Eliciting Informative Feedback: The Peer-Prediction Method

Nolan Miller, Paul Resnick, and Richard Zeckhauser*

June 15, 2004†

Abstract

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 belief 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 on-line recommender and reputation systems.

*Miller and Zeckhauser, Kennedy School of Government, Harvard University; Resnick, School of Information, University of Michigan.

†We 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. We gratefully acknowledge financial support from the National Science Foundation under grant numbers IIS-9977999 and IIS-0308006, and the Harvard Business School for hospitality (Zeckhauser).

1 Introduction

We frequently draw on the experiences of multiple other individuals when making decisions. The process can be informal. Thus, an executive deciding whether to invest in a new business opportunity will typically consult underlings, each of whom has some specialized knowledge or perspective, before committing to a decision. Similarly, journal editors secure independent reviews of papers, and grant review panels and admissions and hiring committees often seek independent evaluations from individual members who provide input to a group decision process.

In other contexts, the process of eliciting and conveying private information can be institutionalized. For example, recommender systems gather and aggregate individual ratings and reviews in order to inform others' choices. On the Internet, eBay invites buyers and sellers to rate each other; MovieLens, Amazon, and ePinions invite ratings of movies, books, and other products on a 1-5 scale; and Zagat Survey solicits restaurant ratings on a 1-30 scale on the separate dimensions of food, decor, and service. The role of institutionalized feedback methods has been greatly enhanced by the Internet, which can gather and disseminate information from vast numbers of individuals at minimal cost.

However formal or informal, 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, not the provider.[1] The second challenge is honesty. Raters may prefer to be nice and withhold negative evaluations.[2] They may fear retaliation or have conflicts of interest. If they care about how they will be perceived, raters may be tempted to report opinions that will improve others' perceptions of them, rather than reporting their honest opinions.

Building an explicit reward system for honest rating and effort is the first step in overcoming these challenges. When objective information will be publicly revealed at a future time, individuals' reports can be compared to that objective information. For example, old evaluations of stocks can be compared to their subsequent price movements, and weather forecasts can be compared against what actually occurs.

This analysis develops methods to elicit feedback effectively when independent, objective outcomes are not available. It may be that no objective outcome exists, as in evaluations of a product's

[1] Despite the rational incentive to free-ride, provision of feedback at eBay is quite common, occurring in more than 50% of transactions for a large sample from 1999 (Resnick and Zeckhauser, 2002).

[2] Dellarocas (2001) analyzes a model in which, so long as harshness in interpretation of feedback is appropriately tied with leniency in giving feedback, leniency offers 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.

"quality." Or, it may be that even though there exists objective information relevant to the issue, such information is not public and verifiable. For example, while breakdown frequency may be an objective measure of product quality, this information is only available to others if it is revealed by the product's current owners. Or, the outcome may be publicly observable, but not independent of the raters' reports. For example, an admissions committee member who can effectively veto the admission of a particular student can always assure agreement between her rating and the outcome by submitting a negative rating. Finally, even when independent, objective outcomes do occur, they may occur too far in the future to provide effective incentives to current raters.

In these situations, one solution is to compare raters' reports to their peers' reports.[3] Such comparison processes occur naturally and informally, as people check whether individual reviewers' opinions differ considerably from those of the rest of the pool. If payoffs are made part of the process, however, dangers arise. 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.[4] 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 turned into 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

[3] 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 meta-moderators to rate the ratings of comments given by regular moderators. Meta-evaluation incurs an obvious inefficiency, since 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, meta-evaluation 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.

[4] Yes Men, who say what they think the boss will say present a related hazard: the additional information they have about the boss's likely perception makes it impossible to fully extract a Yes Man's private information from the report he gives (Prendergast; 1993).

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.[5]

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 when types are correlated. Adapting tools from statistical decision theory, Johnson, Pratt, and Zeckhauser (1990) show how to construct budget-balancing transfer payments based on "proper scoring rules." Johnson, Miller, Pratt, and Zeckhauser (2002) extend those results to the case of multidimensional, continuous private information. Kandori and Matsushima (1998, section 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 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 particularly well suited to Internet-based implementations and could potentially be applied to services such as MovieLens or Amazon.[6] Once ratings are collected and distributed electronically, it is relatively easy to compute posteriors and scores and keep track of payments.

In Section 2 we show that the reviewing application quite naturally fits an informational requirement, which we call stochastic relevance, that is sufficient to allow the center to elicit the rater's private information using a proper scoring rule. We extend the scoring-rule-based approach to address core theoretical issues that arise in this applied problem, such as the elicitation of effort, sequential reporting, and discrete reporting based on continuous signals. In Section 3 we address a number of 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

[5] Similarly, offline point systems that do not provide any tangible reward seem to motivate chess and bridge players to compete harder and more frequently.

[6] 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 varied 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.

about the distributions of product quality and rater types. We also discuss limitations of the mechanism. Section 4 concludes.

2 A Mechanism for Eliciting Honest Feedback

A number of raters experience a product and then rate it for quality. The quality of a product does not vary, 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, with the amount of the transfers determined by the ratings. The center has no independent information, so its scoring decisions can depend only on the information provided by other raters.[7] 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.[8] 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 $t = 1, ..., T$. Let $p(t)$ be the commonly held prior probability assigned to the product's being type t.[9] 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, judging from her own experience, privately observes a signal of the product's type.[10] 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 $S = \{s_1, ..., s_M\}$ be the set of possible signals, and let $f(s_m|t) = \Pr\left(S^i = s_m|t\right)$, where $f(s_m|t) > 0$ for all s_m and t, and $\sum_{m=1}^{M} f(s_m|t) = 1$ for all t. We assume that $f(s_m|t)$ 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 $p(H) = .5$, and two possible signals, h and l, with $f(h|H) = .85$ and $f(h|L) = .45$. Thus, $\Pr(h) = .5 * .85 + .5 * .45 = .65$.

[7] Given independent verifying power, a variant of the system outlined below would be easier to implement. It could simply pay raters by how well they predicted the center's information. Utilizing information from other raters as well as the center would increase the reliability of the mechanism, but would not affect the incentive to report honestly.

[8] We consider the impacts of risk aversion in section 3.1.

[9] We briefly address the issue of non-common priors later.

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

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. Let $a^i \in S$ denote one such announcement, and $a = \left(a^1, ..., a^I\right)$ 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 $\bar{a}^i = \left(a^i_1, ..., a^i_M\right) \in S^M$ denote rater i's announcement strategy. Let $\bar{a} = \left(\bar{a}^1, ..., \bar{a}^I\right)$ denote a vector of announcement strategies. As is customary, let the superscript "$-i$" denote a vector without rater i's component.

Let $\tau_i(a)$ denote the transfer paid to rater i when the raters make announcements a, and let $\tau(a) = (\tau_1(a), ..., \tau_I(a))$ 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:

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

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 $i = 1, ..., I$, and a strict Nash Equilibrium if the inequality in (1) is strict for all $i = 1$,..., 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. Since 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, since players are indifferent between 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. The analysis begins by noting (Lemma 1) that although S^i and S^j are conditionally independent (conditional on the product's type), they are not necessarily independent. Because each signal is drawn from the same distribution with unknown parameter t, S^i and S^j are generally dependent. In fact,

our results rely on a form of dependence, which we call stochastic relevance.[11]

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.

