378.

Hanson, R. (2003). Combinatorial information market design. *Information Systems Frontiers, 5,* 107–119.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics, 14,* 399–425.

Huang, S. W., & Fu, W. T. (2013, February). Enhancing reliability using peer consistency evaluation in human computation. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 639–648). New York, NY: ACM.

Ipeirotis, P. G. (2010). Analyzing the Amazon Mechanical Turk marketplace. *XRDS Crossroads: The ACM Magazine for Students, 17,* 16–21.

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science, 55,* 582–590.

Katz, D., & Allport, F. H. (1931). *Students' attitudes: A Report of the Syracuse University Reaction study*. Syracuse, NY: Craftsman Press.

Marge, M., Banerjee, S., & Rudnicky, A. I. (2010, March). Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing* (pp. 5270–5273). Dallas, TX: ACL.

Marks, G., Graham, J. W., & Hansen, W. B. (1992). Social projection and social conformity in adolescent alcohol use: A longitudinal analysis. *Personality and Social Psychology Bulletin, 18,* 96–101.

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin, 102,* 72.

Marks, G. (1984). Thinking one's abilities are unique and one's opinions are common. *Personality and Social Psychology Bulletin, 10,* 203–208.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44,* 1–23.

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

Mohammad, S. M. (2012). From once upon a time to happily ever after: Tracking Emotions in mail and books. *Decision Support Systems, 53,* 730–741.

Mullen, B., Atkins, J. L., Champion, D. S., Edwards, C., Hardy, D., Story, J. E., & Vanderklok, M. (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology, 21,* 262–283.

Murphy, A. H. (1970). A note on the ranked probability score. *Journal of Applied Meteorology, 10,* 155–156.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences with Dirichlet processes. *Journal of Mathematical Psychology, 50,* 101–122.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15,* 263–280.

Neruda, P. (2007). *100 love sonnets* (bilingual ed.). Holstein, Ontario, Canada: Exile.

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

Othman, A., & Sandholm, T. (2010). Decision rules and decision markets. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems* (pp. 625–632). Budapest, Hungary: AAMAS.

Plous, S. (1993). *The psychology of judgment and decision making*. New York, NY: McGraw-Hill Education.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of "rational" decision theory. *Proceedings of the Royal Society B: Biological Sciences, 276,* 2171–2178.

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

Radanovic, G., & Faltings, B. (2013). A robust Bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence* (pp. 833–839). Bellevue, WA: AAAI Press.

Regan, K., Poupart, P., & Cohen, R. (2006). Bayesian reputation modeling in E-marketplaces sensitive to subjectivity, deception and change. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence* (pp. 1206–1212). Boston, MA: AAAI Press.

Ren, J., Nickerson, J. V., Mason, W., Sakamoto, Y., & Graber, B. (2014). Increasing the crowd's capacity to create: How alternative generation affects the diversity, relevance and effectiveness of generated ads. *Decision Support Systems, 65,* 28–39.

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review, 9,* 32–47.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist, 88,* 313–338.

Ross, L., Green, D., & House, P. (1977). The "false in social consensus perception effect": An egocentric bias and attribution processes. *Journal of Experimental Social Psychology, 13,* 279–301.

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (pp. 275–284). Hangzhou, China: ACM.

Sherman, S. J., Presson, C. C., Chassin, L., Corty, E., & Olshavsky, R. (1983). The false consensus effect in estimates of smoking prevalence underlying mechanisms. *Personality and Social Psychology Bulletin, 9,* 197–207.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 254–263). Honolulu, HI: ACL.

Staël von Holstein, C. A. S. (1970). A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology, 9,* 360–364.

Taylor, J., Taylor, A., & Greenaway, K. (2010). *Little Ann and other poems*. Charleston, SC: Nabu Press.

Torralba, A., Willsky, A. S., Sudderth, E. B., & Freeman, W. T. (2005). Describing visual scenes using transformed Dirichletprocesses. In *Advances in Neural Information Processing Systems* (pp. 1297–1304). Retrieved from https://papers.nips.cc/

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547.

Vlachos, A., Korhonen, A., & Ghahramani, Z. (2009). Unsupervised and constrained Dirichlet-process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* (pp. 74–82). Athens, Greece: ACL.

von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM, 51,* 58–67.

Waggoner, B., & Chen, Y. (2013). Information elicitation sans verification. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content*.

Walker, M. A., Anand, P., Abbott, R. J. E. F. T., Martell, C., & King, J. (2012). That is your evidence? Classifying stance in online political debate. *Decision Support Systems, 53,* 719–729.

Weaver, R., & Prelec, D. (2013). Creating truth-telling incentives with the Bayesian truth serum. *Journal of Marketing Research, 50,* 289–302.

