February 12, 2006

Ownership Structure

and

Enforcement Incentives at Self-regulatory Financial Exchanges

David Reiffen*
U.S. CFTC

and

Michel Robe
U.S. CFTC and
Kogod School of Business,
American University

Abstract:

In the last few years, many of the largest financial exchanges have converted from mutual, not-for-profit organizations to publicly-traded, for-profit firms. In most cases, these exchanges have substantial responsibilities with respect to enforcing various regulations that protect investors from dishonest agents. We examine how the incentives to enforce such regulations change as an exchange converts from mutual to for-profit status. In contrast to some oft-stated beliefs, we find that, in many circumstances, an exchange that maximizes shareholder (rather than member) income has a greater incentive to aggressively enforce these types of regulations.

Keywords: Regulation of financial institutions, Enforcement delegation, Trade practice violations, Governance, Ownership structure
JEL codes: G28, D02, K23

* Corresponding author. Mailing address: U.S. Commodity Futures Trading Commission, 1155 21st Street, NW, Washington, DC 20581. E-mail: dreiffen@cftc.gov. Telephone: (+1) 202-418-5602. We are grateful to Jiro Kondo for helpful discussions and suggestions. We thank Mike Penick, Ria Steiger, and participants to the 2006 Meeting of the Washington Area Finance Association, the 2006 Meeting of the European Association for Law and Economics (Madrid), and the 2006 Conference on Market Structure and Integrity (Toronto) for useful comments. This paper reflects the opinions of its authors only, and not those of the U.S. Commodity Futures Trading Commission (CFTC), the Commissioners, or any of the authors' colleagues upon the Commission staff. All errors and omissions, if any, are the authors' sole responsibility.

"[T]he profit motive of a shareholder-owned SRO (self-regulatory organization) could detract from proper self-regulation. For instance, shareholder-owned SROs may commit insufficient funds to regulatory operations."
U.S. Securities and Exchange Commission, SEC Concept Release, Fall 2004

"(W)hen operated by a management team whose main goal is to create a profit, an exchange may have less interest in devoting resources to its regulatory functions."
IMF Financial Sector Assessment: A Handbook, 2005

"Might a for-profit, publicly-traded SRO attempt to attract volume or increase its profits through lax self-regulation?"
U.S. Commodity Futures Trading Commission, 2005 Request for Comments

I. Introduction

Following the example of the Stockholm Stock Exchange and the Deutsche Börse in the early 1990's, many of the world's major financial exchanges have converted from mutual, not-for-profit organizations to publicly-traded, for-profit firms. Since 2000, institutions such as the London Stock Exchange, the Hong Kong Stock Exchange and the Sydney Futures Exchange have demutualized. In the United States, the two largest stock markets (the New York Stock Exchange and NASDAQ) as well as the three main futures exchanges – the Chicago Mercantile Exchange (CME), the Chicago Board of Trade (CBOT), and the New York Mercantile Exchange (NYMEX) – have all adopted the for-profit form.[1]

In most cases, the mutual exchanges had substantial self-regulatory (SR) authority. Most significantly for investors (i.e., for the customers whose trades are executed on exchanges), these exchanges had legal authority to establish and enforce a variety of rules governing the behavior of exchange members. While some of the newly demutualized entities have established independent subsidiaries for regulatory operations or even completely outsourced them,[2] many for-profit exchanges have retained these self-regulatory responsibilities.

As the pace of demutualization has accelerated, concerns have grown that for-profit exchanges might neglect their self-regulatory responsibilities. In particular, because enforcement

[1] See Aggarwal and Dahiya (2006) for a historical perspective on demutualization.
[2] For example, NASD Regulation oversees and regulates all trading on NASDAQ and in the over-the-counter (OTC) markets, as well as trades in New York Stock Exchange- and Amex-listed securities reported to NASDAQ. Late in 2003, the National Futures Association entered into a Regulatory Services Agreement with Eurex US to provide market surveillance and trade practice surveillance services.

activities are costly, "self-enforcement" could become "too little enforcement" if demutualized exchanges commit insufficient resources to regulatory operations in a bid to maximize profits. Even if an exchange contracts out these duties, the same basic fear remains that it may have incentives to under-fund the subsidiary. This concern is articulated in documents released by the U.S. Securities and Exchange Commission (SEC) and Commodity Futures Trading Commission (CFTC), the International Organization of Securities Commissions (IOSCO), the International Monetary Fund (IMF), as well as in the statements of many commentators on the effects of demutualization.[3]

In this paper, we examine the relationship between self-regulation and SRO ownership structure. We focus on a specific area of concern – whether, once demutualized, exchanges will continue to enforce trade practice regulations with the same vigor. "Trade practice regulations" refer to the rules governing how the end-investors' agents (stock specialists, dealers, futures brokers, etc.) carry out their customers' trades.[4]

We start from the observation that market surveillance and enforcement activities exist *because* investors' agents have incentives to misbehave. If investors are aware of these incentives, they will need reassurance that the exchanges where their trades are carried out adequately monitor agents and enforce penalties for wrongdoing. Hence, an exchange that cuts surveillance and enforcement expenditures runs the risk that investors will decline to trade on that exchange. In other words, cost-cutting could itself be costly.

