MANAGEMENT SCIENCE
Vol. 51, No. 9, September 2005, pp. 1359–1373
ISSN 0025-1909 | EISSN 1526-5501 | 05 | 5109 | 1359

informs

DOI 10.1287/mnsc.1050.0379

Eliciting Informative Feedback: The Peer-Prediction Method

Nolan Miller
Kennedy School of Government, Harvard University, Cambridge, Massachusetts 02138, nolan_miller@harvard.edu

Paul Resnick
School of Information, University of Michigan, Ann Arbor, Michigan 48109-1092, presnick@umich.edu

Richard Zeckhauser
Kennedy School of Government, Harvard University, Cambridge, Massachusetts 02138, richard_zeckhauser@harvard.edu

Many recommendation and decision processes depend on eliciting evaluations of opportunities, products, and vendors. A scoring system is devised that induces honest reporting of feedback. Each rater merely reports a signal, and the system applies proper scoring rules to the implied posterior beliefs about another rater's report. Honest reporting proves to be a Nash equilibrium. The scoring schemes can be scaled to induce appropriate effort by raters and can be extended to handle sequential interaction and continuous signals. We also address a number of practical implementation issues that arise in settings such as academic reviewing and online recommender and reputation systems.

Key words: proper scoring rules; electronic markets; honest feedback
History: Accepted by James E. Smith, decision analysis; received February 21, 2003. This paper was with the authors 1 year and 1 week for 2 revisions.

1. Introduction

Decision makers frequently draw on the experiences of multiple other individuals when making decisions. The process of eliciting others' information is sometimes informal, as when an executive consults underlings about a new business opportunity. In other contexts, the process is institutionalized, as when journal editors secure independent reviews of papers, or an admissions committee has multiple faculty readers for each file. The Internet has greatly enhanced the role of institutionalized feedback methods, since it can gather and disseminate information from vast numbers of individuals at minimal cost. To name just a few examples, eBay invites buyers and sellers to rate each other; NetFlix, Amazon, and ePinions invite ratings of movies, books, etc., on a 1–5 scale; and Zagat Survey solicits restaurant ratings on a 1–30 scale on food, decor, and service.

Any system that solicits individual opinions must overcome two challenges. The first is underprovision. Forming and reporting an opinion requires time and effort, yet the information only benefits others. The second challenge is honesty. Raters' desire to be nice or fear of retaliation may cause them to withhold negative feedback.[1] On the other hand, conflicts of interest or a desire to improve others' perception of them may lead raters to report distorted versions of their true opinions.

An explicit reward system for honest rating and effort may help to overcome these challenges. When objective information will be publicly revealed at a future time, individuals' reports can be compared to that objective information. For example, evaluations of stocks can be compared to subsequent price movements, and weather forecasts can be compared to what actually occurs.

This analysis develops methods to elicit feedback effectively when independent, objective outcomes are not available. Examples include situations where no objective outcome exists (e.g., evaluations of a product's "quality"), and where the relevant information is objective but not public (e.g., a product's breakdown frequency, which is only available to others if the product's current owners reveal it).

[1] Dellarocas (2001) shows that leniency in feedback can offer some advantages in deterring seller opportunism. The problem we are concerned with here is not systematic leniency, but the failure to report negative evaluations, whatever threshold is in use.

In these situations, one solution is to compare raters' reports to their peers' reports and reward agreement.[2] However, dangers arise, if rewards are made part of the process. If a particular outcome is highly likely, such as a positive experience with a seller at eBay who has a stellar feedback history, then a rater who has a bad experience will still believe that the next rater is likely to have a good experience. If she will be rewarded simply for agreeing with her peers, she will not report her bad experience. This phenomenon is akin to the problems of herding or information cascades.

In this paper, we develop a formal mechanism to implement the process of comparing with peers. We label this mechanism the peer-prediction method. The scheme uses one rater's report to update a probability distribution for the report of someone else, whom we refer to as the reference rater. The first rater is then scored not on agreement between the ratings, but on a comparison between the *likelihood* assigned to the reference rater's possible ratings and the reference rater's actual rating. Raters need not perform any complex computations: so long as a rater trusts that the center will update appropriately, she will prefer to report honestly.

Scores can be converted to monetary incentives, either as direct payments or as discounts on future merchandise purchases. In many online systems, however, raters seem to be quite motivated by prestige or privileges within the system. For example, at Slashdot.org, users accumulate karma points for various actions and higher karma entitles users to rate others' postings and to have their own postings begin with higher ratings (Lampe and Resnick 2004); at ePinions.com, reviewers gain status and have their reviews highlighted if they accumulate points. Similarly, offline point systems that do not provide any tangible reward seem to motivate chess and bridge players to compete harder and more frequently.

The key insight—that the correlation in agents' private information can be used to induce truthful revelation—has been addressed, albeit in an abstract way, in the mechanism design literature. Seminal papers by d'Aspremont and Gérard-Varet (1979, 1982) and Crémer and McLean (1985, 1988) demonstrate that it is generally possible to use budget-balancing transfer payments to extract agents' private information. Adapting tools from statistical decision theory, Johnson et al. (1990) show how to construct budget-balancing transfer payments based on "proper scoring rules." Johnson et al. (2003) extend those results to the case of multidimensional, continuous private information. Kandori and Matsushima (1998, §4.2) consider how to enforce cooperation in repeated games through correlated equilibria despite the lack of public information about stage game outcomes, and show how to apply a proper scoring rule to elicit truthful communication of private information about stage game outcomes.

This paper applies the *general* insights on the usefulness of proper scoring rules for eliciting correlated information to the *particular* problem of eliciting honest reviews of products, papers, and proposals. Our mechanism is well suited to Internet-based implementations, and it could potentially be applied to services such as NetFlix or Amazon.[3] Once ratings are collected and distributed electronically, it is relatively easy to compute posteriors and scores and keep track of payments.[4]

In §2 we construct payments based on proper scoring rules that allow the center to elicit the rater's private information and show how the payments can be adapted to address costly effort elicitation and budget balance and voluntary participation requirements. Section 3 extends our approach to scenarios of sequential reporting and of discrete reporting based on continuous signals. In §4 we address practical issues that would arise in implementing proper scoring rules in real systems, including conflicts of interest, estimating the information the mechanism requires from historical reviewing data, and accommodating differences among raters in both tastes and in prior beliefs. We also discuss limitations of the mechanism. Section 5 concludes. Proofs and supporting materials are contained in the appendices.

2. A Mechanism for Eliciting Honest Feedback

A number of raters experience a product and then rate its quality. The product's quality does not vary,

[2] Subjective evaluations of ratings could be elicited directly instead of relying on correlations between ratings. For example, the news and commentary site Slashdot.org allows metamoderators to rate the ratings of comments given by regular moderators. Metaevaluation incurs an obvious inefficiency, because the effort to rate evaluations could presumably be put to better use in rating comments or other products that are a site's primary product of interest. Moreover, metaevaluation merely pushes the problem of motivating effort and honest reporting up one level, to ratings of evaluations. Thus, scoring evaluations in comparsion to other evaluations is preferable.

[3] It could also be extended to eBay or Bizrate, which rate sellers rather than products. Rating sellers, however, complicates the analysis. For example, if sellers strategically vary the quality of service they provide over time, the correlation between one rater's evaluation and future raters' evaluations might be severed, disrupting our scoring mechanism.

[4] Prelec's Information Pump (2001) exploits correlated information and proper scoring rules to elicit honest reports in a different setting, estimating the additional information provided by a sequence of true-false statements about an object.

but is observed with some idiosyncratic error. After experiencing the product, each rater sends a message to a common processing facility called the center. The center makes transfers to each rater, awarding or taking away points based on the raters' messages. The center has no independent information, so its scoring decisions can depend only on the information provided by other raters. As noted above, points may be convertible to money, discounts, or privileges within the system, or merely to prestige. We assume that raters' utilities are linear in points.[5] We refer to a product's quality as its type. We refer to a rater's perception of a product's type as her signal.

Suppose that the number of product types is finite, and let the types be indexed by $t = 1, \ldots, T$. Let $p(t)$ be the commonly held prior probability assigned to the product's being type t.[6] Assume that $p(t) > 0$ for all t, and $\sum_{t=1}^{T} p(t) = 1$.

Let I be the set of raters, where $|I| \geq 3$. We allow for the possibility that I is (countably) infinite. Each rater privately observes a signal of the product's type.[7] Conditional on the product's type, raters' signals are independent and identically distributed. Let S^i denote the random signal received by rater i. Let $S = \{s_1, \ldots, s_M\}$ be the set of possible signals, and let $f(s_m \mid t) = \Pr(S^i = s_m \mid t)$, where $f(s_m \mid t) > 0$ for all s_m and t, and $\sum_{m=1}^{M} f(s_m \mid t) = 1$ for all t. We assume that $f(s_m \mid t)$ is common knowledge, and that the conditional distribution of signals is different for different values of t. Let $s^i \in S$ denote a generic realization of S^i. We use s^i_m to denote the event $S^i = s_m$. We assume that raters are risk neutral and seek to maximize expected wealth.

To illustrate throughout this section, we introduce a simple example. There are only two product types, H and L, with prior $p(H) = 0.5$, and two possible signals, h and l, with $f(h \mid H) = 0.85$ and $f(h \mid L) = 0.45$. Thus, $\Pr(h) = 0.5 * 0.85 + 0.5 * 0.45 = 0.65$.