More technically, 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 $\Pr\left(s^j|s^i\right) \neq \Pr\left(s^j|\hat{s}^i\right)$.[12] Let $g\left(S^j|S^i\right)$ be the distribution of S^j conditional on S^i, and let $g\left(s^j|s^i\right)$ represent $\Pr\left(S^j = s^j|S^i = s^i\right)$.

Lemma 1: *For generic distributions $f(s_m|t)$ and $p(t)$, S^i is stochastically relevant for S^j for any two distinct players i and j.*[13]

In the case of product reviews, we will consider two products to be of different types when they produce different signal distributions. Lemma 1 establishes that in that case, it is only "by coincidence" that two distinct signals, s^i and $\hat{s}^i$, yield the same posterior beliefs about the distribution of another rater's signal. Such coincidences occur with probability zero and can be safely ignored.[14]

Since generically S^i is stochastically relevant for S^j, in the remainder of the paper we will assume that this conditions holds.

Continuing the two-type, two-signal example, suppose that rater i receives the signal l. Recall that $p(H) = .5$, $f(h|H) = .85$, and $f(h|L) = .45$, so that $\Pr\left(s_l^i\right) = .35$. Given i's signal, the probability that rater j will receive a signal h is:

$$g\left(s_h^j|s_l^i\right) = f(h|H)\frac{f(l|H)\,p(H)}{\Pr\left(s_l^i\right)} + f(h|L)\frac{f(l|L)\,p(L)}{\Pr\left(s_l^i\right)} = .85\frac{.15 * .5}{.35} + .45\frac{.55 * .5}{.35} \cong 0.54.$$

If i had instead observed h, then:

$$g\left(s_h^j|s_h^i\right) = f(h|H)\frac{f(h|H)\,p(H)}{\Pr\left(s_h^i\right)} + f(h|L)\frac{f(h|L)\,p(L)}{\Pr\left(s_h^i\right)} = .85\frac{.85 * .5}{.65} + .45\frac{.45 * .5}{.65} \cong 0.51.$$

The elicitation of beliefs about the distribution of S^j from an agent who has observed S^i is the

[11] The concept of stochastic relevance is introduced in Johnson, Miller, Pratt, and Zeckhauser (2002).

[12] This condition is the same as the condition (A4) used in Kandori and Matsushima (1998).

[13] That is, the closure of the set of distributions for which S^i is not stochastically relevant for S^j has Lebesgue measure zero. See Mas-Colell, Whinston, and Green (1995, p. 595) for a discussion of generic conditions. Proofs omitted from the main text are contained in Appendix A.

[14] In the event that $g\left(s^j|s^i\right) \equiv g\left(s^j|\hat{s}^i\right)$ for all s^j given the current prior distribution, for any small perturbation of the prior this condition will no longer hold.

problem for which proper scoring rules were developed.[15] Put simply, suppose agent i privately observes the realization of S^i, which is stochastically relevant for some publicly observable random variable S^j, and agent i is asked to reveal her private information. A scoring rule is a function $R\left(s^j|a^i\right)$ that, for each possible announcement a^i of S^i, assigns a score to each possible realization of S^j. A convenient interpretation is that the scoring rule specifies the payment made (or penalty assessed) to the agent following each realization S^j. A scoring rule $R\left(s^j, a^i\right)$ is strictly proper if rater i uniquely maximizes her expected score by announcing the true realization of S^i.

The literature contains a number of strictly proper scoring rules.[16] The three best known are:

1. Quadratic Scoring Rule: $R\left(s_n^j|a^i\right) = 2g\left(s_n^j|a^i\right) - \sum_{h=1}^{M} g\left(s_h^j|a^i\right)^2$.
2. Spherical Scoring Rule: $R\left(s_n^j|a^i\right) = \frac{g\left(s_n^i|a^i\right)}{\left(\sum_{h=1}^{M} g\left(s_h^j|a^i\right)^2\right)^{\frac{1}{2}}}$.
3. Logarithmic Scoring Rule: $R\left(s_n^j|a^i\right) = \ln g\left(s_n^j|a^i\right)$.

Further, if $R(\cdot|\cdot)$ is a strictly proper scoring rule, then a positive affine transformation of it, i.e., $\alpha R(\cdot|\cdot)+\beta$, is also a strictly proper scoring rule. The ability of the center to manipulate constants α and β is useful in inducing the raters to exert effort and ensuring that their participation constraints are satisfied. Throughout the paper, we will use $R\left(s_n^j|a^i\right)$ 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 continue to hold for any strictly proper scoring rule.

Transfers based on a strictly proper scoring rule can be used to induce truthful revelation by agent i as long as her private information is stochastically relevant for some other publicly available signal. However, in the case we consider each rater's signal is private information, and therefore we can only check players' announcements against other players' announcements, not their actual signals. Nevertheless, if a rater believes that other raters will announce their information truthfully, then transfers based on a strictly proper scoring rule induce the rater to truthfully announce her own information. That is, truthful reporting is a strict Nash Equilibrium.

Proposition 1: *There exist transfers under which truthful reporting is a strict Nash Equilibrium of the reporting game.*

[15] See Cooke (1991) for an introduction to the use of scoring rules.

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

Proof of Proposition 1: For each rater, choose another rater $r(i)$, the reference rater for i, whose announcement i will be asked to predict. Let

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

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

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

Since $R(\cdot|\cdot)$ is a strictly proper scoring rule, (3) is uniquely maximized by announcing $a^i = s^*$, i.e., truthful announcement is a strict best response. Thus, given that player $r(i)$ announces truthfully, player i's best response is to announce truthfully as well.■

We illustrate Proposition 1 using the logarithmic scoring rule. Since $0 < g\left(s_m^j | s_n^i\right) < 1$, $\ln g\left(s_m^j | s_n^i\right) < 0$, and so we refer to τ_i^* as rater i's penalty since it is always negative in this case. Consider the simple example where rater i received the signal l. That signal was unlikely ($\Pr\left(s_l^i\right) = .35$). Moreover, even contingent on that signal it was unlikely that rater j would receive an l signal ($g\left(s_l^j | s_l^i\right) = 1 - 0.54 = .46$). Thus, if rater i were rewarded merely for matching her report to the next report, she would prefer to report h. With the log scoring rule, an honest report of l leads to an expected payoff

$$\ln g\left(s_h^j | l\right) g\left(s_h^j | l\right) + \ln g\left(s_l^j | l\right) g\left(s_l^j | l\right) = \ln(.54)\,.54 + \ln(.46).46 = -0.69.$$

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

$$\ln g\left(s_h^j | h\right) g\left(s_h^j | l\right) + \ln g\left(s_l^j | h\right) g\left(s_l^j | l\right) = \ln(.71)\,.54 + \ln(.29).46 = -0.75.$$

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 next signal, not simply matching a rater's report to the next one. The updating takes into account both the priors and the reported signal, and thus reflects the initial rater's priors. Thus, she has no 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.[17]

2.2 Eliciting Effort

The assumption that evaluation and reporting are costless allowed us to focus on the essence of the scoring-rule based mechanism. However, raters incur costs, including direct costs of effort as well as the opportunity cost of being an early evaluator rather than waiting for better information from other evaluators before deciding whether to use a product. If the expected payoff is less than the sum of these costs, raters will skip the task or provide feedback without doing a good job.[18] 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 of evaluating and reporting given by $c > 0$. To induce effort, the expected value of incurring effort and reporting honestly must exceed the expected value of reporting without receiving any signal. As the proof of Proposition 1 makes clear, the truth-inducing incentives provided by scoring-rule based payments (or any of the scoring rules mentioned above) are unaffected by a positive rescaling of all transfers; if transfers $\tau_i^*\left(a^i, a^{r(i)}\right) = R\left(a^{r(i)}|a^i\right)$ induce truthful reporting, then $\tau_i^*\left(a^i, a^{r(i)}\right) = \alpha R\left(a^{r(i)}|a^i\right)$, where $\alpha > 0$, does as well. Thus, even strictly proper scoring rules offer significant leeway to adapt the transfers. 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 $\alpha \geq 0$ such that when rater i* is *paid according to $\tau_i^*\left(a^i, a^{r(i)}\right) = \alpha R\left(a^{r(i)}|a^i\right)$, her best response is to acquire a signal and report it honestly.*

[17] 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.