To capture these features, we analyze whether for-profit exchanges have greater or lesser incentives to enforce trade practice regulations in the context of a model in which agents have better information than do their customers about the outcomes of trades. Agents can exploit that information to their advantage. The exchange can investigate suspected misrepresentations and metes out penalties to the wrongdoers it identifies, but monitoring is costly. This type of costly state verification model (CSV) has been used extensively to evaluate principal-agent problems in the contexts of debt contracts,[5] various agency relationships, and in self-regulatory organizations

---

[3] See, e.g., SEC (2004); CFTC (2005); IOSCO Technical Committee (2006); IMF (2005); and, Elliot (2002).

[4] For example, regulations prohibit agents from trading ahead of a customer's order (front-running) or from taking a client's order and completing it by routing it to a pre-arranged counter-party.

[5] See, e.g., Townsend (1979), Gale & Hellwig (1985), Boyd & Smith (1994), and references cited therein.

(DeMarzo, Fishman & Hagerty, 2005 – hereafter, DFH). We extend this prior work by analyzing the effect of ownership structure on the stringency of enforcement; the effect of agent wealth on contract terms and the extent of trade practice violations; and the interaction of these effects.

Our principal finding is that, contrary to the fear of many commentators, for-profit SROs have greater incentives to enforce rules than do mutual SROs. Intuitively, the goal of a mutual SRO is to maximize member income, and so it adopts an enforcement policy that creates positive incentives for honest reporting (carrots). In contrast, a for-profit SRO is less interested in member income, and as a result relies to a greater extent on punishment for dishonest behavior (sticks) in the form of greater likelihood of investigation to insure honest reporting. Our results suggest that corporate governance mechanisms meant to ensure that SRO shareholder income is maximized should lead to tighter enforcement at these SROs.[6]

We find that greater agent wealth allows SROs (whether mutual or for-profit) to reduce the frequency of investigations without inducing misreporting. We also evaluate an extension of the model in which agents are heterogeneous. In this augmented framework, we identify conditions under which misreporting can occur in equilibrium. We show that, *ceteris paribus*, misreporting is more likely when the SRO is a mutual exchange.

The remainder of the paper proceeds as follows. Section II outlines our stylized model of self-regulating organizations (SROs). Section III characterizes the optimal strategies for customers, agents and the SRO, when all agents are homogenous. Section IV considers some extensions, including environments in which agents are heterogeneous. Section V concludes.

II. A Model of Self-Regulation in Financial Markets

Our goal is to model the decisions of agents in financial markets regarding whether to honestly represent the interests of their clients, and the impact of an exchange's objectives (i.e., its ownership structure) on these decisions and responses. Specifically, we consider agents who carry out their clients' wishes to trade on organized exchanges. In doing so, we model clients

[6] On January 31, 2007, the CFTC issued a "best practices" guideline for exchange governance, suggesting that 35% or more of board members be independent. This guideline was issued with the intention of strengthening incentives for self-regulation at for-profit derivatives exchanges.

and exchanges as rationally anticipating the behavior of agents given the reward schedule. In the United States and elsewhere, many of these exchanges are self-regulatory organizations, whereby the exchanges enforce rules about permissible trading behavior by customers' agents.[7]

On U.S. futures exchanges, for instance, federal regulations prohibit a variety of trade practices that include "front-running," "bucketing," "changing prices," etc.[8] These regulations are collectively known as "trade practice" rules. The element common to the prohibited practices is that they allow the agent to misrepresent the best available price, to the detriment of the customer. For example, front-running a trade for a client who wants to establish a long position may result in the customer's paying a higher price to establish that position than if the customer's trade had been made prior to the agent's trade.

We employ a stylized model of this environment to evaluate how investors respond to the potential for dishonest agents and how the SRO chooses an enforcement policy. There are three kinds of parties in this model: investors, i.e., customers; agents, who conduct trades on behalf of these investors; and the exchange (SRO) on which the trading takes place. Following DFH, we focus on a single exchange. We model competition across trading venues by assuming that customers do not trade on the exchange unless they expect to at least achieve an exogenous reservation utility level, α, which could represent the utility expected from transacting on an alternative trading platform.

We posit that an agent's true trading cost in carrying out a customer's trade is not observable by the customer. Instead, the customer only receives the agent's report of the cash-flow (net of the agent's trading costs) generated by the trade. We assume that the true cash-flow W takes on one of two values: w_2 with probability π_2, or w_1 with probability π_1, where $w_2 > w_1$, greater meaning that it is more advantageous for the customer. For example, suppose that the customer is selling a share of stock. Then w_2 is the high realization of the price received, and w_1

[7] In many countries, a government agency (such as the SEC or the CFTC in the United States) has oversight of the SRO's enforcement practices, including the ability to conduct additional inspections and to potentially sanction the SRO for failing to enforce rules. The potential for government intervention, which brings about a more stringent enforcement policy by a mutual SRO (see DFH), does not affect our main conclusions.

[8] "Front-running" refers to trades made by the agent (broker or futures commission merchant) on the same side of a market prior to executing an order that has already been placed by a client. "Bucketing" refers to an agent's taking the opposite side of a customer's order. "Changing prices" refers to an agent's misrepresenting the actual trading price. See, e.g., Johnson and Hazen (1999) for a detailed discussion of trade practice rules. See also Cummins and Johan (2006) for an empirical analysis of surveillance activities in 25 of the world's largest financial exchanges.

is the low realization. This captures the general notion of unobservable states of the world in a tractable model. To avoid trivial cases, we assume that $w_1 < \alpha < w_2$, so that the customer is willing to trade, but only if she expects to receive a payment above w_1.