In the mechanism we propose, the center asks each rater to announce her signal. After all signals are announced to the center, they are revealed to the other raters and the center computes transfers. We refer to this as the simultaneous reporting game. Let $a^i \in S$ denote one such announcement, and $a = (a^1, \ldots, a^I)$ denote a vector of announcements, one by each rater. Let $a^i_m \in S$ denote rater i's announcement when her signal is s_m, and $\bar{a}^i = (a^i_1, \ldots, a^i_M) \in S^M$ denote rater i's announcement strategy. Let $\bar{a} = (\bar{a}^1, \ldots, \bar{a}^I)$ denote a vector of announcement strategies. As is customary, let the superscript "$-i$" denote a vector without rater i's component.

[5] We consider the impacts of risk aversion in §4.1.

[6] We briefly address the issue of noncommon priors in §4.5.

[7] We refer to raters as female and to the center as male.

Let $\tau_i(a)$ denote the transfer paid to rater i when the raters make announcements a, and let $\tau(a) = (\tau_1(a), \ldots, \tau_I(a))$ be the vector of transfers made to all agents. An announcement strategy $\bar{a}^i$ is a best response to $\bar{a}^{-i}$ for player i if, for each m,

$$E_{S^{-i}}[\tau_i(\bar{a}^i_m, \bar{a}^{-i}) \mid s^i_m] \geq E_{S^{-i}}[\tau_i(\hat{a}^i, \bar{a}^{-i}) \mid s^i_m] \quad \text{for all } \hat{a}^i \in S. \quad (1)$$

That is, a strategy is a best response if, conditional on receiving signal s_m, the announcement specified by the strategy maximizes that rater's expected transfer, where the expectation is taken with respect to the distribution of all other raters' signals conditional on $S^i = s_m$. Given transfer scheme $\tau(a)$, a vector of announcement strategies $\bar{a}$ is a Nash equilibrium of the reporting game if (1) holds for $i = 1, \ldots, I$, and a strict Nash equilibrium if the inequality in (1) is strict for all $i = 1, \ldots, I$.

Truthful revelation is a Nash equilibrium of the reporting game if (1) holds for all i when $a^i_m = s_m$ for all i and all m, and is a strict Nash equilibrium if the inequality is strict. That is, if all the other players announce truthfully, truthful announcement is a strict best response. Because raters receive no direct return from their announcement, if there were no transfers at all, then any strategy vector, including truthful revelation, would be a Nash equilibrium. However, because players are indifferent among all strategies when there are no transfers, this Nash equilibrium is not strict.

2.1. The Base Case

Our base result defines transfers that make truthful revelation a strict Nash equilibrium. Because all raters experience the same product, it is natural to assume that their signals are dependent. Our results rely on a form of dependence we call stochastic relevance.[8]

Definition. Random variable S^i is *stochastically relevant* for random variable S^j if and only if the distribution of S^j conditional on S^i is different for different realizations of S^i. That is, S^i is stochastically relevant for S^j if for any distinct realizations of S^i, call them s^i and $\hat{s}^i$, there exists at least one realization of S^j, call it s^j, such that $\Pr(s^j \mid s^i) \neq \Pr(s^j \mid \hat{s}^i)$.

Stochastic relevance is almost always satisfied when different types of products generate different signal distributions, as we assumed above, and so throughout the paper we assume that stochastic relevance holds for all S^i and S^j.[9]

[8] The term "stochastic relevance" is introduced in Johnson et al. (2003). It is the same as condition (A4) used in Kandori and Matsushima (1998).

[9] In Miller et al. (2005), we show that stochastic relevance is generically satisfied in product-rating environments.

Continuing the two-type, two-signal example, suppose that rater i receives the signal l. Recall that $p(H) = 0.5$, $f(h \mid H) = 0.85$, and $f(h \mid L) = 0.45$, so that $\Pr(s_l^i) = 0.35$. Given i's signal, the probability that rater j will receive a signal h is

$$g(s_h^j \mid s_l^i) = f(h \mid H)\frac{f(l \mid H)p(H)}{\Pr(s_l^i)} + f(h \mid L)\frac{f(l \mid L)p(L)}{\Pr(s_l^i)}$$
$$= 0.85\frac{0.15 * 0.5}{0.35} + 0.45\frac{0.55 * 0.5}{0.35} \cong 0.54.$$

If i had instead observed h, then

$$g(s_h^j \mid s_h^i) = f(h \mid H)\frac{f(h \mid H)p(H)}{\Pr(s_h^i)} + f(h \mid L)\frac{f(h \mid L)p(L)}{\Pr(s_h^i)}$$
$$= 0.85\frac{0.85 * 0.5}{0.65} + 0.45\frac{0.45 * 0.5}{0.65} \cong 0.71.$$

A scoring rule is a function $R(s^j \mid a^i)$ that for each possible announcement a^i of S^i, assigns a score to each possible realization of S^j. A scoring rule is strictly proper if rater i uniquely maximizes her expected score by announcing the true realization of S^i.

The literature discusses a number of strictly proper scoring rules.[10] The three best known are as follows.

1. Quadratic scoring rule:

$$R(s_n^j \mid a^i) = 2g(s_n^j \mid a^i) - \sum_{h=1}^{M} g(s_h^j \mid a^i)^2.$$

2. Spherical scoring rule:

$$R(s_n^j \mid a^i) = \frac{g(s_n^i \mid a^i)}{(\sum_{h=1}^{M} g(s_h^j \mid a^i)^2)^{1/2}}.$$

3. Logarithmic scoring rule:

$$R(s_n^j \mid a^i) = \ln g(s_n^j \mid a^i).$$

Further, if $R(\cdot \mid \cdot)$ is a strictly proper scoring rule, then a positive affine transformation of it, i.e., $\alpha R(\cdot \mid \cdot) + \beta$, $\alpha > 0$, is also a strictly proper scoring rule. The ability of the center to manipulate α and β is useful in inducing the raters to exert effort and satisfying their participation constraints (see §2.2). We will use $R(s_n^j \mid a^i)$ to denote a generic strictly proper scoring rule. At times we will illustrate our results using the logarithmic rule because of its intuitive appeal and notational simplicity. However, unless otherwise noted, all results hold for any strictly proper scoring rule.

Transfers based on a strictly proper scoring rule induce truthful revelation by agent i as long as her private information is stochastically relevant for some other publicly available signal. However, in our case, each rater's signal is private information, and therefore we can only check players' announcements against other players' announcements, not their actual signals. For each rater, we will choose a reference rater $r(i)$, whose announcement i will be asked to predict. Let

$$\tau_i^*(a^i, a^{r(i)}) = R(a^{r(i)} \mid a^i). \tag{2}$$

Proposition 1. *For any mapping r that assigns to each rater i a reference rater $r(i) \neq i$, and for any proper scoring rule R, truthful reporting is a strict Nash equilibrium of the simultaneous reporting game with transfers τ_i^*.*

Proof of Proposition 1. Assume that rater $r(i)$ reports honestly: $a^{r(i)}(s_m) = s_m$ for all m. S^i is stochastically relevant for $S^{r(i)}$, and $r(i)$ reports honestly, so S^i is stochastically relevant for $r(i)$'s report as well. For any $S^i = s^*$, player i chooses $a^i \in S$ in order to maximize

$$\sum_{n=1}^{M} R(s_n^{r(i)} \mid a^i) g(s_n^{r(i)} \mid s^*). \tag{3}$$

Because $R(\cdot \mid \cdot)$ is a strictly proper scoring rule, (3) is uniquely maximized by announcing $a^i = s^*$. Thus, given that rater $r(i)$ is truthful, rater i's best response is to be truthful as well. □

We illustrate Proposition 1 using the logarithmic scoring rule. Because $0 < g(s_m^j \mid s_n^i) < 1$, $\ln g(s_m^j \mid s_n^i) < 0$; we refer to τ_i^* as rater i's penalty because it is always negative in this case. Consider the simple example where rater i received the relatively unlikely signal l ($\Pr(s_l^i) = 0.35$). Even contingent on observing l, it is unlikely that rater j will also receive an l signal ($g(s_l^j \mid s_l^i) = 1 - 0.54 = 0.46$). Thus, if rater i were rewarded merely for matching her report to that of rater j, she would prefer to report h. With the log scoring rule, an honest report of l leads to an expected payoff

$$\ln g(s_h^j \mid l) g(s_h^j \mid l) + \ln g(s_l^j \mid l) g(s_l^j \mid l)$$
$$= \ln(0.54)0.54 + \ln(0.46)0.46 = -0.69.$$

If, instead, she reports h, rater i's expected score is

$$\ln g(s_h^j \mid h) g(s_h^j \mid l) + \ln g(s_l^j \mid h) g(s_l^j \mid l)$$
$$= \ln(0.71)0.54 + \ln(0.29)0.46 = -0.75.$$

As claimed, the expected score is maximized by honest reporting.

The key idea is that the scoring function is based on the updated beliefs about the reference rater's signal, given the rater's report. The updating takes into account both the priors and the reported signal, and thus reflects the initial rater's priors. Thus, she has no

[10] See Cooke (1991, p. 139) for a discussion of strictly proper scoring rules. Selten (1998) provides proofs that each of the three rules is strictly proper and discusses other strictly proper scoring rules.

reason to shade her report toward the signal expected from the priors. Note also that she need not perform any complex Bayesian updating. She merely reports her signal. As long as she trusts the center to correctly perform the updating and believes other raters will report honestly, she can be confident that honest reporting is her best action.[11]

Note that while Proposition 1 establishes that there is a truthful equilibrium, it is not unique, and there may be nontruthful equilibria. To illustrate, in the example we have been considering, two other equilibria are (1) report h all the time, and (2) report l all the time.[12] While such nontruthful equilibria exist, it is reasonable to think that the truthful equilibrium will be a focal point, especially when communication among raters is limited, or when some raters are known to have a strong ethical preference for honesty. In addition, the center can punish all the raters if he detects a completely uninformative equilibrium such as all h or all l.