[18] At the news and commentary site Slashdot, 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."

The same scaling ideas can be generalized to a situation where raters can choose to work harder to obtain better information.[19] 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.[20] Assuming that the cost of acquiring a sample is increasing and convex in the sample's size, we can ask when and how it is possible for the center to induce the raters to acquire samples of a particular size.

We relegate the technical presentation to Appendix B. However, the basic idea is straightforward. For any sample size, stochastic relevance continues to hold for generic distributions. 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 desired sample size.[21]

We return to the question of eliciting effort in Section 2.5.1, where, due to assuming information is normally distributed, we are able to present the theory more parsimoniously.

2.3 Voluntary Participation and Budget Balance

The transfers constructed in the previous sections induce raters to report truthfully and exert costly effort. However, 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, since the logarithmic score is always negative. However, this problem is easily addressed. Since adding a constant to all payments (i.e., letting the transfer be given by $\alpha_i R\left(a^{r(i)}|a^i\right) + k_i$) will not affect on the incentives for effort or honest

[19] Clemen (2002) undertakes a similar investigation in a principal-agent model.

[20] Later, we discuss the case where raters' beliefs and signals are normally distributed, in which case it is natural to think of better information in terms of reducing the signal variance.

[21] Clemen (2002) provides examples of a number of distributions for which $V^*(x)$ is concave.

reporting, the constant k_i can be chosen to satisfy either ex ante participation constraints (i.e., each agent must earn a non-negative expected return), interim participation constraints (i.e., each agent must earn a positive return conditional on any observed signal), or ex post participation constraints (i.e., the agent must earn a positive expected return for each possible $\left(s^j, s^i\right)$ pair). To illustrate using the logarithmic case, since there are a finite number of possible reports, there exists some maximum penalty. Let $\tau_0 = \min_{s_m, s_n \in S} (\alpha \ln g(s_m|s_n))$. Define $\tau^+ = \tau^* - \tau_0$. If α is chosen to induce a desired effort level, transfers τ^+ will attract voluntary (ex post) participation while still inducing effort and honest reporting.

It is often desirable for the center to balance its budget. Clearly, this is important if scores are converted into monetary payments. Even if scores are merely karma points or some other currency that the center can generate at will, uncontrolled inflation would make it hard for users to interpret point totals. As long as there are at least three raters, the center can balance the budget by reducing the base transfer τ^* to each rater by an amount equal to some other rater's base transfer. Though all the transactions actually occur between raters and the center, this creates the effect of having the raters settle the transfers among each other.[22] 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 $r(b(i)) \neq i$. The net transfer to rater i is then:

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

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. Since 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. Each player's expected gain from honest reporting is zero so long as the expected value of τ_i^* is the same for all players. If, however, a player knows that her own signals are less precise than those of other players, then her expected base transfer τ_i^* from the proper scoring rule will be less than the expected $\tau_{b(i)}^*$. Even if the expected gain is zero or positive, in particular cases it may be negative. One way to assure ex-post voluntary participation is to collect bonds or entry fees in advance, and use the collected

[22] Since 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^+.

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 the bond prevents 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. Of course, when payments are made in points maintained by the system, as with karma points or chess ratings, the system need not secure raters' acquiescence when changing their point totals, and ex-post voluntary participation is not a constraint.

2.4 Sequential Interaction

In the scenario above, raters report their experiences simultaneously. Sequential reporting may be desirable, since it makes superior use of information by allowing later raters to make immediate use of the information provided by their predecessors. The mechanism adapts readily to sequential situations.[23] The transfer to rater i can be determined using any subsequent rater as a reference rater. To balance the budget, the transfer can be funded by any subsequent rater other than rater i's reference rater.

More formally, suppose an infinite sequence of raters, indexed by $i = 1, 2, ...$, interacts with the product in order. We will designate rater $i+1$ as rater i's reference rater, i.e., i's report is used to predict the distribution of rater $i+1$'s report. Initially, the commonly held prior distribution for the product's type is given by $p(t)$. Let $p_1\left(t|s^1\right)$ denote the posterior distribution after rater 1 receives signal s^1. That is,

$$p_1\left(t|s^1\right) = \frac{f\left(s^1|t\right)p(t)}{\Pr\left(s^1\right)}, \tag{5}$$

where $\Pr\left(s^1\right) = \sum_{t=1}^{T} f\left(s^1|t\right)p(t)$. Rater 1's posterior belief about the probability that $S^2 = s^2$ is given by

$$g\left(s^2|s^1\right) = \sum_{t=1}^{T} f\left(s^2|t\right)p_1\left(t|s^1\right). \tag{6}$$

Given the distribution specified in (6), rater 1 can be induced to truthfully reveal s^1 using the scoring rule specified in Proposition 1. Rater 1 announces her signal. The center computes a

[23] 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.

posterior distribution of rater 2's announcements. The penalty is determined by rater 2's actual report.

By iteratively updating beliefs about the product using Bayes' rule (as in (5)), each player's announcement feeds into the information used to score subsequent players. Let $p_{i-1}(t)$ be the prior distribution over types computed from the announcements of the first $i-1$ raters. Although $p_{i-1}(t)$ depends on the history of announcements $s^1, ..., s^{i-1}$, we suppress this dependence for notational simplicity. Conditional on observing signal s^i, rater i's posterior beliefs about the distribution of types is given by:

$$p_i\left(t|s^i, p_{i-1}\right) = \frac{f\left(s^i|t\right) p_{i-1}\left(t|s^1, ..., s^{i-1}\right)}{\Pr\left(s^i|p_{i-1}\right)}, \tag{7}$$

where $\Pr\left(s^i|p_{i-1}\right) = \sum_{t=1}^{T} f\left(s^i|t\right) p_{i-1}(t)$. Rater i's computed posterior beliefs about the distribution of rater $i+1$'s announcement is given by:

$$g\left(s^{i+1}|s^i\right) = \sum_{t=1}^{T} f\left(s^{i+1}|t\right) p_i\left(t|s^i, p_{i-1}\right). \tag{8}$$

The sequential elicitation game has each rater i observe s^i, report it, and then be scored based on the implied distribution of the subsequent rater's (i.e., rater $i+1$'s) signal. We continue to assume that, conditional on product type, signals are independent, so that prior raters' announcements affect how current players are scored, but not the signals they receive. Transfers constructed according to (2) using the conditional distribution specified in (8) elicit truthful announcement. This announcement then becomes common knowledge and is used to update beliefs about the product according to (7). Incentives to rater $i+1$ are then constructed using a scoring rule that incorporates these updated beliefs.