The SRO oversees the agent's behavior, in the sense that it observes the agent's report of the realized state, can choose to investigate whether the report was accurate and, if it was not, the size of the fine X to impose on the agent. We assume that the agent has limited liability so that the fine cannot exceed the agent's resources, i.e., the sum of the ill-gotten gains and of the agent's pre-trade wealth, γ.[9] Any fine thus levied is paid to the injured customer. The customer is fully informed as to the parameters of the SRO's enforcement policy – both the probabilities p_i ($i=1,2$) that the SRO will review a transaction that is reported to be in *state i* and the penalties x_{ij} ($i,j=1,2$) to be meted out if the agent reports state w_j and the SRO finds out that the true state was w_i. The client chooses with which agent to trade and the terms of the contract, given the enforcement policy $\{P, X\}$ of the exchange. In this context, the contract consists of a schedule specifying the state-contingent transfer to the customer, $Z(W)$, and agent's fee, $W\text{-}Z(W)$.

We assume that there are a large number of agents competing for each client's business and, hence, that the client chooses a contract to maximize her own surplus from trading, subject to the agent's receiving non-negative profits. In Section III, all agents are posited risk-neutral and otherwise identical. We relax this assumption in Section IV by allowing agents to differ with respect to their wealth and, hence, the maximum fine they can pay.

The SRO sets its enforcement policy in anticipation of the behavior of customers and agents, to maximize its own objective function. A key goal of our analysis is to compare the enforcement policy of a not-for-profit, mutual exchange with that of a for-profit, demutualized exchange. To do so, we take the self-regulatory exchange's decision to demutualize as a given,[10] and then investigate the implications of that decision on the SRO's optimal enforcement policy.

---

[9] In a dynamic environment, the SRO would also have the option of suspending the agent's trading privileges, which has an effect similar to that of a fine. The static framework adopted here captures the essence of the trade-off inherent in the agent's decision of whether to honestly report the realized state. In the dynamic environment, γ could represent the future earnings loss to a trader whose trading privileges are suspended.

[10] For example, Hart and Moore (1996, 1998) and Pirrong (2000) conjecture that demutualization may help win over certain traders opposed to modernization because they would not otherwise benefit from technological innovation.

We posit, in line with DFH, that a mutual SRO seeks to maximize agent income (subject to customers' expecting to receive their reservation utility α) and sets its transaction fee t_{NFP} to cover its expected enforcement costs: $t_{NFP} = (P_{NFP}\ \Pi)c$, where P_{NFP} is the mutual exchange's vector of investigation probabilities, Π is the vector of states-of-the-world probabilities, and c is the unit investigation cost. By contrast, we assume that a demutualized exchange seeks to maximize its shareholders' income (trading fees net of expected investigation costs), subject to customers and agents expecting to receive their respective reservation utilities.[11] This choice of objective function: $t_{FP} – (P_{FP}\ \Pi)c$ captures the concern that a for-profit exchange has incentives to curtail its enforcement expenditures.

The next section assesses whether this fear is warranted, given that an exchange must optimize its objective function subject to the constraints it faces – in particular, the constraint that customers will not trade on that exchange unless their expected gains from trading there, which depend on its enforcement policies, are high enough.

III. Enforcement Policies under Alternative Ownership Structures

As noted above, we model the behavior of customers who offer fee schedules to agents who represent them, given the enforcement policy chosen by the SRO. If a customer's expected income at the optimum is less than her opportunity cost α, the customer does not trade. Given a fee schedule and the enforcement policy of the exchange, agents decide which message to send (e.g., which price to report), conditional on the exogenous true state. The SRO chooses its enforcement policy in anticipation of the behavior of customers and agents. The remainder of this section derives the subgame perfect equilibrium of this game.

A. The Customer's Optimization:

As is standard in CSV models, a customer contracts with an agent to make a trade (e.g., sell a share of stock) and the customer cannot observe the realization of the trade. All the agents

[11] The assumption that the exchange collects revenues by charging per-unit-trade fees matches practices at the major U.S. futures exchanges.

are risk-neutral and otherwise identical (we relax this assumption in Section IV below). The customer, who takes the exchange's enforcement policy as given, wants to set a fee schedule that induces the agent to tell the truth about the realized trade. Specifically, the risk-neutral customer wants to set a schedule of fees to maximize her expected income from the trade, subject to the constraints that (i) the agent tells the truth (agent incentive compatibility constraint – AIC); (ii) the agent is better off serving the customer than not (agent participation, or individual rationality constraint – AIR); (iii) the agent earn a non-negative return whenever he correctly reports the true state (no loss condition – NLC). Formally, the customer's problem (CP) is to:[12]

$$\max_{\{z(w1),z(w2)\}} \ \pi_2 z(w_2) + \pi_1 z(w_1) - t \qquad \text{(CP)}$$

subject to

AIC (1) $w_2 - z(w_2) \geq p_1(\max\{w_2 - z(w_1) - x_{12}, -\gamma\}) + (1-p_1)(w_2 - z(w_1))$

AIR (2) $\pi_2 (w_2 - z(w_2)) + \pi_1(w_1 - z(w_1)) \geq A_0$

NLC (3) $w_i - z(w_i) \geq 0$; $i = 1,2$

where:

π_i	is the likelihood of state i, with $1 > \pi_2 = 1-\pi_1 > 0$;
$z_i \equiv z(w_i)$	is the customer's return in state i (e.g., how much the customer receives from the sale of a stock when the agent announces that the state is state i);
t	is the transaction fee charged by the SRO;
$w_i - z(w_i)$	is the agent's fee in state i (i.e.., how much the customer pays the agent when the latter reports *state i*);
p_i	is the probability that the SRO will review a transaction that is reported to be in state i ($i = 1,2$);
x_{ij}	is the penalty to the agent if he announces state i, but state j is the true state and the SRO catches the agent misreporting;
$A_0 \geq 0$	is the agent's opportunity cost;
$\gamma \geq 0$	is the agent's pre-transaction wealth;[13]

[12] Condition (3) implies that the AIR condition (2) is redundant if $A_0 = 0$, as in DFH. We impose both constraints in this Section to set up the parallel with the weaker version of condition (3) that we analyze in Section IV.B. There, we consider optimal contracts in a more general model, in which condition (2) is not redundant even if $A_0 = 0$. The qualitative results of the present section carry through to that alternative model.