2.2. Eliciting Effort and Deterring Bribes

Assuming costless evaluation and reporting allowed us to focus on the essence of the scoring-rule-based mechanism. However, raters' willingness to exert effort will depend on the direct costs of effort as well as the opportunity cost of being an early evaluator rather than free riding off the evaluations of others. Avery et al. (1999) explore how market mechanisms can elicit costs and determine appropriate compensation levels, but the assumption that raters will exert effort once they accept compensation is problematic.[13] Here, we use a scoring rule to induce effort. We begin by assuming a fixed cost of rating. We then move on to consider how the center can induce raters to select an optimal effort level when additional costly effort leads to more precise signals.

Suppose there is a fixed cost, $c > 0$, of evaluating and reporting. To induce effort, the expected value of incurring effort and reporting honestly must exceed the expected value of reporting without a signal. As the proof of Proposition 1 makes clear, the truth-inducing incentives provided by scoring-rule based payments are unaffected by a positive rescaling of all transfers: if transfers $\tau_i^*(a^i, a^{r(i)}) = R(a^{r(i)} \mid a^i)$ induce truthful reporting, then $\tau_i^*(a^i, a^{r(i)}) = \alpha R(a^{r(i)} \mid a^i)$, where $\alpha > 0$, does as well. Since the rater is better informed if she acquires a signal than if she doesn't, and better information always increases the expected value of a decision problem (Savage 1954, Lavalle 1968), increasing the scaling factor increases the value of effort without affecting the incentives for honest reporting once effort is expended.

Proposition 2. *Let $c > 0$ denote the cost of acquiring and reporting a signal. If other raters acquire and report their signals honestly, there exists a scalar $\alpha > 0$ such that when rater i is paid according to $\tau_i^*(a^i, a^{r(i)}) = \alpha R(a^{r(i)} \mid a^i)$, her best response is to acquire a signal and report it honestly.*[14]

Scaling can be used to induce raters to work harder to obtain better information. Without putting additional structure on the distributions under consideration, the natural notion of "better" information is to think about the rater's experience as being a random sample, with better information corresponding to greater sample size. If the cost of acquiring a sample is increasing and convex in its size, we can ask when and how the center can induce the raters to acquire samples of a particular size.

Because of space considerations, we relegate the technical presentation to Appendix B. However, the basic idea is straightforward.[15] For any sample size, stochastic relevance continues to hold. Thus, when the rater is paid according to a strictly proper scoring rule, she maximizes her expected score by truthfully announcing her information (if all other raters do as well). When a rater increases her sample size from, say, x to $x + 1$, the additional observation further partitions the outcome space. Using well-known results from decision theory (Savage 1954, Lavalle 1968), this implies that the rater's optimized expected score increases in the sample size. Let $V^*(x)$ denote optimized expected score as a function of sample size. The question of whether the center can induce the rater to choose a particular sample size, x^*, then comes down to whether there exists a scaling factor, α^*, such that

$$x^* \in \arg\max_x \alpha^* V^*(x) - c(x).$$

If $V^*(x)$ is concave in x and $c(x)$ satisfies certain regularity conditions (i.e., $c'(0) = 0$, and $\lim_{x\to\infty} c'(x) = \infty$), it is possible to induce the agent to choose any

[11] In an experiment, Nelson and Bessler (1989) show that even when the center does not perform the updating for them, with training and feedback subjects learn that truthful revelation is a best response when rewards are based on a proper scoring rule.

[12] To verify the "always play h equilibrium," note that if the reference rater always reports high, the rater expects $\ln(0.54)1 + \ln(0.46)0 = -0.61619$ if she reports l, and $\ln(0.71)1 + \ln(0.29)0 = -0.34249$ if she reports h. Similar reasoning verifies the "always play l equilibrium."

[13] At the news and commentary site Slashdot.org, where users earn karma points for acting as moderators, staff have noticed that occasionally ratings are entered very quickly in succession, faster than someone could reasonably read and evaluate the comments. They call this "vote dumping."

[14] Proofs not included in the main text are in Appendix A.

[15] Clemen (2002) undertakes a similar analysis in the context of a principal-agent problem.

desired sample size. We return to the question of eliciting effort in §3.2.1, where, because it is assumed that information is normally distributed, we are able to present the theory more parsimoniously.

Scaling can also be used to overwhelm individuals' outside preferences, including bribes that may be offered for positive ratings. For example, if a bribe has been offered for a positive rating, the constant c can be interpreted to include the potential opportunity cost of acquiring a negative signal and then reporting it.

2.3. Voluntary Participation and Budget Balance

In some cases, the expected payment from truthful reporting (and optimal effort) may be insufficient to induce the rater to participate in the mechanism in the first place. This is most apparent when the logarithmic rule is employed, because the logarithmic score is always negative. However, this problem is easily addressed. Because adding a constant to all payments (i.e., letting the transfer be $\alpha_i R(a^{r(i)} \mid a^i) + k_i$) does not affect incentives for effort or honest reporting, the constant k_i can be chosen to satisfy either ex ante participation constraints (i.e., each agent must earn a nonnegative expected return), interim participation constraints (i.e., each agent must earn a nonnegative expected return conditional on any observed signal), or ex post participation constraints (i.e., the agent must earn a nonnegative expected return for each possible (s^j, s^i) pair). To illustrate using the logarithmic case, let $\tau_0 = \min_{s_m, s_n \in S}(\alpha \ln g(s_m \mid s_n))$ and define $\tau^+ = \tau^* - \tau_0$. Transfers τ^+ will attract voluntary (ex post) participation while still inducing effort and honest reporting.

It is often desirable for the center to balance his budget. Clearly, this is important if scores are converted into monetary payments. Even if scores are merely points that the center can generate at will, uncontrolled inflation would make it hard for users to interpret point totals. If there are at least three raters, the center can balance the budget by reducing each rater's base transfer τ^* by some other rater's base transfer. Though all transactions actually occur between raters and the center, this creates the effect of having the raters settle the transfers among themselves.[16] Let $b(i)$ be the rater whose base transfer i settles (paying if τ^* is positive, and collecting if it is negative), and let $b(i)$ be a permutation such that $b(i) \neq i$ and $r(b(i)) \neq i$. Rater i's net transfer is

$$\tau_i(a) = \tau_i^*(a^i, a^{r(i)}) - \tau_{b(i)}^*(a^{b(i)}, a^{r(b(i))}). \tag{4}$$

These transfers balance. The only raters whose reports can influence the second term are $b(i)$ and rater $b(i)$'s reference rater, $r(b(i))$, and by construction of $b(\cdot)$, they are both distinct from rater i. Because all reports are revealed simultaneously, rater i also cannot influence other players' reports through strategic choice of her own report. Thus, the second term in (4) does not adversely affect rater i's incentive to report honestly or put forth effort.

The balanced transfers in (4) do not guarantee voluntary participation. In some cases, a rater's net transfer may be negative. One way to assure ex post voluntary participation is to collect bonds or entry fees in advance, and use the collected funds to ensure that all transfers are positive. For example, with the logarithmic scoring rule, $\min \tau \leq \min \tau^* = \tau_0$. If $-\tau_0$ is collected from each player in advance and then returned with the transfer τ, each player will receive positive payments after the evaluations are reported. Some raters will still incur net losses, but their bonds prevent them from dropping out after they learn of their negative outcome. Alternatively, it may be sufficient to threaten to exclude a rater from future participation in the system if she is unwilling to act as a rater or settle her account after a negative outcome.

3. Extensions

We now consider two extensions to the base model. In the first, raters report sequentially rather than simultaneously. In the second, their types and signals are continuous rather than discrete.

3.1. Sequential Interaction

Sequential reporting may be desirable because it allows later raters to make immediate use of the information provided by their predecessors. The mechanism adapts readily to sequential situations.[17] Rater i's transfer can be determined using any subsequent rater as a reference rater. To balance the budget, the transfer can be settled by any subsequent rater other than rater i's reference rater.

For example, suppose an infinite sequence of raters, indexed by $i = 1, 2, \ldots$, interacts with the product. Let rater $i+1$ be rater i's reference rater, i.e., i's report is used to predict the distribution of rater $i+1$'s report. Let $p(t)$ be the initial, commonly held prior distribution for the product's type. Let $p_1(t \mid s^1)$ denote the posterior distribution after rater 1 receives signal s^1. This can be computed using Bayes' rule in the usual way. Rater 1's posterior belief about the probability

[16] Because each player will receive her own base transfer and fund one other player's, the addition of τ_0 to each has no net effect, so we phrase the discussion in terms of the raw penalties τ^* rather than the net payments τ^+.

[17] Hanson (2002) applies a scoring-rule-based approach in a model in which a number of experts are sequentially asked their belief about the distribution of a random event, whose realization is revealed after all experts have reported. In our model, the product's type is never revealed, and therefore we must rely on other agents' reports to provide incentives.

that $S^2 = s^2$ when $S^1 = s^1$ is then given by $g(s^2 \mid s^1) = \sum_{t=1}^{T} f(s^2 \mid t)p_1(t \mid s^1)$. Using this distribution (and still assuming stochastic relevance), rater 1 can be induced to truthfully reveal s^1 using the scoring rule specified in Proposition 1. After rater 1 announces her signal, this information is made public and is used to update beliefs about the product's type.