To balance the budget, let rater i's transfer be paid by rater $i+2$. For all raters after the first two, the net transfer will be $\tau_i^* - \tau_{i-2}^*$. Raters 1 and 2 present the only complication since they do not pay transfers to anyone. For example, under the logarithmic rule these raters pay a penalty (since the logarithmic score is negative) but do not receive one from another player. To ensure voluntary participation, the center can provide an additional payment of τ_0 to them, as in the previous section. The budget remains nearly though not exactly in balance. After transfers are paid to the ith rater ($i \geq 2$), the budget will be out of balance by the amount $\tau_i^* - 2q$. Presumably, if necessary, the center could recover the initial expenditure of $-2q$ by charging a small fixed fee to later raters.

When a finite string of raters experience the product, the process of checking each rater's announcement against that of a later rater could unravel. The last rater has no incentive to lie, but also none to tell the truth, since there is no future signal upon which to base her reward. If the final announcement is unreliable, the previous raters cannot be induced to report truthfully, and so on up the line. Worse, if reporting is costly, the final rater may not submit a report at all, again leading the system to collapse.

Fortunately, the center can group some raters together and treat the groups as if they report simultaneously, creating reporting "rings" that provide appropriate incentives for every rater. 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. The transfers can also be made within the ring in order to balance the budget for the ring. To balance the overall budget exactly, rather than approximately as in the infinite case, multiple rings could be created, e.g., $\{1, 2, 3, 4\}$, $\{5, 6, 7\}$, and $\{8, 9, 10\}$. Within each ring, reports would be revealed simultaneously, but the reports of each ring would be available to the next ring, maintaining some of the advantages of sequential reporting.

There are other ways to avoid unraveling and to balance the budget exactly. For example, the center could withhold the reports of the first rater and the third-to-last rater until after the final rater reports. The last two raters are scored against the report of rater 1, and otherwise rater i is scored against rater $i + 2$. Rater i's transfer is paid by rater $i + 1$, except for the last rater, whose transfer is paid by rater 1. Since all penalties are assessed against reports that have not yet been revealed, the transfers induce effort and honest reporting. Since all transfers are paid by a rater who has not yet reported at the time the transfer is computed, the transfers do not affect incentives for effort or honest reporting.

2.5 Continuous Signals

Until now, we have considered a model where type and signal spaces are discrete. 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 $\ln\left(g\left(s^j|s^i\right)\right)$, where $g\left(s^j|s^i\right)$ is now the posterior *density* of s^j given 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.

2.5.1 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 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, the rater's posterior belief about q is that q is distributed normally with mean $\hat{\mu}$ and precision $\theta_q + \theta_i$, where:[24]

$$\hat{\mu} = \frac{\left(\mu\theta_q + s^i\theta_i\right)}{\left(\theta_q + \theta_i\right)}. \tag{9}$$

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, rater i's posterior belief about the distribution of s^j is that s^j is normally distributed with mean $\hat{\mu}$ and precision θ, where $\theta = \theta_q + \theta_i + \theta_j$.

Since different observation-precision combinations lead to different posterior beliefs about the distribution of s^j, our stochastic relevance condition is satisfied, and 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, and this is achieved by honestly reporting her observation and her precision, allowing the center to correctly compute her posterior.[25]

We assume that by exerting effort, raters can increase the precision of their signals. Let $c\left(\theta_i\right)$ represent the cost of acquiring a signal of precision $\theta_i \geq 0$, where $c'\left(\theta_i\right) > 0$, $c'\left(0\right) = 0, c'\left(\infty\right) = \infty$, and $c''\left(\theta_i\right) \geq 0$. To compare the logarithmic, quadratic, and spherical scoring rules, it is necessary

[24] See Pratt, Raiffa, and Schlaifer (1965).

[25] Ottavianni and Sorensen (2003) 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.

to ensure that the rater is choosing the same signal precision under each rule. As suggested by our analysis in Section 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. Under the logarithmic scoring rule, the maximized expected utility as a function of precision (i.e., when the rater announces truthfully) is given by:

$$\begin{aligned} v_l(\theta_i) &= \int \log(f(x)) f(x)\, dx = \int \left\{ -\frac{\theta(x-\mu)^2}{2} - \log\left(\sqrt{\frac{2\pi}{\theta}}\right) \right\} f(x)\, dx \\ &= \frac{-\theta E(x-\mu)^2}{2} - \log\left(\sqrt{\frac{2\pi}{\theta}}\right) = -\frac{1}{2} + \frac{1}{2}\log\left(\frac{\theta}{2\pi}\right). \end{aligned}$$

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 $\alpha = 2(\theta_q + \theta_i + \theta_j)\, c'(\theta_i) \equiv \alpha_l$ induces precision θ_i under the logarithmic rule.

Under the quadratic rule, the rater's expected score if he announces truthfully is:

$$v_q(\theta_i) = \int 2f(x)^2\, dx - \int f(x)^2\, dx = \frac{\sqrt{\theta}}{2\sqrt{\pi}} = \frac{\sqrt{\theta}}{2\sqrt{\pi}},$$

since $\int f(x)^2\, dx = \frac{\sqrt{\theta}}{2\sqrt{\pi}}$. Again, the maximized expected score, $v_q(\theta_i)$, is increasing and concave in θ_i. To find the optimal scaling factor, solve:

$$\max_{\theta_i} \alpha\left(\frac{\sqrt{\theta}}{2\sqrt{\pi}}\right) - c(\theta_i),$$

which yields that choosing $\alpha = 4\sqrt{\pi}(\theta_q + \theta_i + \theta_j)^{\frac{1}{2}} c'(\theta_i) \equiv \alpha_q$ induces precision θ_i under the quadratic rule.

Repeating the same computation for the spherical rule, the expected score is:

$$v_s(\theta_i) = \left(\frac{\sqrt{\theta}}{2\sqrt{\pi}}\right)^{\frac{1}{2}},$$

which is increasing and concave in θ_i. Solving

$$\max_{\theta_i} \alpha \left(\frac{\sqrt{\theta}}{2\sqrt{\pi}}\right)^{\frac{1}{2}} - c(\theta_i)$$

yields that $\alpha = 4\sqrt{2}\pi^{1/4}(\theta_q + \theta_i + \theta_j)^{3/4} c'(\theta_i) \equiv \alpha_s$ induces precision θ_i under the spherical rule.

Thus, when information is normally distributed the center can induce a desired level of precision and truthful revelation using any of these rules. And, since a constant can be added to any of the scoring rules without affecting either truth-telling or effort-inducing incentives (though possibly participation incentives), the expected payment under the different rules cannot be used to discriminate between them. One salient dimension over which the rules differ is the variability of the payments needed to induce a particular effort level. Variability may become important if raters have limited liability or are risk averse. The variance and range of the transfers needed to induce a particular precision (effort) level under each of the rules is:[26]

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

Several important features emerge from this analysis. First, the quadratic and spherical rules have the same variance and range of payments. This is because when information is normally distributed both rules specify scores that are linear in $f(x)$. Thus, when the two rules are scaled to induce the same precision, they differ only by an additive constant. Second, the logarithmic rule has the smallest variance ($\frac{16}{3}\left(2\sqrt{3}-3\right) \simeq 2.4752$). If scores are used to evaluate the raters (for example, to decide whether to invite them back as reviewers in the future), a lower variance will allow more reliable evaluation based on fewer trials, and thus the logarithmic rule may be preferred. On the other hand, the range of payments is infinite with the logarithmic scoring

[26] Supporting computations for this table are available from the authors upon request.

rule because $\lim_{x\to 0} \ln(x) = -\infty$. Thus, the logarithmic rule may not be an attractive option when probabilities become small and raters' limited liability is a concern. However, while small-probability events play a role when information is normally distributed, they need not always be important, and the logarithmic rule may be preferable to the other rules in particular cases. We return to the relative merits of the rules in Section 3.2.