[13] In the static model in this paper, one could interpret wealth as the agent's earnings from his other clients. In a dynamic model, γ could reflect the future earnings of agents, which can be reduced by SRO action (e.g., by suspending trading privileges)

Implicit in the statement of the AIC constraint is the fact that the agent has no incentive to misrepresent the poor outcome, w_1. Embedded in the statement of the AIR constraint is the result that the SRO will set penalties $x_{ij} = 0$ for all i, j except perhaps for x_{12} (i.e., when the high return w_2 is realized but the agent pretends the low return w_1 has been realized). This penalty structure is optimal for both the mutual and for-profit SROs.[14]

The customer's problem is similar to that in DFH, except that we allow the agent wealth $\gamma \geq 0$, rather than restrict γ to 0. Figure 1 illustrates graphically three of the constraints faced by the customer in this environment. The agent incentive-compatibility constraint (1) is depicted by the upward-sloping AIC line. The line, drawn for the case where $w_2 - z(w_1) + \gamma \geq x_{12}$, can be written as $z(w_2) = z(w_1) + p_1\, x_{12}$.[15] To ensure incentive compatibility, $z(w_2)$ must lie on or below this line. The individual rationality constraint (2) is represented by the downward-sloping AIR line, $z(w_2) = w_2 - \pi_1 w_1/\pi_2 - \pi_1 z(w_1)/\pi_2$. To ensure agent participation, $z(w_2)$ must lie on or below this line. Finally, the limited liability constraint (3) is shown for $i = 1$ by the vertical NLC line at $z(w_1) = w_1$; $z(w_1)$ must lie to the left of this line.[16] The shaded five-sided area depicts the combinations of $z(w_1)$ and $z(w_2)$ that meet all three constraints (AIC, AIR and NLC).

The customer's income in Figure 1 is just the mirror image of the agent's income, in that the expected aggregate income to the two parties is always $\Sigma \pi_i w_i$, minus the SRO fee t. In other words, the customer's iso-income lines are parallel to the agent participation constraint in Figure 1. The solid AIR line represents the maximal customer income (gross of enforcement costs) that is consistent with the agent-participation constraint (2). Whether the AIR constraint is binding or not (i.e., whether maximized customer income falls short of this amount or not) depends on the exogenous parameters p, x, and γ and on the exchange's objective function. In either case, the customer's constrained optimization is to set $z^*(w_1) = w_1$.[17] Setting instead $z(w_1) < w_1$ would not only lower the customer's payment in state 1, but would also lower the maximum $z(w_2)$ that is incentive-compatible – i.e., that is consistent with (1). Given this constraint, the highest income

---

[14] It is straightforward to extend arguments in DFH to show that this result also holds for the mutual SRO with the weaker version of (3) used in Section IV.B. As discussed below, the same result holds for the for-profit SRO.

[15] If it were optimal for the SRO to set $x_{12} > w_2 - z(w_1) + \gamma$, then the AIC would take the form $z(w_2) = (1 - p_1)\, z(w_1) + p_1\, (w_2 + \gamma)$. In equilibrium, however, it does not matter whether the SRO selects $x_{12} > w_2 - z(w_1) + \gamma$ or $x_{12} = w_2 - z(w_1) + \gamma$, because the $z(w_i)$ chosen are the same in either case. We therefore focus here on the latter case.

[16] As noted in footnote 10, the AIR is redundant in this case if $A_0 = 0$, but can be binding for $A_0 > 0$.

[17] In Section IV.B, we evaluate customer choice under the weaker condition that $z(w_1) \leq w_1 + \gamma$. As shown there, the qualitative results derived here are largely unaffected when one imposes the weaker constraint.

the customer can obtain is depicted by the downward sloping line going through the intersection of the incentive compatibility constraint (1) and the vertical NLC line representing the no-loss constraint (3). At this point, constraint (2) is not binding (for $A_0 = 0$). Thus, as in DFH, the customer will set $z^*(w_2, \gamma) = p_1 x_{12} + w_1$. Lemma 1 summarizes these results:

Lemma 1: Given the exchange's enforcement policy (i.e., given investigation likelihood p_1 and penalty for wrongdoing x_{12}), the customer sets $z^*(w_1) = w_1$ and $z^*(w_2) = p_1 x_{12} + w_1$.

B. Mutual SRO's Optimization

Given this behavior by the customer, a mutual SRO seeks to maximize agent income using the p_i's and x_{ij}'s as instruments, subject to the constraint that customers expect to earn their reservation levels of income (customer individual rationality – CIR).