This process can be iterated. When rater i is asked to announce her signal, the "prior" distribution over types takes into account all previous announcements. Incentives to rater i are constructed using a scoring rule that incorporates these updated beliefs, i.e., rater i is scored using a strictly proper scoring rule applied to the distribution implied by rater i's announcement and the current beliefs about the product's type (which incorporates the announcements of the first $i-1$ raters). To balance the budget, rater i's transfer could be paid by rater $i+2$.

When a finite string of raters experience the product, the last rater has no incentive to lie, but also none to tell the truth, because there is no future signal upon which to base her reward. Thus, there is a danger of the whole process unravelling. Fortunately, the center can solve this problem by grouping some raters together and treating group members as if they report simultaneously. For example, suppose there are 10 raters. Consider the last three—8, 9, and 10. The center can score rater 8 based on 9's announcement, 9 based on 10's, and 10 based on 8's. As long as the center can avoid revealing these three raters' announcements until all three have announced, effective incentives can be provided using our earlier techniques, and the chain will not unravel. Transfers can also be made within the ring to balance the budget for the ring.

3.2. Continuous Signals

Until now, we have considered discrete type and signal spaces. All of our results translate to the continuous case in a natural way (e.g., density functions replace discrete distributions, integrals replace sums, etc.). For example, if rater i reports signal s^i, the logarithmic score is computed as $\ln(g(s^j \mid s^i))$, where $g(s^j \mid s^i)$ is now the posterior *density* of $S^j = s^j$ given $S^i = s^i$. Most importantly, the scoring rules we have discussed continue to be strictly proper in the continuous case.

In this section, we briefly consider two particularly interesting aspects of the problem with continuous signals and product-type spaces, a comparison of the three scoring rules when prior and sample information are normally distributed, and the problem of eliciting discrete information when signals are continuous.

3.2.1. Effort Elicitation with Normally Distributed Noise: A Comparison of Scoring Rules. Let q denote the unknown quality of the good, and suppose that raters have prior beliefs that q is normally distributed with mean μ and precision θ_q, where precision equals 1/variance. Suppose each rater observes a real-valued signal S^i of the object's quality that is normally distributed with mean q and precision θ_i. That is, each rater receives a noisy but unbiased signal of the object's quality. Conditional on observing $S^i = s^i$, the rater's posterior belief about q is that q is distributed normally with mean $\hat{\mu} = (\mu\theta_q + s^i\theta_i)/(\theta_q + \theta_i)$ and precision $\hat{\theta} = \theta_q + \theta_i$.[18]

Suppose that rater j observes signal S^j on the object's quality, where S^j is normally distributed with mean q and precision θ_j. Conditional on observing $S^i = s^i$, rater i's posterior belief about the distribution of S^j is that S^j is normally distributed with mean $\hat{\mu}$ and precision $\theta = \hat{\theta}\theta_j/(\hat{\theta} + \theta_j)$.[19]

Because different observation-precision combinations lead to different posterior beliefs about the distribution of S^j, assuming stochastic relevance continues to be reasonable in the continuous case. If we make this assumption, then payments based on a proper scoring rule can induce effort and honest reporting. As before, rater i will prefer to be scored on her posterior for the reference rater j, which is achieved by honestly reporting her observation and her precision, allowing the center to correctly compute her posterior.[20]

We assume that by exerting effort, raters can increase the precision of their signals. Let $c(\theta_i)$ represent the cost of acquiring a signal of precision $\theta_i \geq 0$, where $c'(\theta_i) > 0$, $c'(0) = 0$, $c'(\infty) = \infty$, and $c''(\theta_i) \geq 0$. To compare the logarithmic, quadratic, and spherical scoring rules, it is necessary to ensure that the rater is choosing the same signal precision under each rule. As suggested by our analysis in §2.2, the center can induce the rater to choose more or less effort by multiplying all transfers by a larger or smaller constant.

Let $f(x)$ be the probability density function of a normal random variable with mean μ and precision θ.

[18] See Pratt et al. (1965).

[19] The variance of S^j conditional on S^i is the sum of the variance of the posterior belief about q, $1/\hat{\theta}$, and the variance of S^j conditional on q, $1/\theta_j$, which implies precision $\theta = \hat{\theta}\theta_j/(\hat{\theta} + \theta_j)$.

[20] Ottaviani and Sørensen (2004) consider a related model, with normally distributed information of fixed precision for each rater. In their analysis, however, each rater attempts to convince the world of their expertise (i.e., that they have precise signals). With that objective function, there is no equilibrium where signals are fully revealed. By contrast, we introduce an explicit scoring function that is not based solely on the inferred or reported precision of raters' signals, and full information revelation can be induced.

Table 1

Rule	Variance of transfers	Min	Max	Range
Log	$2A^2c'(\theta_i)^2$	$-\infty$	$A\log\left(\frac{\theta}{2\pi}\right)c'(\theta_i)$	∞
Quadratic	$\frac{16(2\sqrt{3}-3)}{3}A^2c'(\theta_i)^2$	$-2Ac'(\theta_i)$	$2(2\sqrt{2}-1)Ac'(\theta_i)$	$4\sqrt{2}Ac'(\theta_i)$
Spherical	$\frac{16(2\sqrt{3}-3)}{3}A^2c'(\theta_i)^2$	0	$4\sqrt{2}Ac'(\theta_i)$	$4\sqrt{2}Ac'(\theta_i)$

where $A=\frac{(\theta_i+\theta_q)(\theta_i+\theta_j+\theta_q)}{\theta_j}$.

Under the logarithmic scoring rule, the maximized expected utility as a function of precision (i.e., when the rater announces truthfully) is given by

$$v_l(\theta_i)=\int \log(f(x))f(x)dx=-\frac{1}{2}+\frac{1}{2}\log\left(\frac{\theta}{2\pi}\right).$$

It is straightforward to verify that $v_l(\theta_i)$ is increasing and concave in θ_i. Thus, as in the discrete case, by varying the multiplicative scaling factor, the center can induce the rater to choose any particular level of precision.

The scaling factor α that induces a particular θ_i is found by solving

$$\max_{\theta_i}\alpha\left(-\frac{1}{2}+\frac{1}{2}\log\left(\frac{\theta}{2\pi}\right)\right)-c(\theta_i).$$

Setting the derivative of this expression equal to zero yields that choosing $\alpha=\frac{2}{\theta_j}(\theta_q+\theta_i)(\theta_q+\theta_i+\theta_j)\cdot c'(\theta_i)\equiv\alpha_l$ induces precision θ_i under the logarithmic rule. Analogous calculations for the quadratic and spherical scoring rules find that to induce precision θ_i, $\alpha=(4\pi^{1/2}/\theta_j^{3/2})(\theta_q+\theta_i)^{1/2}(\theta_q+\theta_i+\theta_j)^{3/2}c'(\theta_i)$ and $\alpha=((4\sqrt{2}\pi^{1/4})/\theta_j^{5/4})(\theta_q+\theta_i)^{3/4}(\theta_q+\theta_i+\theta_j)^{5/4}c'(\theta_i)$ respectively.

Based on these choices for α, the variance and range of the transfers under each of the rules is shown in Table 1.[21]

Two notable features emerge from this analysis. First, the quadratic and spherical rules have the same variance and range of payments. This is because both rules specify scores that are linear in $f(x)$, and so, once scaled to induce the same precision, they differ only by an additive constant. Second, while the logarithmic rule has the smallest variance ($16/3(2\sqrt{3}-3)\simeq 2.4752$), its the range of payments is infinite because $\lim_{x\to 0}\ln(x)=-\infty$. We refer to these results in §4.2, where we discuss how to choose among the scoring rules in particular application contexts.

[21] Supporting computations for Table 1 are available from the authors upon request.

3.2.2. Eliciting Coarse Reports. Raters' information is often highly nuanced. Yet, systems often employ coarser measures of quality, such as 1 to 5 stars. In this section, we consider situations where the center offers raters a choice between several "coarse" reports, and analyze whether it is possible to design payments that induce people to be as truthful as possible, i.e., to choose the admissible report closest to their true signal.

The coarse reporting problem is both subtle and complex. Proper scoring rules induce people to truthfully announce their exact information. One might hope that in a sufficiently smooth environment, a rater offered a restricted set of admissible reports will choose the one that is "closest" to her true information. However, this intuition relies on two assumptions: that closeness in signals corresponds to closeness in posteriors over product types, and that close beliefs in product-type space correspond to close beliefs about the distribution of a reference rater's announcement. Although it remains an open question whether these assumptions hold in general, it is possible to show that they hold when there are only two types of products.

Suppose raters receive signals drawn from the unit interval and that there are only two types of objects, good (type G) and bad (type B). Their signal densities are $f(s\mid G)$ and $f(s\mid B)$. Let $p\in(0,1)$ denote the prior probability (commonly held) that the object is good. We assume that densities $f(s\mid G)$ and $f(s\mid B)$ satisfy the Monotone Likelihood Ratio Property (MLRP), i.e., $f(s\mid G)/f(s\mid B)$ is strictly increasing in s.

MLRP implies the distribution for type G first order stochastically dominates the distribution for B (see Gollier 2001). If rater i observes signal $S^i=s^i$, she assigns posterior probability $p(G\mid s^i)=pf(s^i\mid G)/(pf(s^i\mid G)+(1-p)f(s^i\mid B))$ to the object's being good. MLRP ensures that $p(G\mid s^i)$ is strictly increasing in s^i. Thus, MLRP embodies the idea that higher signals provide stronger evidence that the object is good.