2.5.2 Eliciting Coarse Reports

Raters' information is often highly complex. For example, it may take pages of description to completely summarize all of the relevant information about a person's experience in a restaurant. Although a proper scoring rule could elicit such information (as long as it is stochastically relevant), it is often impractical to do so. Further, complex information will likely be more difficult for subsequent users to interpret than coarser measures of quality, such as 1 to 5 stars, etc. In this section, we consider situations where the center offers raters a choice between several "coarse" reports, and asks when 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.

In general, the problem of coarse reporting is both subtle and complex. Proper scoring rules induce people to truthfully announce their exact information. One might hope that if the environment were sufficiently smooth, then, when offered a restricted set of admissible reports to announce, a rater 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 $f(s|G)$ and $f(s|B)$. Let $p \in (0,1)$ denote the prior probability (commonly held) that the object is good. We assume that densities $f(s|G)$ and $f(s|B)$ satisfy the monotone likelihood ratio property (MLRP):

$$\frac{f(s|G)}{f(s|B)} \text{ is strictly increasing in } s.$$

MLRP implies the distribution for type G first-order stochastically dominates the distribution for

B (see Gollier, 2001). If a rater observes signal s^i, then she assigns posterior probability $p\left(G|s^i\right)$ to the object's being good, where

$$p\left(G|s^i\right) = \frac{pf\left(s^i|G\right)}{pf\left(s^i|G\right) + (1-p)\, f\left(s^i|B\right)}.$$

MLRP ensures that $p\left(G|s^i\right)$ is strictly increasing in s. 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 it is a best response for a rater 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 score. 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 Proposition 3 is as follows. After observing 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, $p\left(G|s^i\right)$. We begin by dividing the space of PT beliefs into L equal-sized bins. Since $p\left(G|s^i\right)$ is monotone, these PT-belief bins translate to intervals in the rater's signal space that we refer to as signal bins. Note that the signal bins need not be of equal 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 that we score using a quadratic scoring rule.

Since 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, $p\left(G|s^i\right)$. 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 the quadratic scoring rule induces truthful (albeit coarse) bin announcements.

One interesting feature of the construction is 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 $p\left(G|0.5\right) = 0.3$. It is possible that $p\left(G|0.4\right) = 0.2$ while $p\left(G|0.6\right) = 0.35$. 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.[27]

Even in the simple case of 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.

3 Issues in Practical Application

The previous section provides 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.

3.1 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 induce truthful revelation. Risk aversion can be addressed in a number of ways; we present three.

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

[27] Our construction not unique. Others may work as well, and it is likely that the equal-sized bins of signals approach works for some specifications of the underlying distributions.

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. Since von-Neumann Morgenstern utility functions are linear in probabilities, an expected-utility maximizer will also maximize the expected probability of winning the lottery. Thus this procedure induces individuals with unknown non-linear 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.[28]

A third method of dealing with risk averse raters uses the fact that raters' risk aversion is 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 variability of the rater's final payment (measured in terms of its variance) can be reduced if the rater is scored against multiple raters and paid the average of those scores.[29] As the number of reference raters becomes large, the Law of Large numbers implies that, given the object's true type, the rater's payment from reporting truthfully under this scheme converges to the expected score of a truthful report for that type of object. Thus, by paying the rater the average score 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.[30]

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

[29] Since the score from being scored against a particular reference rater depends only on that rater's announcement, it is straightforward to show that truthful revelation remains a best response when scored against multiple reference raters. That this procedure reduces the variance of final payments follows from the Cauchy-Schwartz inequality. At least in the normal information case (Section 2.5.1), the scaling factor needed to induce a particular effort level is the same whether the rater is scored against a single reference rater or is paid the average score from a number of reference raters.

[30] Interestingly, if the rater is scored against the average report of a group of reference raters (as opposed to being scored independently against a number of raters and then paid the average score), this can actually increase the variability in the rater's payment. Holding effort constant, the distribution of the average report of a group of reference raters tends to become more concentrated as the number of reference raters grows. This, however, increases the variability of the *density* of reference rater announcements (as announcements near the median reference-rater announcement become very likely and extreme announcements become very unlikely). Since the scoring rules studied here depend on the level of the density (as opposed to the level of the random report), this tends to increase the variability in the rater's payment.

3.2 Choosing a Scoring Rule

Given a choice, which of the three scoring rules we have discussed is best? Since the logarithmic, quadratic, and spherical rules are all strictly proper, each will elicit truthful revelation. Each rule has its relative strengths and weaknesses, and none emerges as clearly superior.

Of the three, the logarithmic rule is the simplest, giving it a modest advantage in comprehension and computational ease. The logarithmic rule is also "relevant" in the sense that it depends only on the likelihood of events that actually occur, and our results in Section 2.5.1 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.[31] On the other hand, as we mentioned earlier, $\log(x)$ decreases to $-\infty$ as x decreases to zero, which may present problems if raters have limited liability, or if the support of the raters' posterior distributions changes with their information. On a related note, under the logarithmic 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 Section 2.5.2. However, 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

[31] Relevance is important in Bayesian models of comparing different probability assessors (Winkler 1969; Staël Von Holstein 1970).

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), in which cases the logarithmic score is undesirable, there is no clear reason to prefer one scoring rule over the others.

3.3 Conflicts of Interest

Raters may have conflicts of interest. The restaurant owner's friends and family may prefer to give favorable reviews, even if they think the food is mediocre. A vendor might offer bribes to reviewers. In reviewing papers, proposals, or candidates, a reviewer may have a personal or disciplinary interest in the outcome. Advisors may stand to gain or lose financially if an investment is made. In some situations, random selection of reviewers, anonymous reporting, and recusing of reviewers will reduce these conflicts.

It is also possible to adjust the scoring function to try to overwhelm individuals' outside preferences. The same parameter α that was used to scale payments τ^* in order to induce effort can be used to counteract outside preferences. Suppose that a rater gains utility c from a particular report. Although c represented the cost of effort in Proposition 2, the proof of Proposition 2 applies equally well here. Recall that the maximum expected value of any report made without acquiring a signal is $\alpha Z_i(0)$. The proof of Proposition 2 shows how to choose α so that getting a signal and reporting it honestly is worth at least c more than making the rater's externally preferred report. Thus, raters can be induced to ignore any conflict of interest up to the utility they get from some constant c that the system designer chooses. Of course, since there is only one parameter α, if raters can exert variable effort, then setting α to overcome conflicts of interest may also create incentives for raters to exert more than optimal effort.

3.4 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 so as to start the rating 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,..,9, with $f(h|i) = \frac{i}{10}$. 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.[32]

Table 1 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 .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 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.

after signal	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$	$p(7)$	$p(8)$	$p(9)$	$pr(h)$
	.05	.1	.1	.1	.3	.1	.1	.1	.05	.5
h	.01	.04	.06	.08	.3	.12	.14	.16	.09	.59
h, l	.02	.08	.1	.12	.36	.12	.1	.08	.02	.5

Table 1: Initial and updated probabilities of nine types
defined by their probability of yielding signal h.