Weiss, R. R. J. (2009). *Optimally aggregating elicited expertise: A proposed application of the Bayesian truth serum for policy analysis* (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Cambridge, MA.

Winkler, R. L., & Murphy, A. H. (1968). "Good" probability assessors. *Journal of Applied Meteorology, 7,* 751–758.

Witkowski, J., & Parkes, D. C. (2012a). A robust Bayesian truth serum for small populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. Toronto, Ontario, Canada: AAAI press.

Witkowski, J., & Parkes, D. C. (2012b). Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (pp. 964–981). Valencia, Spain: ACM.

Yinon, Y., Mayraz, A., & Fox, S. (1994). Age and the false-consensus effect. *The Journal of Social Psychology, 134,* 717–725.

(*Appendix follows*)

Appendix

Description of the Texts and Agents/Experts' Reported Answers

In this Appendix, we describe the texts we used in our content-analysis experiment, the distributions of agents' reported answers, and the gold-standard answers reported by five professors and tutors from the English and Literature Department at the University of Waterloo, Waterloo, Ontario, Canada, henceforth referred to as the *experts*.

Text 1

An excerpt from the "Sonnet XVII" by Neruda (2007). Intentionally misspelled words are highlighted in bold.

I do not love you as if you **was** salt-rose, or topaz,
or the **arrown** of carnations that spread fire:
I love you as certain dark things are loved,
secretly, between the **shadown** and the soul

Table A1 shows the distributions of agents' reported answers. Table A2 shows the experts' reported answers. The gold-standard answer for each criterion is the median/mode of experts' reported answers.

Table A1
Distributions of Agents' Reported Answers Per Group and Criterion for Text 1

	CG			BG			BIG		
Criterion	0	1	2	0	1	2	0	1	2
Grammar	.28	.56	.16	.38	.50	.12	.22	.68	.10
Clarity	.20	.46	.34	.14	.54	.32	.06	.50	.44
Relevance	.06	.18	.76	.04	.14	.82	.02	.16	.82

Note. CG = control group; BG = bonus group; BIG = bonus and information group.

Table A2
Answers Reported by the Experts for Text 1

Criterion	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Median/ Mode
Grammar	1	0	1	0	1	1
Clarity	2	2	2	1	2	2
Relevance	2	2	2	2	2	2

Text 2

An excerpt from "The Cow" by Taylor et al. (2010). Intentionally misspelled words are highlighted in bold.

THANK you, **prety** cow, that made
Plesant milk to soak my bread,
Every day and every night,
Warm, and fresh, and sweet, and white.

Table A3 shows the distributions of agents' reported answers. Table A4 shows the experts' reported answers. The gold-standard answer for each criterion is the median/mode of experts' reported answers.

Table A3
Distributions of Agents' Reported Answers Per Group and Criterion for Text 2

	CG			BG			BIG		
Criterion	0	1	2	0	1	2	0	1	2
Grammar	.18	.62	.20	.30	.56	.14	.20	.64	.16
Clarity	.06	.42	.52	.08	.34	.58	.06	.26	.68
Relevance	.58	.34	.08	.40	.56	.04	.54	.36	.10

Note. CG = control group; BG = bonus group; BIG = bonus and information group.

(*Appendix continues*)

Table A4
Answers Reported by the Experts for Text 2

Criterion	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Median/ Mode
Grammar	1	1	1	1	1	1
Clarity	2	2	2	1	2	2
Relevance	1	0	0	1	1	1

Text 3

Random words in a semistructured way. Each line starts with a noun followed by a verb in a wrong verb form. All the words in the same line start with a similar letter in order to mimic a poetic writing style.

Baby bet binary boundaries bubbles
Carlos cease CIA conditionally curve
Daniel deny disease domino dumb
Faust fest fierce forced furbished

Table A5 shows the distributions of agents' reported answers. Table A6 shows the experts' reported answers. The gold-standard answer for each criterion is the median/mode of experts' reported answers.

Table A5
Distributions of Agents' Reported Answers Per Group and Criterion for Text 3

	CG			BG			BIG		
Criterion	0	1	2	0	1	2	0	1	2
Grammar	.34	.30	.36	.50	.24	.26	.50	.22	.28
Clarity	.84	.16	.00	.88	.10	.02	1.00	.00	.00
Relevance	.84	.12	.04	.92	.04	.04	.92	.06	.02

Note. CG = control group; BG = bonus group; BIG = bonus and information group.

Table A6
Answers Reported by the Experts for Text 3

Criterion	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Median/ Mode
Grammar	0	1	0	0	0	0
Clarity	0	0	0	0	0	0
Relevance	0	1	0	0	0	0

Received March 10, 2015
Revision received December 2, 2015
Accepted December 17, 2015 ■