We divide the signal space into a finite number of intervals, which we call bins, and construct a scoring rule such that rater i's best response is to announce

the bin in which her signal lies, if she believes that all other raters will do the same. The construction of reporting bins and a scoring rule capitalizes on a special property of the quadratic scoring rule. Friedman (1983) develops the notion of "effective" scoring rules. A scoring rule is effective with respect to a metric if the expected score from announcing a distribution increases as the announced distribution's distance from the rater's true distribution decreases. When distance between distributions is measured using the L_2 metric, the quadratic scoring rule has this property. Also, when there are only two types, the L_2 distance between two distributions of reference raters' announcements is proportional to the product type beliefs that generate them (if such beliefs exist).

Proposition 3. *Suppose there are two types of objects with signal densities that satisfy MLRP. Then, for any integer L, there exists a partition of signals into L intervals and a set of transfers that induce Nash equilibrium truthful reporting when agents can report only in which interval their signal lies.*

The essence of the proof of Proposition 3, which appears in Appendix A, is as follows. After observing $S^i = s^i$, rater i's belief about the product's type (PT belief) is summarized by rater i's posterior probability that the product is good, $p(G \mid s^i)$. We begin by dividing the space of PT beliefs into L equal-sized bins. Since $p(G \mid s^i)$ is monotone, these PT-belief bins translate to intervals in the rater's signal space, which we refer to as signal bins. Signal bins can differ in size. A rater who announces her signal is in the l^{th} bin of signals is treated as if she had announced beliefs about the product type at the midpoint of the l^{th} PT bin, which implies some distribution for the reference rater's announcement (RRA). Each signal bin announcement thus maps to PT beliefs and then to an RRA distribution. The RRA distribution is scored using the quadratic rule.

Because the quadratic scoring rule is effective, given a choice among this restricted set of admissible RRA distributions, the rater chooses the RRA distribution nearest (in the L_2 metric) to her true one. This turns out to be the one with PT belief nearest her true PT belief. If s^i is in the l^{th} signal bin, the closest available PT belief is the midpoint of the l^{th} PT bin. Thus, given coarse bins, the quadratic scoring rule induces truthful (albeit coarse) bin announcements.

Note that the bins are constructed by dividing the PT space rather than the signal space into equal-sized bins. While closeness of PT beliefs corresponds to closeness of RRA beliefs, close signals do not translate linearly to close PT beliefs. For example, suppose a rater observes signal $s^i = 0.5$, and that $p(G \mid 0.5) = 0.3$. It is possible that $p(G \mid 0.4) = 0.2$ while $p(G \mid 0.6) = 0.35$. Thus, although the distance between signals 0.5 and 0.6 is the same as the distance between signals 0.5 and 0.4, the PT beliefs (and therefore the RRA beliefs) are closer for the first pair than for the second.

Even in the simple case with only two product types, it is somewhat complicated to show that raters will want to honestly reveal their coarse information. It remains an open question whether it is possible to elicit honest coarse reports in more complex environments.

4. Issues in Practical Application

Sections 2 and 3 provide a theoretical framework for inducing effort and honest reporting. Designers of practical systems will face many challenges in applying it. Many of these challenges can be overcome with adjustments in the transfer payment scheme, computation of parameters based on historical data, and careful choice of the dimensions on which raters are asked to report.

4.1. Risk Aversion

Until now, we have assumed that raters are risk neutral, i.e., that maximizing the expected transfer is equivalent to maximizing expected utility. If raters are risk averse, then scoring-rule-based transfers will not always induce truthful revelation. We present three ways to address risk aversion.

If the center knows the rater's utility function, the transfers can be easily adjusted to induce truthful reporting. If $U()$ is the rater's utility function and R is a proper scoring rule, then choosing transfers $\tau = U^{-1}(R)$ induces truthful reporting, because $U(U^{-1}(R)) \equiv R$ (Winkler 1969).

If the rater's utility function is not known, risk-neutral behavior can be induced by paying the rater in "lottery tickets" for a binary-outcome lottery instead of in money (Smith 1961, Savage 1971). In effect, the score assigned to a particular outcome gives the probability of winning a fixed prize. Because von Neumann-Morgenstern utility functions are linear in probabilities, an expected-utility maximizer will seek to maximize the expected probability of winning the lottery. Thus the lottery-ticket approach induces individuals with unknown nonlinear utility functions to behave as if they are risk neutral. Experimental evidence suggests that, while not perfect, the binary-lottery procedure can be effective in controlling for risk aversion, especially when raters have a good understanding of how the procedure works.[22]

A third method of dealing with risk-averse raters capitalizes on the fact that raters' risk aversion is

[22] See Roth (1995, pp. 81–83) and the references therein.

likely to be less important when the variability in payments is small. Although we have presented our results for the case where each rater is scored against a single reference rater, the idiosyncratic noise in the rater's final payment (measured in terms of its variance) can be reduced by scoring the rater against multiple raters and paying her the average of those scores. By averaging the scores from a sufficiently large number of reference raters, the center can effectively eliminate the idiosyncratic noise in the reference raters' signals. However, the systematic risk due to the object's type being unknown cannot be eliminated.

4.2. Choosing a Scoring Rule

Which of the three scoring rules we have discussed is best? Each rule has its relative strengths and weaknesses and none emerges as clearly superior.

The logarithmic rule is the simplest, giving it a modest advantage in comprehension and computational ease. It is also "relevant" in the sense that it depends only on the likelihood of events that actually occur.[23] In addition, our results in §3.2.1 show that the payments needed to induce a particular effort level have lower variance under the logarithmic rule than under either of the other two rules, at least when information is normally distributed. If scores are used to evaluate the raters (for example, to decide whether to invite them back as reviewers in the future), this lower variance enables the logarithmic rule to provide a more reliable evaluation given the same number of trials.

On the other hand, the fact that $\log(x)$ goes to $-\infty$ as x decreases to zero renders the log rule unattractive when probabilities become small and raters' limited liability is a concern, or if the support of the raters' posterior distributions changes with their information. On a related note, under the log rule, small changes in low-probability events can significantly affect a rater's expected score, which may be undesirable if raters have difficulty properly assessing low-probability events. A final disadvantage to the logarithmic score is that, in contrast to the quadratic rule, there is no metric with respect to which the logarithmic rule is effective (Nau 1985). That is, a rater's expected score from announcing a particular distribution need not increase as its distance (as measured by any valid metric) from the true distribution decreases.

As discussed above, the quadratic rule is effective with respect to the L_2 metric, which is what allowed us to solve the coarse reporting problem in §3.2.2. The quadratic rule is not relevant, so it can have the perverse property that, given two distributions, the quadratic score may be higher for the distribution that assigns lower probability to the event that actually occurs (Winkler 1996). The spherical rule shares many properties with the quadratic rule (although its payments are always positive). As we saw in the normal-information case, once the spherical and quadratic rules are scaled to induce the same rating effort, they become identical up to an additive constant. The spherical rule is effective with respect to a renormalized L_2 metric (see Friedman 1983).

Jensen and Peterson (1973) compare the three scoring rules in head-to-head experimental trials. They conclude that there is essentially no difference in the probabilities elicited from raters. They do note that subjects seem to have trouble understanding scoring rules involving both positive and negative payments; while the quadratic rule has this property, it is easily addressed by adding a constant to all payments. Thus, except for situations where some events have low probability or raters' information affects the set of possible events (i.e., moving support), factors that make the logarithmic score undesirable, there is no clear reason to prefer one scoring rule over the others.

4.3. Estimating Types, Priors, and Signal Distributions

In many situations, there will be sufficient rating history available for the center to estimate the prior probabilities of alternative types and signals to start the scoring process. One technique would define the product types in terms of the signal distributions they generate. For example, suppose that there are only two signals, h and l. Products are of varying quality, which determines the percentage of users who submit h ratings for the product. The type space is continuous in principle, but in practice the site could approximately capture reality by defining a set of discrete types that partitions the space. For illustrative purposes, we define a fairly coarse partition of types, $1, \ldots, 9$, with $f(h \mid i) = i/10$. That is, products of type 1 get rated h 10% of the time, and those of type 7 get rated h 70% of the time. The site would then estimate the prior distribution function $p(i)$ based on how many products in the past accumulated approximately $10i\%$ ratings.[24]

Table 2 illustrates updating of beliefs about the probability that a product is of any of the nine types. Note that the initial distribution is symmetric about type 5, implying that initial probability of h is 0.5. After receiving a report h, types that have higher frequencies of h signals become more likely, as shown in the second row of the table. After receiving two

[23] Relevance is important in Bayesian models of comparing different probability assessors (Winkler 1969, Staël Von Holstein 1970), although this is not important for our application.

[24] Obviously, the partition could be finer, for example with types 1–99 defined by percentage of raters rating the product h. In addition, the partition need not be uniform: More types could be defined in the region that occur most often on a particular site.

Table 2 Initial and Updated Probabilities of Nine Types Defined by Their Probability of Yielding Signal h.

After signal	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$pr(h)$
	0.05	0.1	0.1	0.1	0.3	0.1	0.1	0.1	0.05	0.5
h	0.01	0.04	0.06	0.08	0.3	0.12	0.14	0.16	0.09	0.59
h, l	0.02	0.08	0.1	0.12	0.36	0.12	0.1	0.08	0.02	0.5

conflicting reports, h and l, the distribution is again symmetric about type 5, but the extreme types are now seen as less likely than they were initially.