Formally, the mutual SRO's problem (MP) is to maximize

$$\pi_2 (w_2 - z(w_2)) + \pi_1 (w_1 - z(w_1)) \qquad \text{(MP)}$$

with respect to the enforcement parameters (x_{ij} and p_i) subject to the customer's expecting income of at least α:

CIR (4) $\pi_2 z(w_2) + \pi_1 z(w_1) - t \geq \alpha$

as well as two agent constraints:

AIC (1) $w_2 - z(w_2) \geq p_1[\max\{w_2 - z(w_1) - x_{12}, -\gamma\}] + (1-p_1)(w_2 - z(w_1))$

NLC (3) $w_i - z(w_i) \geq 0, i = 1,2$

where t is the fee that the exchange charges customers per transaction. Reflecting the non-profit nature of a mutual SRO, we follow DFH and assume that t is set equal to the expected number of inspections times the unit cost, or $\pi_1 p_1 c$.

As discussed above, the customer's choice of $z(w_i)$ reflects the constraints (1) and (3). Furthermore, $t = \pi_1 p_1 c$. Thus, we can replace these two constraints with the single constraint:

CIR' (4') $\pi_2 z^*(w_2) + \pi_1 z^*(w_1) - \pi_1 p_1 c \geq \alpha$

where $z^*(w_i, \gamma)$ reflects the optimized value of $z(w_i)$ subject to constraints (1) and (3).

We have assumed, following DFH, that $\alpha > w_1$.[18] It is worth noting that this assumption puts bounds on how large the cost c can be. Specifically, the customer receives $\pi_2\, z(w_2) + \pi_1\, z(w_1) - t = p_1\, \pi_2\, (w_2 - w_1 + \gamma) + w_1 - \pi_1\, p_1\, c$, which must be greater than α for the CIR to hold. Given $\alpha > w_1$, no solution exists unless $c \leq (w_2 - w_1 + \gamma)\pi_2/\pi_1$.[19]

Lemma 2 derives the SRO's optimal enforcement parameters.

<u>Lemma 2:</u> If $\pi_2\, (w_2 - w_1 + \gamma) - \pi_1\, c \geq \alpha - w_1$, then the mutual SRO will choose:

$$p_1 = p_M = \frac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma) - \pi_1 c}$$

$p_2 = 0$, $x_{12} = w_2 - w_1 + \gamma$ and $x_{ij} = 0$ for all other i,j.

Proof: See Appendix. □

Intuitively, Lemma 2 shows that the SRO creates an enforcement environment – *via* positive fines for misreporting (which are paid to the customer) and a positive probability of detection – such that customers choose to give their agents some incentive for honesty. This SRO policy maximizes agent income, subject to the constraint that customers expect an income of α. For any given expected fine (i.e., for any $p_1 x_{12}$), it is profit-maximizing for the SRO to set x_{12} as high as possible (i.e., to choose $x_{12} = w_2 - w_1 + \gamma$) because the concomitant decrease in the probability of investigation, p_1, reduces enforcement expenditures – which, in turn, allows for higher agent fees in equilibrium. Lemma 2 thus also implies that $\partial p_M/\partial\gamma < 0$; that is, higher agent wealth allows the exchange to select higher penalties, which allows it to reduce p_1 while holding $p_1 x_{12}$ fixed.

The condition that $\pi_2\, (w_2 - w_1 + \gamma) - \pi_1\, c \geq \alpha - w_1$ in Lemma 2 (which is implicit in DFH, for $\gamma = 0$, and hence holds for all $\gamma < \alpha - w_1$) requires that there be some p_1 that allows the SRO to

[18] In the trivial case where $\alpha \leq w_1$, then it is optimal for the SRO to set $p_1 = 0$, in which case the agent always reports that state 1 has occurred, and the customer receives w_1, regardless of which state actually occurred.

[19] This result strengthens the conclusion in Proposition 2 of DFH. These authors find that a welfare-maximizing SRO would either set $p_1 = 0$ or $p_1 = 1$ depending on the sign of $c - (w_2 - w_1)\pi_2/\pi_1$. The analysis here implies that an SRO seeking to maximize customer surplus would set $p_1 = 1$.

provide an enforcement regime in which customers can earn an income greater than or equal to α. The interpretation of this inequality is that $\alpha – w_1$ would be the customer's loss in income from trading if p_1 were equal to 0 (recalling that $z^*(w_2) = z^*(w_1) = w_1$ if $p_1 = 0$). The left-hand side, $\pi_2 (w_2 – w_1 + \gamma) – \pi_1 c$, is the gain in customer income from a unit increase in p_1. Thus, if $\pi_2 (w_2 – w_1 + \gamma) – \pi_1 c < \alpha – w_1$, increases in p_1 are insufficient to make up for the entire income loss.

Figure 2 shows how changing the investigation probability p_1 affects the customer's expected income. At p_1 = p', customers can only reach the income level C_1. If $C_1 < \alpha$, then condition (4) is not satisfied. In order to meet the customer's participation constraint, the SRO needs to increase p_1 to a level, say p'', that enables customers to reach income level $C_0 = \alpha$.

Figure 3 shows the relation between the customer's expected income and the exogenous agent wealth, γ. As γ rises (for a fixed w_1), the customer can move to higher income from C_0 to C_2. If p_1 were held constant at p'', this move would reduce agent income. The SRO therefore lowers p_1 (to p') as γ increases, so that px is kept constant and the customer's expected income remains equal to α. The agent's income correspondingly rises. Furthermore, because combined net income (i.e., agent income plus customer income minus enforcement costs) rises as p_1 falls, the agent is strictly better off with higher γ.