3.5 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, $f_A(l|t) > f_B(l|t)$. 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 reversed for fans of comedies. Similarly, economists on a review panel might tend to perceive economics proposals more favorably than computer science proposals, and vice versa for the computer scientists.

When tastes differ systematically, the center will need to model rater types explicitly. Given a particular rater's known type or the distribution of types that the rater is drawn from, and given for each rater type the signal densities conditional on product types, the center can compute from a reported signal the posterior distribution of product types. Given a known type for another rater

[32] 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.

or a distribution from which the other rater's type will be drawn, the center can then compute a distribution of signals for the other rater implied by the first rater's report.

As in the simpler case in section 3.4, 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 her type is drawn from.

For example, 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.[33] 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 (or each rater type) for each product not yet rated. That is, instead of predicting that rater j will like movie X 4.27 on a $1-5$ scale, the algorithm could predict that the probability j would score the movie a 5 is .4, the probability of a 4 is .3, and so on. When an additional rating is added from rater i, the predicted distributions for each other rater for that product would be updated.

3.6 Non-Common Priors and Other Private Information

Individuals' private information may affect the ratings they are likely to report. For example, suppose that the National Science Foundation creates a new program solicitation on a topic (homeland security, say) that has been the subject of few proposals in the past. The program officer and each reviewer will form beliefs about the distribution of quality of the initial round of submissions, but their assessments may differ. When the program gets established, there will be a shared prior history, but individuals may have private histories that they weight strongly in forming their own prior beliefs. Similarly, individual raters may have beliefs about the distribution of rater types or of certain types' signal distributions. For example, a reviewer may recognize the names and affiliations of other reviewers on the panel and conclude that 70% are economists even though the center has not classified so many that way.

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

[33] Some algorithms adopt an explicitly Bayesian approach, while others compute weights for other raters in a neighborhood. Others decompose the matrix of ratings, identifying implict underlying dimensions and expressing raters and products as linear combinations of the underlying types. See, for example, Breese, Heckerman, and Kadie (1998) and Sarwar et al. (2000).

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 so as 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.[34] By exactly the same arguments that made honest reporting of signals optimal in the base case, raters will prefer to honestly report their signals and other relevant priors: honest reports will lead the center to score based on a correct posterior distribution of signals for the next rater, and a correct posterior maximizes the score.

Once a rater has received a signal about the current product's type, she may unconsciously update her priors about the distribution of product types or of signals conditional on product types. Ideally, however, she should want to report her unupdated priors because that is the information that will lead the center to make the best predictions about the next rater's distribution of signals. Thus, when raters' priors over product types or distributions of signals contingent on product types is to be reported, it is preferable to elicit this information before the rater is exposed to the product.

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.

3.7 Other Potential Limitations

Three other potential limitations could interfere with the smooth functioning of a scoring system based on the peer-prediction method. 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 will not be possible for all of the raters to be better off through collusive

[34] 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. It would not work to have raters compute the posteriors themselves and report them, because the center would be unable to extract the raters' signal from the other private information she used to calculate her posterior, and thus would be unable to use the rater's signal as an outcome in computing a previous rater's score. In any case, reporting private information separately will often be far easier for a rater and lead to far more accurate computation of posteriors than if raters computed the posteriors themselves.

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, so that collusion would be harder to coordinate. In addition, the center may be able to detect suspicious rating patterns through statistical analysis, and then employ an outside expert to independently evaluate the product.[35].

A second potential limitation may arise when raters perceive multidimensional signals. Our scoring system can generalize easily to handle multiple dimensions by eliciting reports on several dimensions, such as food, decor, and service for restaurants. 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 would 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 partially independent 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 Section 3.5. 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 in order to make the center more closely approximate her true beliefs.

Canice 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

[35] 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.

what she thinks the boss will think.

The best approach to this problem is 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 so 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.

Given the computational power and the information resources available to the center, it will not always be necessary 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 be unable to determine whether giving a false report on the dimensions that the center elicits would increase or decrease 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. The threat of public disclosure of false claims, or even litigation, might reinforce the belief that honesty was the best policy.

4 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 a conceptual roadmap, not a detailed implementation plan, and only for those design decisions that involve incentives for effort and honest reporting. It is an important roadmap, however, because the most obvious approach to peer comparison, simply rewarding for agreement in reviews, does not offer the right 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, so 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.

References

[1] Avery, Chris, Paul Resnick, and Richard Zeckhauser (1999): "The Market for Evaluations," *American Economic Review*, 89(3) 564-584.

[2] Breese, J., D. Heckerman, and C. Kadie, (1998): "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, July, 1998*. Morgan Kaufmann Publisher.

[3] Brynjolfsson, E. and M. Smith (2000): "Frictionless Commerce? A comparison of the Internet and Conventional Retailers," *Management Science*, 46, 563-585.

[4] Clemen, Robert (2002): "Incentive Contracts and Strictly Proper Scoring Rules," *Test*, 11, 195-217.

[5] Congdon, Peter (2001): *Bayesian Statistical Modelling*, Wiley: Chichester, England.

[6] Cooke, Roger M. (1991): *Experts in Uncertainty: Opinion and Subjective Probability in Science,* Oxford University Press: New York.

[7] Crémer, J., and R. McLean (1985): "Optimal Selling Strategies Under Uncertainty for a Discriminating Monopolist When Demands Are Interdependent," *Econometrica*, 53, 345-361.

[8] _____(1988): "Full Extraction of Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica,* 56, 1247-1257.

[9] d'Aspremont, C., and L.-A. Gérard-Varet (1979): "Incentives and Incomplete Information," *Journal of Public Economics,* 11, 25-45.

[10] _____(1982): "Bayesian Incentive Compatible Beliefs," *Journal of Mathematical Economics*, 10, 83-103.

[11] Dellarocas, Chrysanthos (2000): "Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior," ACM Conference on Electronic Commerce EC-00, Minneapolis, http://ccs.mit.edu/dell/ec00reputation.pdf.

[12] _____ (2001): "Analyzing the Economic Efficiency of eBay-like Online Reputation Reporting Mechanisms." Proceedings of the 3rd ACM Conference on Electronic Commerce, Tampa, FL, October 14-16, 2001.

[13] Friedman, Daniel (1983): "Effective Scoring Rules for Probabilistic Forecasts," *Management Science*, Vol. 29(4) 447-454.

[14] Gollier, Cristian (2001): *The Economics of Risk and Time,* MIT Press, Cambridge, MA.

[15] Hanson, Robin (2002): "Logarithmic Market Scoring Rules for Modular Combinatoral Information Aggregation," Working Paper, George Mason University Department of Economics, http://hanson.gmu.edu/mktscore.pdf.

[16] Jensen, Floyd and Cameron Peterson (1973): "Psychological Effects of Proper Scoring Rules," *Organizational Behavior and Human Performance,* 9, 307-317.

[17] Johnson, Scott, Nolan Miller, John Pratt, and Richard Zeckhauser (2002): "Efficient Design with Interdependent Valuations and an Informed Center," Kennedy School Working Paper, RWP02-025, http://ksgnotes1.harvard.edu/research/wpaper.nsf/rwp/RWP02-025.

[18] Johnson, Scott, John Pratt, and Richard Zeckhauser (1990): "Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case," *Econometrica,* 58, 873-900.

[19] Kandori, Michihiro (1992): "Social Norms and Community Enforcement," *Review of Economic Studies,* Vol. 59(1) 63-80.