4.4. Taste Differences Among Raters

Suppose that raters differ systematically in their tastes. For example, raters of type A might be generally harsher in their assessments than those of type B, so that with binary signals, they would be more likely to perceive goods of any particular type as being low quality, $f_A(l \mid t) > f_B(l \mid t)$. The same problems could arise if the differences among raters' perceptions covaried with the product types. For example, an action movie aficionado might perceive most action movies to be h and most romantic comedies to be l; perceptions would be reversed for fans of comedies.

When tastes differ systematically, the center will need to model rater types explicitly. As in the simpler case in §4.3, given a sufficient history the center can estimate the distribution of user types and for each type the signal distributions. An individual rater's history provides additional information for inferring the distribution from which her type is drawn.[25]

4.5. Noncommon Priors and Other Private Information

The incentives for effort and honest reporting depend critically on the center's ability to compute a posterior distribution for another rater's signal that the current rater would agree with, if only she had the information and computational ability available to the center. Problems may arise if raters have relevant private information beyond their own signals. Knowing that the center will not use that other private information, the rater will no longer be confident that an honest report of her signal will lead to scoring based on her true posterior beliefs about the distribution of another rater's signals. If she can intuit the correct direction, she may distort her reported signal to cause the center to score her based on posterior beliefs closer to what she would compute herself.

Fortunately, the mechanisms in this paper easily adapt if raters can report any private information they have about the distribution of product types, rater types, or signals contingent on product and rater types.[26] The center will use the reference rater's report to compute two scores. The first comes from the distribution implied by the reported private priors; the second is based on the posteriors computed from the priors and the reported signal. An honest report of priors maximizes the first score. The second is maximized when the center calculates accurate posteriors, and that occurs when both priors and signal are honestly reported. Thus, honest reports maximize either score.

In most practical situations, it will not be necessary to elicit all possible private information. Where the center has a sufficient history of past ratings, most raters will trust the center's inferences about the distribution of product types, rater types, and signals conditional on product and rater types. In those cases, raters need only report what they saw. However, when raters may have beliefs that diverge from the center's, it will be useful to offer raters an opportunity to report those beliefs, lest the unreported beliefs create incentives for distorting signal reports.

4.6. Other Potential Limitations

Other potential limitations could interfere with the smooth functioning of a scoring system based on the peer-prediction method. We mention three. First, while we have shown there is a Nash equilibrium involving effort and honest reporting, raters could collude to gain higher transfers. Of course, with balanced transfers it would not be possible for all of the raters to be better off through collusive actions, and it is unclear whether a subset of the raters could collude to gain at the expense of the remaining raters who exerted effort and reported honestly. For example, one rater can gain by knowing what a colluding reference rater will report, but it is not clear whether the gain would outweigh the losses for the colluding reference rater when she is scored against some other, honest rater. Even if such collusion were profitable, the center has two approaches available to deter it. The selection of who will serve as a reference rater for each rater can be randomized and delayed until after ratings are reported, which would make collusion harder to coordinate. In addition, the center may be able to detect suspicious rating patterns through

[25] A variety of recommender systems or collaborative filtering algorithms rely on the past ratings of a set of users to make personalized predictions of how well each individual will like products they have not yet rated. See Breese et al. (1998) and Sarwar et al. (2000) for reviews. Often these algorithms merely predict a scalar value for an individual's rating, but they could be extended to predict a distribution over signals for each rater for each product not yet rated. When an additional rating is added from rater i, the predicted distributions for each other rater for that product would be updated.

[26] Note that for peer-prediction scoring to work, we need to compare one rater's posterior to another rater's reported signal, so it is critical to elicit raters' signals separately from any other information that is also elicited from them.

statistical analysis, and then employ an outside expert to independently evaluate the product.[27]

A second potential limitation may arise when raters perceive multidimensional signals. Our scoring system can easily generalize to handle multiple dimensions by eliciting reports on several dimensions, such as a restaurant's food, decor, and service. Scores can then be computed based on implied distributions for reports on one or all of the dimensions. If, however, some dimensions are not elicited, two problems emerge. First, information may not be captured that could be valuable to consumers. More troubling, in some situations the information not elicited from a rater may be useful in predicting the next report, in which case the rater may be tempted to manipulate the report that is requested.

Consider, for example, an interdisciplinary review panel. An economist with some knowledge of computer science may evaluate proposals as other economists do, but may perceive some additional signal about how computer scientists will perceive the proposals. Suppose she is asked to report only her perception of the proposal's quality. The center then computes an updated distribution of signals for the next rater, accounting for both raters' types as in §4.4. But the economist's secondary signal about how well computer scientists will like the proposal may allow her to compute a more accurate distribution than the center can, and thus she will sometimes want to report dishonestly to make the center more closely approximate her true beliefs.[28]

One solution to this problem would be to find a set of dimensions on which raters are asked to report such that any other signals the raters get are not relevant for predicting the next player's report. For example, if restaurant reviewers are asked to report separately on food, decor, and service, the transfer payments can induce honest reporting as long as any other independent signals that reviewers may receive (such as the number of people in the restaurant that night) are not useful in predicting how other raters will perceive food, decor, or service. On an interdisciplinary review panel, reviewers might be asked to separately report quality from the perspective of each of the disciplines involved. When scores are computed, they can be based on the probabilities for another player's report on any one dimension, or on all of them. Again, because honest reporting will cause the center to correctly compute the rater's beliefs about the reference rater's signal, honest reporting will be an equilibrium. Unfortunately, it may be difficult in practice to find a set of rating dimensions such that unreported signals for a rater are irrelevant to computing beliefs about reported signals for a reference rater.

Given the computational power and the information resources available to the center, it may not be necessary in practice to elicit from raters all of their weakly stochastically relevant signals. For example, suppose the center performs a complex collaborative filtering algorithm to predict the next rater's distribution, and the individual rater either lacks the computational resources or the history of everyone's previous ratings, or does not know in advance which rater she will be scored against. Although an additional private signal might make rater i think that, say, signal h is more likely for some raters than the center would otherwise compute, she will often be unable to determine which false report on the dimensions that the center elicits would raise her payoff.

A third potential limitation is trust in the system: people may not believe that effort and honest reporting are optimal strategies. In individual instances, raters who follow that strategy will have negative transfers, and they may incorrectly attribute such outcomes to their strategy rather than to the vagaries of chance. Few raters will be willing or able to verify the mathematical properties of the scoring system proven in this paper, so it will be necessary to rely on outside attestations to ensure public confidence. Professional experts could be invited to investigate the working of the systems, or independent auditors could be hired.

5. Conclusion

Buyers derive immense value from drawing on the experience of others. However, they have the incentive to shirk from the collective endeavor of providing accurate information about products, be they microwave ovens or movies, academic papers or appliances. Peer-prediction methods, capitalizing on the stochastic relevance between the reports of different raters, in conjunction with appropriate rewards, can create incentives for effort and honest reporting.

Implementors of such systems will face a number of design choices, ranging from rating dimensions and procedures for selecting reviewers, to technology platforms and user interfaces. This paper provides only

[27] This would be analogous to a university provost who normally accepts promotion and tenure recommendations with a minimal review, but may undertake the costly option of personally evaluating the portfolios of candidates from units whose recommendation patterns are suspicious, or employing an outside expert to evaluate those portfolios.

[28] Prendergast's (1993) model of yes-men is one example of this type of situation. In that model, the first rater receives one signal about the expected value of a business action and another signal about how well the next rater (the boss) will like that action. There is no scoring function that will elicit reports from which the center can infer just the rater's direct signal as opposed to her signal about the boss' signal. Thus, she will become, at least partially, a yes-man who says what she thinks the boss will think.

a conceptual road map, not a detailed implementation plan, and only for those design decisions that involve incentives for effort and honest reporting. It is an important road map, however, because the most obvious approach to peer comparison, simply rewarding for agreement in reviews, offers inappropriate incentives.

The basic insight is to compare implied posteriors (rather than an actual report) to the report of a reference rater. A rater need not compute the implications of her own signal for the distribution of the reference rater, as long as she trusts the center to do a good job of computing those implications. There remain many pitfalls, limitations, and practical implementation issues, for which this paper provides conceptual design guidance.

Recommender and reputation systems require that ratings be widely collected and disseminated. To overcome incentive problems, raters must be rewarded. Whether those rewards are monetary or merely grades or points in some scoring system that the raters care about, intense computational methods are required to calibrate appropriate rewards. The upward march of information technology holds promise.

Acknowledgments

The authors thank Alberto Abadie, Chris Avery, Miriam Avins, Chris Dellarocas, Jeff Ely, John Pratt, Bill Sandholm, Lones Smith, Ennio Stachetti, Steve Tadelis, Hal Varian, two referees, and two editors for helpful comments. They gratefully acknowledge financial support from the National Science Foundation under grant numbers IIS-9977999, IIS-0428868, and IIS-0308006, and Zeckhauser thanks the Harvard Business School for hospitality.

Appendix A. Proofs

Proof of Proposition 2. Let

$$Z_i(0) = \arg\max_a \sum_{n=1}^{M} R(s_n^{r(i)} \mid a) f(a),$$

so that the maximum expected value of any report made without acquiring a signal is $\alpha Z_i(0)$. Let

$$\begin{aligned} Z_i(1) &= E_{s_m^i}\left(E_{s_n^{r(i)}} R(s_n^{r(i)} \mid s_m^i)\right) \\ &= \sum_{m=1}^{M} f(s_m^i) \sum_{n=1}^{M} g(s_n^{r(i)} \mid s_m^i) R(s_n^{r(i)} \mid s_m^i), \end{aligned}$$

so that the expected value of getting a signal and reporting it is $\alpha Z_i(1)$. Savage's analysis of the partition problem (1954, Chapter 7) shows that acquiring the signal strictly increases the buyer's expected score whenever it changes the rater's posterior belief about the other raters' announcements (see also Lavalle 1968). Thus, $Z_i(1) > Z_i(0)$ when stochastic relevance holds.