C. The Profit-maximizing SRO's Optimization

The profit-maximizing exchange earns its revenues through the transactions fee, t. Given that the SRO is risk-neutral and $A_0 = 0$ (as in DFH), its maximization problem is

$$\max_{\{t,\, p,\, x\}} \; t - \pi_1 c p$$

subject to

CIR' (4') $\pi_2 z^*(w_2) + \pi_1 z^*(w_1) - t \geq \alpha$

Again, $z^*(w_1)$ and $z^*(w_2)$ incorporate constraints (1) and (3). As shown above, $z^*(w_1) = w_1$ and $z^*(w_2) = w_1 + p_1 x_{12}$. The CIR constraint must bind as well, otherwise t could be increased (thereby raising the SRO's objective) without inducing customers or agents to exit. Hence:

Lemma 3: As long as $w_1 + \pi_2 (w_2 – w_1) – \alpha > \pi_1 c (w_2 – w_1) / (w_2 – w_1 + \gamma)$, the for-profit SRO sets

$$p_1 = p_F = \frac{w_2 - w_1}{w_2 - w_1 + \gamma}$$

$t = w_1 + \pi_2 (w_2 – w_1) – \alpha; p_2 = 0$; $x_{12} = w_2 – w_1 + \gamma$, and $x_{ij} = 0$ for all other i,j.

Proof: See Appendix. □

Intuitively, the for-profit SRO chooses values for *p, x,* and *t* so that both agents and customers only receive their reservation values. The SRO sets the penalty *x* as high as it can given agent-liability limits (i.e., $x = w_2 – w_1 + \gamma$), as otherwise *x* could be increased and *p* reduced, which would lower SRO costs. It then chooses *p* so that, when customers optimally choose $z(w_2)$ and $z(w_1)$, an agent's expected income is 0 (these choices of *p* and *x* lead to $z^*(w_i) = w_i$, i = 1,2).

As was the case for the mutual SRO, $\partial P_F/\partial\gamma < 0$. The intuition for $\partial P_F/\partial\gamma < 0$ is similar to that for the mutual SRO: higher agent wealth increases the maximum penalty that can be imposed on the agent, which allows the SRO to satisfy the AIC with a smaller *p*. Note that, when $\gamma = 0$, $P_F = 1$; since the agent's payment is 0 in both states, the agent has no incentive to honestly report for any $p < 1$.

Lemma 3 implies that the fee *t* is independent of γ in equilibrium. Still, because the SRO spends less on enforcement to obtain the same *t* as γ increases, its profits rise with γ.

Figure 4 illustrates the for-profit SRO's decision. C_0 is the income level associated with the customer's participation constraint, i.e., with condition (4). A_0 is the income level associated with the agent's participation constraint, i.e., with condition (2). As before, $z^*(w_2) = z^*(w_1) + p (w_2 – w_1 + \gamma)$ is the agent's incentive compatibility constraint (1), and the vertical NLC line at w_1 reflects the no-loss constraint (3) in state 1. Figure 4 shows that, if the exchange set $p_1 = p' < P_F$, then agent's expected income would be $A_1 > A_0$, and *t* would equal $A_1 – C_0$, the value that maximizes exchange profits when $p_1 = p'$. By increasing *p* towards P_F, the exchange makes the agent's income falls towards the reservation value A_0, and *t* can be increased without violating condition (4). Thus, the equilibrium fee *t* is the vertical distance between A_0 and C_0, achieved with $p_1 = P_F$. As γ increases, x_{12} rises so that the AIR shifts upward, and the value of P_F that leads to the agent income level A_0 falls. The distance between A_0 and C_0 is independent of γ (both lines have the same slope π_2/π_1) and hence t does not change.

In sum, the for-profit exchange uses p and t to extract surplus from both customer and agent. By increasing p, the exchange reduces the agent's surplus. In doing so, however, the exchange allows the customer to attain higher levels of expected income. By increasing the transaction fee t in turn, the exchange is able to extract those rents from the customer.

D. Comparison of Ownership Structures

The principal conclusion that follows from the foregoing analysis is that the for-profit SRO devotes more resources to enforcement than does the mutual SROs. Formally, we have:

Proposition 1: If $\pi_2 (w_2 - w_1) - \pi_1 c > \alpha - w_1$, then the for-profit SRO spends at least as much on enforcement as does the mutual SRO.

Proof: See Appendix. □

The logic behind Proposition 1 is that, because the agent's compensation is equal to $E[W] - \alpha - \pi_1 t P_M$ under the mutual form, there would be no revenue for the owners of the for-profit SRO if agent compensation were not lower with a for-profit SRO. With either a for-profit or a mutual SRO, rents are earned by agents only if state 2 occurs. Hence, it follows that a for-profit SRO must reduce agent compensation in the high state (state 2), and thus the difference between the agent's state-1 income and state-2 income must be lower with a for-profit SRO. This, in turn, implies that P_F must be higher than P_M in order to induce honest reporting.

IV. Extensions

In Section III, we assumed that customers and agents are homogeneous in all respects. This simplifying assumption helps bring to light the intuition behind our results. At the same time, previous research suggests that agent heterogeneity is an important reason why exchanges have traditionally used the mutual, not-for-profit form of organization (see, e.g., Hart and Moore, 1996, 1998; Pirrong, 2000). Likewise, anecdotal evidence points to the differential impact of technological change on different types of agents as a major driver of exchange demutualization in recent years (e.g., Karmel, 2002; Aggarwal and Dahiya, 2006). In Section IV.A, we analyze the impact of cross-agent differences in wealth or future profitability (captured *via* heterogeneity in γ) on self-regulation by mutual versus demutualized exchanges.[20] In Section IV.B, we then establish the robustness of our main results to the assumption that agent income cannot be negative, i.e., that $z(w_i)$ is restricted to being less than or equal to w_i ($i=1,2$). Finally, in Section IV.C, we examine the assumption that the exchange can credibly precommit to all the aspects of its enforcement policy and discuss the potential for *ex-post* opportunistic SRO behavior with respect to enforcement activities.