[20] Kandori, Michihiro and Hitoshi Matsushima (1998): "Private Observation, Communication and Collusion," *Econometrica,* 66(3) 627-652.

[21] Kollock, Peter (1999): "The Production of Trust in Online Markets," *Advances in Group Processes,* Vol. 16.

[22] Kreps, David, and Robert Wilson (1982): "Reputation and Imperfect Information," *Journal of Economic Theory,* 27(2) 253-279.

[23] Lampe, Cliff and Paul Resnick (2004): "Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space," CHI 2004, ACM Conference on Human Factors in Computing Systems, CHI Letters 6(1), 543-550.

[24] Lavalle, I (1968): "On Cash Equivalents and Information Evaluation in Decisions Under Uncertainty: Part I: Basic Theory," *Journal of the American Statistical Association*, 63(321), 252-276.

[25] Mas-Colell, A., M. Whinston, and J. Green (1995): *Microeconomic Theory,* Oxford University Press, New York.

[26] Nau, Robert (1985): "Should Scoring Rules Be Effective?" *Management Science*, 34(5) 527-535.

[27] Nelson, Robert and David Bessler (1989): "Subjective Probabilities and Scoring Rules: Experimental Evidence," *American Journal of Agricultural Economics,* 71, 363-369.

[28] Ottaviani, M. and P. N. Sørensen (2003): "Professional Advice: The Theory of Reputational Cheap Talk," Working Paper, University of Copenhagen, http://www.econ.ku.dk/sorensen/Papers/pa.pdf.

[29] Pratt, John, Howard Raiffa, and Robert Schlaifer (1965): *Introduction to Statistical Decision Theory,* McGraw-Hill, New York.

[30] Prendergast, C. (1993): "A Theory of Yes Men," *American Economic Review,* 83 (4) 757-770.

[31] Resnick, Paul and Hal Varian (1997): "Recommender Systems," *Communications of the ACM*, 40(3) 56-58.

[32] Resnick, Paul, Richard Zeckhauser, Eric Friedman, and Ko Kuwabara (2000): "Reputation Systems: Facilitating Trust in Internet Interactions," *Communications of the ACM* 43(12) 45-48.

[33] Resnick, Paul, and Richard Zeckhauser (2002): "Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System," In *The Economics of the Internet and E-Commerce*, Michael R Baye, editor. Volume 11 of Advances in Applied Microeconomics. Amsterdam, Elsevier Science.

[34] Roth, Alvin (1995): "Introduction to Experimental Economics," in John H. Kagel and Alvin E. Roth (Eds.), *The Handbook of Experimental Economics,* 3-110.

[35] Savage, Leonard (1954): *Foundations of Statistics,* Dover Publications, New York.

[36] Savage, Leonard (1971): "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association,* 66 (336) 783-801.

[37] Sarwar, B. M., G. Karypis, J. A. Konstan, and J. Riedl, (2000): "Analysis of Recommender Algorithms for E-Commerce," In *Proceedings of the ACM E-Commerce 2000 Conference.* Oct. 17-20, 2000, pp. 158-167.

[38] Selten, Reinhard (1998): "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, 1(1) 43-62.

[39] Shapiro, Carl (1982): "Consumer Information, Product Quality, and Seller Reputation," *Bell Journal of Economics,* 13 (1) 20-35.

[40] Shuford, E., A. Albert, and H. Massengil (1966): "Admissible Probability Measurement Procedures," *Psychometrika,* 31, 125-145.

[41] Smith, Cedric (1961): "Consistency in Statistical Inference and Decision," *Journal of the Royal Statistical Society, Series B (Methodological),* 23(1) 1-37.

[42] Staël von Holstein, Carl-Axel (1970): "Measurement of Subjective Probability," *Acta Psychologica,* 34 (Subjective Probability) 146-159.

[43] Tadelis, Steven (2002): "The Market for Reputations as an Incentive Mechanism," *Journal of Political Economy*, 110(4), 854-882.

[44] Winkler, Robert (1969): "Scoring Rules and the Evaluation of Probability Assessors," *Journal of the American Statistical Association,* 64, 1073-1078.

[45] Winkler, Robert (1996): "Scoring Rules and the Evaluation of Probabilities," *Test*, 5(1) 1-60.

A Proofs

Proof of Lemma 1: We argue by contradiction. By Bayes' Rule:

$$g\left(s^j|s^i\right) = \sum_{t=1}^{T} f\left(s^j|t\right) \frac{f\left(s^i|t\right) p\left(t\right)}{\Pr\left(s^i\right)},$$

where $\Pr\left(s^i\right) = \sum_{t=1}^{T} f\left(s^i|t\right) p\left(t\right)$.

If there exists $s^i$and $\hat{s}^i$ such that $g\left(s^j|s^i\right) = g\left(s^j|\hat{s}^i\right)$ for all s^j, then the following must hold for each s_m^j :

$$\begin{aligned}
g\left(s_m^j|s^i\right) - g\left(s_m^j|\hat{s}^i\right) &= 0 \\
\sum_{t=1}^{T} f\left(s_m^j|t\right) \frac{f\left(s^i|t\right) p\left(t\right)}{\Pr\left(s^i\right)} - \sum_{t=1}^{T} f\left(s_m^j|t\right) \frac{f\left(\hat{s}^i|t\right) p\left(t\right)}{\Pr\left(\hat{s}^i\right)} &= 0 \\
\sum_{t=1}^{T} f\left(s_m^j|t\right) \left(\frac{f\left(s^i|t\right) p\left(t\right)}{\Pr\left(s^i\right)} - \frac{f\left(\hat{s}^i|t\right) p\left(t\right)}{\Pr\left(\hat{s}^i\right)} \right) &= 0 \\
\sum_{t=1}^{T} f\left(s_m^j|t\right) \Delta\left(t\right) &= 0, \qquad (10)
\end{aligned}$$

where $\Delta\left(t\right) = \left(\frac{f\left(s^i|t\right)p(t)}{\Pr(s^i)} - \frac{f\left(\hat{s}^i|t\right)p(t)}{\Pr(\hat{s}^i)} \right)$. Since (10) is an inner product, the set of distributions $f\left(s_m|t\right)$ and $p\left(t\right)$ that satisfy it is closed and has Lebesgue measure zero.[36]■

Proof of Proposition 2: Let $Z_i(0) = \arg\max_a \sum_{n=1}^{M} R\left(s_n^{r(i)}|a\right) f\left(a\right)$, so that the maximum expected value of any report made without acquiring a signal is $\alpha Z_i(0)$. Let

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

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 $Z_i\left(1\right) > Z_i\left(0\right)$ when stochastic relevance holds.

Pick $\alpha > \frac{c}{Z_i(1) - Z_i(0)}$. 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 product type (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 $B_l = [\frac{l-1}{L}, \frac{l}{L})$, and $B_L = \left[\frac{L-1}{L}, 1\right]$. Given these bins, the rater's PT belief induces a reference rater bin announcement (RRA) belief. Let $P_G^l = \int_{\frac{l-1}{L}}^{l} f\left(s|G\right) ds$ and $P_B^l = \int_{\frac{l-1}{L}}^{l} f\left(s|B\right) ds$, 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

[36] To show this, note that (10) is satisfied only if all $\Delta t = 0$, or, failing this, $f\left(s_m|t\right)$ satisfies a linear equation of the form $\sum_{t=1}^{K} x_t \Delta t = 0$. It is straightforward to show that these restrictions are only satisfied non-generically.

the l^{th} bin is:

$$P^l_{s^i} = \int_{\frac{l-1}{L}}^{l} p\left(G|s^i\right) f\left(s|G\right) + \left(1 - p\left(G|s^i\right)\right) f\left(s|B\right) ds = p\left(G|s^i\right) P^l_G + \left(1 - p\left(G|s^i\right)\right) P^l_B,$$

Let $P_{s^i} = \left(P^1_{s^i}, ..., P^L_{s^i}\right)$ denote the RRA distribution of a rater who has observed s^i.