Pick $\alpha > c/(Z_i(1) - Z_i(0))$. Thus $\alpha Z_i(1) - \alpha Z_i(0) > c$, so the best response is to pay the cost c to acquire a signal and report it. □

Proof of Proposition 3. Divide the space of PT beliefs, which are just probabilities that the product is of the good type, into L equal-sized bins, with the l^{th} bin being $B_l = [(l-1)/L, l/L]$ and $B_L = [(L-1)/L, 1]$. Given these bins, the rater's PT belief induces a RRA belief. Let $P_G^l = \int_{(l-1)/L}^{l} f(s \mid G)\, ds$ and $P_B^l = \int_{(l-1)/L}^{l} f(s \mid B)\, ds$, the probabilities assigned to the reference rater announcing the l^{th} bin if the object is known to be good or bad, respectively. If the rater observes s^i, the likelihood of the reference rater's announcing the l^{th} bin is

$$\begin{aligned} P_{s^i}^l &= \int_{(l-1)/L}^{l} p(G \mid s^i) f(s \mid G) + (1 - p(G \mid s^i)) f(s \mid B)\, ds \\ &= p(G \mid s^i) P_G^l + (1 - p(G \mid s^i)) P_B^l. \end{aligned}$$

Let $P_{s^i} = (P_{s^i}^1, \ldots, P_{s^i}^L)$ denote the RRA distribution of a rater who has observed s^i.

Because $p(G \mid s)$ is monotone in s, the inverse function $\pi(p)$ is well defined. Let $\widetilde{B}_l = [\pi((l-1)/L), \pi(l/L)]$ be the l^{th} bin of signals and $\widetilde{B}_L = [\pi((L-1)/L), \pi(1)]$; i.e., raters observing signals in $\widetilde{B}_l$ have PT beliefs in B_l. A rater who announces that her signal is in $\widetilde{B}_l$ is paid using the quadratic scoring rule based on the RRA distribution for a rater who has PT belief $m_l = (2l-1)/2L$. Thus, if a rater always prefers to be scored on the PT bin that contains her true beliefs, she will report the signal bin that contains her true signal. The remainder of the proof is to show that it is optimal for a rater to be scored against the midpoint of the PT bin that contains her true posterior PT belief.

First, we show that closeness of PT beliefs corresponds to closeness of RRA beliefs. The distance between two PT beliefs p_1 and p_2 is simply their absolute difference, $|p_1 - p_2|$. For the distance between two RRA distributions, we use the L_2 metric. That is, if P and $\widehat{P}$ denote two RRA distributions, the L_2 distance between them is given by $d(P, \widehat{P}) = (\sum_l (P^l - \widehat{P}^l)^2)^{1/2}$.

A rater who observes signal s^i assigns probability $P_{s^i}^l = p(G \mid s^i) P_G^l + (1 - p(G \mid s^i)) P_B^l$ to the reference rater announcing bin l. The distance between the posterior distributions of a rater observing s^i and a rater observing $\hat{s}^i$ is therefore given by

$$\begin{aligned} d(P_{s^i}, P_{\hat{s}^i}) &= \left(\sum_l (P_{s^i}^l - P_{\hat{s}^i}^l)^2\right)^{1/2} \\ &= |p(G \mid s^i) - p(G \mid \hat{s}^i)| \left(\sum_l (P_G^l - P_B^l)^2\right)^{1/2}. \quad (5) \end{aligned}$$

Expression (5) establishes that the L_2 distance between two RRA distributions is proportional to the distance between the PT beliefs that generate them.

The final step is to show that, given the choice between being scored based on the RRA distribution for $m_1, \ldots, m_L$, a rater observing s^i maximizes her expected quadratic score by choosing the m_l that is closest to $p(G \mid s^i)$, i.e., her true PT beliefs. This follows from a result from Friedman (1983, Proposition 1), who shows that the expected quadratic score of a rater with true RRA P is larger from reporting $\widehat{P}$ than from reporting $\widetilde{P}$ if and only if $d(\widehat{P}, P) < d(\widetilde{P}, P)$.[29] Thus, Friedman's result, in conjunction with (5), establishes that if a rater believes the reference rater will truthfully announce her bin, then she maximizes her expected quadratic score by selecting the PT bin that contains her true beliefs. □

[29] Friedman (1983) calls metric-scoring-rule pairs that have this property "effective."

Appendix B. Eliciting Effort

To consider the issue of effort elicitation, the rater's experience with the product is encoded not as a single outcome, but as a sequence of outcomes generated by random sampling from distribution $f(s_m \mid t)$. Greater effort corresponds to obtaining a larger sample. Let x_i denote the number of outcomes observed by rater i, i.e., her sample size. We require the rater to put forth effort to learn about her experience, letting $c_i(x_i)$ be the cost of observing a sample of size x_i, where $c_i(x_i)$ is strictly positive, strictly increasing, and strictly convex, and assumed to be known by the center.

For a rater who already observes a sample of size x, learning the $x+1^{\text{st}}$ component further partitions the outcome space, i.e., larger samples correspond to better information. We begin by arguing that holding fixed the agents' sample sizes, scoring-rule-based payments can elicit this information. We then ask how the mechanism can be used to induce agents to acquire more information, even though such acquisition is costly.

For any fixed x_i, the information content of two possible x_i component sequences depends only on the frequencies of the various outcomes and not on the order in which they occur. Consequently, let $Y^i(x_i)$ be the M-dimensional random variable whose m^{th} component counts the number of times outcome s_m occurs in the first x_i components of the agent's information.[30] Let $y^i = (y^i_1, \ldots, y^i_M)$ denote a generic realization of $Y^i(x_i)$, where y^i_m is the number of times out of x_i that signal s_m is received, and note that $\sum_{m=1}^{M} y^i_M = x_i$. Rater i's observation of $Y^i(x_i)$ determines her posterior beliefs about the product's type, which are informative about the expected distribution of the other players' signals. Because different realizations of $Y^i(x_i)$ yield different posterior beliefs about the product's type, it is also natural to assume that $Y^i(x_i)$ is stochastically relevant for $Y^j(x_j)$, and we make this assumption throughout this appendix. In the remainder of this section, we let $g(y^j(x_j) \mid y^i(x_i))$ denote the distribution of $Y^j(x_j)$ conditional on $Y^i(x_i)$.

Lemma 1. *Consider distinct players i and j and suppose x_i, $x_j \geq 0$ are commonly known. If agent i is asked to announce a realization of $Y^i(x_i)$ and is paid according to the realization of $Y^j(x_j)$ using a strictly proper scoring rule, i.e., $R(y^j(x_j) \mid y^i(x_i))$, then the rater's expected payment is uniquely maximized by announcing the true realization of $Y^i(x_i)$.*

Proof. Follows from the definition of a strictly proper scoring rule.

Proposition 4 restates Proposition 1 in the case where the sizes of the raters' samples are fixed and possibly greater than 1, i.e., $x_i \geq 1$ for $i = 1, \ldots, I$. It follows as an immediate consequence of Lemma 1.

Proposition 4. *Suppose rater i collects $x_i \geq 1$ signals. There exist transfers under which truthful reporting is a strict Nash equilibrium of the reporting game.*

Proof of Proposition 4. The construction follows that in Proposition 1, using $Y^i(x_i)$ for the information received by rater i and constructing transfers as in (2) and (4). Under the equilibrium hypothesis, $j = r(i)$ announces truthfully. Let a^i denote rater i's announcement of the realization of $Y^i(x_i)$, and let transfers be given by

$$\tau_i^*(y^j \mid a^i) = R(y^j \mid a^i). \tag{6}$$

Under these transfers, truthful announcement is a strict best response. □

Proposition 4 establishes that truthful reporting remains an equilibrium when raters can choose how much information to acquire. We next turn to the questions of how and whether the center can induce a rater to choose a particular x_i. Let j denote the rater whose signal player i is asked to predict (i.e., let $r(i) = j$), and suppose rater j has a sample of size x_j and that she truthfully reports the realization of $Y^j(x_j)$. (For simplicity, we omit argument x_j in what follows.) Further, suppose that rater i is paid according to the scoring-rule-based scheme described in (6). Because x_i affects these transfers only through rater i's announcement, it is optimal for rater i to truthfully announce $Y^i(x_i)$ regardless of x_i.

Because x_i is chosen before observing any information, rater i's incentive to choose x_i depends on her ex ante expected payoff before learning her own signal. This expectation is written as $Z_i(x_i) = E_{Y^i}(E_{Y^j} R(Y^j \mid Y^i(x_i)))$.

Lemma 2 establishes that raters benefit from better information, and is a restatement of the well-known result in decision theory that every decision maker benefits from a finer partition of the outcome space (Savage 1954).