A. Agent Heterogeneity

The previous section showed that when agents and their customers are homogeneous, customer welfare is the same with profit-maximizing or mutual SROs as long as $\pi_2 (w_2 - w_1) - \pi_1 c > \alpha - w_1$. Under either form of ownership, customers only receive their reservation value from trading, and there is no misreporting in equilibrium. In this subsection, we analyze the effect of agent heterogeneity on outcomes. Specifically, we present a highly stylized extension of the basic model in which misreporting sometimes occurs. We then compare the equilibrium amounts of misreporting with mutual and for-profit SROs.

In this augmented model, agents are heterogeneous with respect to the exogenous wealth parameter, γ. We posit that these differences in wealth are not observable by customers prior to deciding which agent to hire, although an agent's wealth is costlessly verifiable by the exchange during an investigation of possible wrongdoing by that agent. In the context of the financial intermediaries we analyze, one could imagine that agents own portfolios for which the positions taken are not easily determined by outsiders, and that the values of those portfolios may be subject to considerable variation. We assume that customers do know the distribution of agents' wealth. We make the simplest possible representation of such a distribution, by assuming that a fraction s of all agents have wealth γ_H, while the rest ($1-s$) have wealth γ_L, with $\gamma_L < \gamma_H$. Analogously to Section III, we assume that $(w_2 - w_1 + \gamma_L)(\pi_1 w_1 + \pi_2 w_2 - \alpha) > \pi_1 c (w_2 - w_1)$,

[20] Throughout this analysis, we maintain the assumption that customers are homogeneous. See DFH (2005) for an investigation of the importance of customer heterogeneity with respect to α in the context of a mutual exchange.

i.e., that there exists a contract under which all industry participants' individual rationality constraints can be met.

1. *The customer's decision*

Given this knowledge, customers again take the exchange's enforcement policy *{P, X}* and transaction fee *t* as given and choose a fee schedule to maximize their expected income from trading. We abstract from the possibility of offering a menu of schedules that would bring about separation of agents by wealth. Instead, we posit that a single fee schedule must be offered to all agents, and that this one-size-fits-all schedule must guarantee participation by both agent types.

In such an environment, the key change introduced by heterogeneity is that a fee schedule which induces an agent with wealth γ_L to honestly report the true state may not induce an agent with wealth γ_H to honestly report, and vice-versa.

As in Section III, the customer sets $z^*(w_1) = w_1$. It is also straightforward to show that she will choose $z(w_2)$ so that the AIC is binding for at least one type of agent. Intuitively, the trade-off in choosing $z(w_2)$ is that "high" values of $z(w_2)$ (i.e., set so that incentive-compatibility is binding for high-wealth agents) maximize the payment from high-wealth agents, but lead to misreporting by low-wealth agents – who, given limited liability, are undeterred by high penalties. Formally, we write the AIC for each type of agent:

AIC-k (10) $$w_2 - z(w_2) \geq p \max\{w_2 - w_1 - x, -\gamma_L) + (1-p)(w_2 - w_1) \qquad (k=L,H)$$

and observe that:

Lemma 4: AIC-H is binding whenever AIC-L is binding. The reverse is not true.

Proof: See Appendix. □

There are two cases to consider in the customer's choice of $z(w_2)$, depending on the SRO's choice for the penalty x. First, if the SRO has set $x \leq x_L \equiv w_2 - w_1 + \gamma_L$, then γ_L does not enter the AIC. This case is similar to the situation in Section III, with the customer choosing a schedule that binds on both agent types: $z^*(w_2) = w_1 + px$ and no misreporting in equilibrium.

The more interesting situations arise when the SRO has instead set $x > x_L \equiv w_2 - w_1 + \gamma_L$, so that the right-hand side of AIC-L equals $-p\gamma_L + (1-p)(w_2-w_1)$. In this second case, it follows from Lemma 4 that the customer must decide for which type(s) of agents the AIC constraint should bind. That is, should the fee schedule be incentive-compatible for both types of agents or for high-wealth agents only?

On the one hand, by agreeing to a sufficiently low payment from the agent in the good state, the customer can ensure that even low-wealth agents have no incentive to misreport it. Precisely, the AIC binds on low-wealth agents (and, hence, also on high-wealth agents) if the customer sets $z(w_2)$ "low," i.e., if $z(w_2) = w_1 + p\, x_L$.[21] On the other hand, given that the SRO has set $x > x_L$, the customer can get $p(x-x_L)$ more from high-wealth agents when the good state occurs by increasing $z(w_2)$ to $z(w_2) = w_1 + px$, which is the value that maximizes the payment the customer receives from high-wealth agents. At the same time, though, the perspective of paying a greater amount to the customer in the good state tempts low-wealth agents into misreporting because limited liability effectively sets an upper bound x_L on the penalty that these agents face when caught lying.[22] In other words, in the good state, the payment from a low-wealth agent to a customer who has set $z(w_2) = w_1 + px$ will be $w_1 + px_L$ (as the customer receives the amount x_L only if misreporting is detected).[23]

Because the customer is small, she does not take into account the possible impact of her choice for $Z(W)$ on the equilibrium amount of misreporting and, thereby, on the exchange's enforcement expenditures and (possibly) trading fees. Given that $z^*(w_1) = w_1$ in all cases, she will therefore choose to set $z(w_2) = w_1 + px$ and to face a positive probability of misreporting (rather than to lower $z(w_2)$ to $w_1 + px_L$ and to discourage all misreporting) as long as

$$(11) \qquad s\,(w_1 + px) + (1-s)\,[w_1 + px_L] > w_1 + px_L$$

---

[21] Any lower $z(w_2)$ would reduce the payment to the customer without changing the agent's incentive to misreport, and hence is strictly dominated by $z(w_2) = w_1 + p\,(w_2 - w_1 + \gamma_L)$.