Since $p\left(G|s\right)$ is monotone in s, the inverse function $\pi\left(p\right)$ is well-defined. Let $\tilde{B}_l = [\pi\left(\frac{l-1}{L}\right), \pi\left(\frac{l}{L}\right))$ be the l^{th} bin of signals and $\tilde{B}_L = \left[\pi\left(\frac{L-1}{L}\right), \pi\left(1\right)\right]$; i.e., raters observing signals in $\tilde{B}_l$ have PT beliefs in B_l. A rater who announces that her signal is in $\tilde{B}_l$ is paid using the quadratic scoring rule based on the RRA distribution for a rater who has PT belief $m_l = \frac{2l-1}{2L}$. 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 $\hat{P}$ denote two RRA distributions, the L_2-distance between them, $d\left(P, \hat{P}\right)$, is given by:

$$d\left(P, \hat{P}\right) = \left(\sum_l \left(P^l - \hat{P}^l\right)^2\right)^{1/2}.$$

A rater who observes signal s^i assigns probability $P^l_{s^i} = p\left(G|s^i\right) P^l_G + \left(1 - p\left(G|s^i\right)\right) P^l_B$ 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:

$$d\left(P_{s^i}, P_{\hat{s}^i}\right) = \left(\sum_l \left(P^l_{s^i} - P^l_{\hat{s}^i}\right)^2\right)^{1/2} = \left|p\left(G|s^i\right) - p\left(G|\hat{s}^i\right)\right| \left(\sum_l \left(P^l_G - P^l_B\right)^2\right)^{1/2}. \quad (11)$$

Expression (11) 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 $m_1, ..., m_L$, a rater observing s^i maximizes her expected quadratic score by choosing the m_l that is closest to $p\left(G|s^i\right)$, i.e., her true PT beliefs. This follows from a result due to Friedman (1983, Proposition 1), who shows that the expected quadratic score of a rater with true RRA P is larger from reporting $\hat{P}$ than from reporting $\tilde{P}$ if and only if $d\left(\hat{P}, P\right) < d\left(\tilde{P}, P\right)$.[37] Thus Friedman's result, in conjunction with (11), 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.■

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 $f\left(s_m|t\right)$. Greater effort corresponds to obtaining a larger sample. Let x_i denote the number of

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

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.[38]

For a rater who already observes a sample of size x, learning the $x+1^{st}$ component further partitions the outcome space, i.e., larger samples correspond to better information.[39] 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.[40] Let $y^i = \left(y_1^i, ..., y_M^i\right)$ denote a generic realization of $Y^i(x_i)$, where y_m^i is the number of times out of x_i that signal s_m is received, and note that $\sum_{m=1}^{M} y_M^i = x_i$. 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. Since different realizations of $Y^i(x_i)$ yield different posterior beliefs about the product's type, we can extend Lemma 1 to the multiple signal case. In the remainder of this section, we let $g\left(y^j(x_j)\,|y^i(x_i)\right)$ denote the distribution of $Y^j(x_j)$ conditional on $Y^i(x_i)$.

Lemma 2: *Consider distinct players i and j, and suppose x_i, $x_j \geq 0$ are commonly known. For generic distributions $f(s_m|t)$ and $p(t)$, $Y^i(x_i)$ is stochastically relevant for $Y^j(x_j)$. 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., $R\left(y^j(x_j)\,|y^i(x_i)\right)$, then the rater's expected payment is uniquely maximized by announcing the true realization of $Y^i(x_i)$.*

Proof: The proof of the first part is analogous to the proofs of Lemma 1. The second part 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 $i = 1,..., I$. It follows as an immediate consequence of Lemma 2.

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:

$$\tau_i^*\left(y^j|a^i\right) = R\left(y^j|a^i\right). \tag{12}$$

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

[38] In a single-agent context, Clemen (2002) examines the incentive problem when the center does not know $c_i(x_i)$.

[39] Savage (1954) formally studied this approach, which he calls the "partition problem."

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

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 (12). Since 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.

Since 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:

$$Z_i(x_i) = E_{Y^i}\left(E_{Y^j} R\left(Y^j | Y^i(x_i)\right)\right).$$

Lemma 3 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 3: *$Z_i(x_i)$ is strictly increasing in x_i.*

Proof of Lemma 3: Fix x_i and let y^i be a generic realization of $Y^i(x_i)$. Conditional upon observing y^i, rater i maximizes her expected transfer by announcing distribution $g\left(Y^j | y^i\right)$ for rater j's information. Suppose rater i observes the $x_i + 1^{st}$ component of her information. By Lemma 2, i's expected transfer is now strictly maximized by announcing distribution $g\left(Y^j | \left(y^i, s_m\right)\right)$, and rater i increases her expected value by observing the additional information. Since this is true for every y^i, it is true in expectation, and $Z_i(x_i + 1) > Z_i(x_i)$.■

Lemma 3 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 (12), 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 (12) induce rater i to collect a sample of size $x_i^* \in \arg\max\left(Z_i(x_i) - cx_i\right)$.

Rater i's incentives to truthfully report are unaffected by a uniform scaling of all transfers in (12). 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 (13) extends the transfers described in (12) to allow for multiple signals and a rescaling of all payments by multiplier $\alpha_i > 0$:

$$\tau_i^*\left(a^i, y^{r(i)}\right) = \alpha_i R\left(y^{r(i)} | a^i\right). \tag{13}$$

Under transfers (13), 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 $\hat{x}_i \in \arg\max \hat{\alpha} Z_i(x_i) - c(x_i)$. 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 (13), rater i chooses sample size $\hat{x}_i$.*

Proof of Proposition 5: Since $Z_i(x)$ is concave, sample size $\hat{x}_i$ is optimal if there exists $\hat{\alpha}_i$

satisfying

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

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

$$\begin{aligned}\hat{\alpha}_i &\leq \frac{c_i(\hat{x}_i + 1) - c_i(\hat{x}_i)}{Z_i(\hat{x}_i + 1) - Z_i(\hat{x}_i)}, \text{ and} \\ \hat{\alpha}_i &\geq \frac{c_i(\hat{x}_i) - c_i(\hat{x}_i - 1)}{Z_i(\hat{x}_i) - Z_i(\hat{x}_i - 1)}.\end{aligned}$$

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.[41] 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.[42]

[41] While we are not aware of any general results pertaining to the shape of the $Z()$ function, Clemen (2002) provides a number of examples of cases in which $Z_i(x_i + 1) - Z_i(x_i)$ decreases in x_i. The problem of finding the set of sample sizes that maximize $\alpha Z_i(x_i) - c(x_i)$ for some α when $Z_i(x_i)$ is not concave is isomorphic to the problem in production theory of finding the set of outputs that maximize profit for some output price when the production function is not concave. The solution involves finding the set of outputs that remain on the convex closure of the technology set. See, for example, Mas-Colell, Whinston, and Green (1985, Section 5.D).

[42] 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.