Lemma 2. *$Z_i(x_i)$ is strictly increasing in x_i.*

Proof of Lemma 2. Fix x_i and let y^i be a generic realization of $Y^i(x_i)$. Conditional on observing y^i, rater i maximizes her expected transfer by announcing distribution $g(Y^j \mid y^i)$ for rater j's information. Suppose rater i observes the $x_i + 1^{\text{st}}$ component of her information. By Lemma 1, i's expected transfer is now strictly maximized by announcing distribution $g(Y^j \mid (y^i, s_m))$, and rater i increases her expected value by observing the additional information. This is true for every y^i, so it is true in expectation, and $Z_i(x_i + 1) > Z_i(x_i)$. □

Lemma 2 establishes that as x_i increases, rater i's information becomes more informative regarding rater j's signal as x_i increases. Of course, the direct effect of rater i's gathering more information is to provide her with better information about the product, not about rater j. Nevertheless, as long as rater i's information is stochastically relevant for that of rater j, better information about the product translates into better information about rater j.

When transfers are given by (6), the expected net benefit to rater i from collecting a sample of size x_i and truthfully reporting her observation is $Z_i(x_i) - c(x_i)$. Hence, transfers (6) induce rater i to collect a sample of size $x_i^* \in \arg\max(Z_i(x_i) - cx_i)$.

Rater i's incentives to truthfully report are unaffected by a uniform scaling of all transfers in (6). Therefore, by a judicious rescaling of the payments to rater i, the center may be able to induce the agent to acquire more or less information. Expression (7) extends the transfers described in (6) to allow for multiple signals and a rescaling of all payments by multiplier $\alpha_i > 0$:

$$\tau_i^*(a^i, y^{r(i)}) = \alpha_i R(y^{r(i)} \mid a^i). \tag{7}$$

Under transfers (7), the maximal expected benefit from a sample of size x_i is $\alpha_i Z_i(x_i)$. Hence, the center can induce rater i to select a particular sample size, $\hat{x}_i$, if and only if there is some multiplier $\hat{\alpha} > 0$ such that

[30] $Y^i(x_i)$ is a multinomial random variable with x_i trials and M possible outcomes. On any trial, the probability of the m^{th} is $f(s_m \mid t)$, where t is the product's unknown type.

$\hat{x}_i \in \arg\max \hat{\alpha} Z_i(x_i) - c(x_i)$. The simplest case has $Z_i(x_i)$ concave, i.e., where $Z_i(x_i+1) - Z_i(x_i)$ decreases in x_i.

PROPOSITION 5. *If $Z_i(x_i+1) - Z_i(x_i)$ decreases in x_i, then for any sample size $\hat{x}_i \geq 0$ there exists a scalar $\hat{\alpha}_i \geq 0$ such that when paid according to* (7)*, rater i chooses sample size $\hat{x}_i$.*

PROOF OF PROPOSITION 5. $Z_i(x)$ is concave, so sample size $\hat{x}_i$ is optimal if there exists $\hat{\alpha}_i$ satisfying

$$\hat{\alpha}_i Z_i(\hat{x}_i) - c_i(\hat{x}_i) \geq \hat{\alpha}_i Z_i(\hat{x}_i+1) - c_i(\hat{x}_i+1), \quad \text{and}$$
$$\hat{\alpha}_i Z_i(\hat{x}_i) - c_i(\hat{x}_i) \geq \hat{\alpha}_i Z_i(\hat{x}_i-1) - c_i(\hat{x}_i-1).$$

Solving each condition for $\hat{\alpha}_i$ yields

$$\frac{c_i(\hat{x}_i) - c_i(\hat{x}_i-1)}{Z_i(\hat{x}_i) - Z_i(\hat{x}_i-1)} \leq \hat{\alpha}_i \leq \frac{c_i(\hat{x}_i+1) - c_i(\hat{x}_i)}{Z_i(\hat{x}_i+1) - Z_i(\hat{x}_i)}.$$

Such an $\hat{\alpha}_i$ exists if and only if

$$\frac{Z_i(\hat{x}_i) - Z_i(\hat{x}_i-1)}{Z_i(\hat{x}_i+1) - Z_i(\hat{x}_i)} \geq \frac{c_i(\hat{x}_i) - c_i(\hat{x}_i-1)}{c_i(\hat{x}_i+1) - c_i(\hat{x}_i)}.$$

By our assumptions, this expression is always true. □

If $Z_i(x_i+1) - Z_i(x_i)$ does not decrease in x_i, then there may be some sample sizes that are never optimal.[31] Nevertheless, increasing the scaling factor never decreases optimal sample size, and so while the center may not be able to perfectly control the raters' effort choices, it can always induce them to put forth greater effort if it wishes.

In practice, the center will not know each individual's cost of procuring additional information. However, the center may be able to estimate costs, and then pick a scaling factor that, in expectation, induces each rater to acquire an optimal size sample.[32]

References

Avery, C., P. Resnick, R. Zeckhauser. 1999. The market for evaluations. *Amer. Econom. Rev.* **89**(3) 564–584.

Breese, J., D. Heckerman, C. Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. *Proc. Fourteenth Conf. Uncertainty Artificial Intelligence*. Morgan Kaufmann Publisher, Madison, WI.

Clemen, R. 2002. Incentive contracts and strictly proper scoring rules. *Test* **11**(1) 195–217.

Congdon, P. 2001. *Bayesian Statistical Modelling.* Wiley, Chichester, UK.

Cooke, R. M. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science.* Oxford University Press, New York.

Crémer, J., R. McLean. 1985. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica* **53**(2) 345–361.

Crémer, J., R. McLean. 1988. Full extraction of surplus in Bayesian and dominant strategy auctions. *Econometrica* **56**(6) 1247–1257.

d'Aspremont, C., L.-A. Gérard-Varet. 1979. Incentives and incomplete information. *J. Public Econom.* **11**(1) 25–45.

[31] Clemen (2002) provides a number of examples of cases in which $Z_i(x_i+1) - Z_i(x_i)$ decreases in x_i.

[32] The center chooses the scale that induces the optimal ex ante precision. Ex post, if raters know their costs, they will tend to choose lower precision if they are high cost and vice versa.

d'Aspremont, C., L.-A. Gérard-Varet. 1982. Bayesian incentive compatible beliefs. *J. Math. Econom.* **10**(1) 83–103.

Dellarocas, C. 2001. Analyzing the economic efficiency of eBay-like online reputation reporting mechanisms. *Proc. 3rd ACM Conf. Electronic Commerce*, ACM Press, New York.

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

Gollier, Cristian. 2001. *The Economics of Risk and Time.* MIT Press, Cambridge, MA.

Hanson, R. 2002. Logarithmic market scoring rules for modular combinatoral information aggregation. Working paper, Department of Economics, George Mason University, Fairfax, VA.

Jensen, F., C. Peterson. 1973. Psychological effects of proper scoring rules. *Organ. Behavior Human Performance* **9**(2) 307–317.

Johnson, S., J. Pratt, R. Zeckhauser. 1990. Efficiency despite mutually payoff-relevant private information: The finite case. *Econometrica* **58**(4) 873–900.

Johnson, S., N. Miller, J. Pratt, R. Zeckhauser. 2003. Efficient design with interdependent valuations and an informed center. Working paper RWP03-020, Kennedy School of Government, Harvard University, Cambridge, MA.

Kandori, M., H. Matsushima. 1998. Private observation, communication and collusion. *Econometrica* **66**(3) 627–652.

Lampe, C., P. Resnick. 2004. Slash(dot) and burn: Distributed moderation in a large online conversation space. *CHI Lett.* **6**(1) 543–550.

Lavalle, I. 1968. On cash equivalents and information evaluation in decisions under uncertainty: Part I: Basic theory. *J. Amer. Statist. Assoc.* **63**(321) 252–276.

Miller, N., P. Resnick, R. Zeckhauser. 2005. A sufficient condition for correlated information in mechanism design. Mimeo, Harvard University, Cambridge, MA.

Nau, R. 1985. Should scoring rules be effective? *Management Sci.* **34**(5) 527–535.

Nelson, R., D. Bessler. 1989. Subjective probabilities and scoring rules: Experimental evidence. *Amer. J. Agricultural Econom.* **71**(2) 363–369.

Ottaviani, M., P. N. Sørensen. 2004. The strategy of professional forecasting. Finance Research Unit working paper 2004–2005, University of Copenhagen, Copenhagen, Denmark.

Pratt, J., H. Raiffa, R. Schlaifer. 1965. *Introduction to Statistical Decision Theory.* McGraw-Hill, New York.

Prelec, D. 2001. A two-person scoring rule for subjective reports. Working paper, Marketing Center, MIT Sloan School, Cambridge, MA.

Prendergast, C. 1993. A theory of yes-men. *Amer. Econom. Rev.* **83**(4) 757–770.

Roth, A. 1995. Introduction to experimental economics. J. Kagel, A. Roth, eds. *The Handbook of Experimental Economics,* Princeton University Press, Princeton, NJ, 3–110.

Sarwar, B. M., G. Karypis, J. A. Konstan, J. Riedl. 2000. Analysis of recommender algorithms for e-commerce. *Proc. 2nd ACM Conf. Electronic Commerce*, ACM Press, New York, 158–167.

Savage, L. 1954. *Foundations of Statistics.* Dover Publications, New York.

Savage, L. 1971. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66**(336) 783–801.

Selten, R. 1998. Axiomatic characterization of the quadratic scoring rule. *Experiment. Econom.* **1**(1) 43–62.

Smith, C. 1961. Consistency in statistical inference and decision. *J. Roy. Statist. Soc., Series B (Methodological)* **23**(1) 1–37.

Staël von Holstein, C.-A. 1970. Measurement of subjective probability. *Acta Psych.* **34**(1) 146–159.

Winkler, R. 1969. Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* **64**(327) 1073–1078.

Winkler, R. 1996. Scoring rules and the evaluation of probabilities. *Test* **5**(1) 1–60.