[22] When the SRO has set $x > w_2 - w_1 + \gamma_L$, and the customer has chosen $z(w_2) = w_1 + px$, the low-wealth agent's AIC becomes: $w_2 - w_1 - px \geq -p\,\gamma_L + (1-p)\,(w_2 - w_1)$, or $w_2 - w_1 + \gamma_L > x$. Since $x > w_2 - w_1 + \gamma_L$ in this case, it follows that the low-wealth agent's AIC does not hold, and that he will misreport when state 2 occurs.

[23] The results are substantively the same if the SRO receives the revenue generated by fines rather than the customer. If the SRO receives fine revenues, then in order to induce customers to trade, the transaction fee is lower than the case considered in the text. Expected revenues are the same in both case, but the ex-post wealth of customers who deal with low-wealth (*high-wealth*) agents is lower (*higher*).

The left-hand of (11) represents the customer's expected revenue in state 2 when some misreporting takes place. It is in fact always greater than the right-hand side when $x > x_L$. Thus, given that the amount received from low-wealth agents is the same regardless of whether or not these agents lie (given misreporting is detected with probability p), it is optimal to set the fee schedule so as to maximize the revenue received from high-wealth types.[24] Put differently, when $x > x_L$, the customer chooses $z^*(w_2) = w_1 + px$ (thereby allowing some misreporting) rather than setting $z(w_2)$ sufficiently low to discourage all misreporting. In sum, across all cases:

Lemma 5: When agents are heterogeneous, customers set $z^*(w_1) = w_1$ and $z^*(w_2) = w_1 + px$.

Corollary 1: If $x > x_L$, then the optimal payment schedule $Z^*(W)$ brings about misreporting by low-wealth agents. If $x = x_L$, then $Z^*(W)$ induces truth telling by all agents.

Lemma 5 and its Corollary show that the SRO's choice of penalty, x, is key to whether all agents report honestly. In what follows, we examine in turn the levels of x chosen by mutual and for-profit SROs.

2. The mutual SRO's decision

As in Section III, the mutual SRO selects p and x to maximize the incomes of its member agents, subject to ensuring customer participation. However, while in Section III all agents were homogenous and member income was unequivocally defined, in this Section the heterogeneity of members raises the question of whose income the SRO should maximize (e.g., is the median agent a low-wealth or a high-wealth individual?). As emphasized in the extant literature (e.g., Hart and Moore; 1998; Pirrong, 2000), agent heterogeneity may impact SRO policy. In the case at hand, if different agents prefer different choices of enforcement variables, then the values of p and x chosen by the mutual exchange could depend on whose wealth is being maximized.

Regardless of possible cross-exchange differences, we first note that x will take on one of only two values: either $x = x_L \equiv w_2 - w_1 + \gamma_L$ or $x = x_H \equiv w_2 - w_1 + \gamma_H$.

---

[24] Because the customer takes the exchange's policies as exogenous, she focuses on the marginal impact of allowing misreporting on her revenues (11) rather than her net income and ignores the fact that, if misreporting takes place in equilibrium, then the exchange may change its probability of investigation and the magnitude of the fees it charges.

Lemma 6: Both high- and low-wealth agents would choose $x_L \leq x \leq x_H$.

Proof: See Appendix. □

The logic behind the lower bound is similar to that in Section III. Specifically, as long as $x < x_L$, an increase in x accompanied by a reduction in p that leaves consumer's net income unchanged will raise agent income, because it allows for a reduction in enforcement costs without inducing misreporting. The upper bound for x is a simple consequence of the limited liability condition for high-wealth agents: increases in x beyond x_H have no effect on any agent's incentives to misreport, since the most they can lose if their misreporting is detected is x_H. The next Lemma shows how the SRO's choice of p depends on the value of x chosen.

Lemma 7: If $x = x_L$, then the mutual SRO chooses $P_{M,1} = \dfrac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma_L) - \pi_1 c}$

If $x = x_H$, then $P_{M,2} = \dfrac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma_L) + \pi_2 s(\gamma_H - \gamma_L) - [\pi_1 + \pi_2(1-s)]c} < P_{M,1}$

In the latter case, $\partial p / \partial s < 0$.

Proof: See Appendix. □

We know from Corollary 1 that, if the SRO sets $x > x_L$, the result is misreporting by low-wealth agents. If x is such that misreporting does occur (e.g., if $x = x_H$), then to ensure customer participation the SRO must adjust p upward as the proportion s of low-wealth individuals rises.

Setting $x > x_L$ increases low-wealth agents' expected incomes but can reduce high-wealth agents' incomes. Hence, whether misreporting is allowed to take place in equilibrium can depend on which type of agent's income the mutual SRO maximizes.

Proposition 2: Let $x_L \equiv w_2 - w_1 + \gamma_L$ and $x_H \equiv w_2 - w_1 + \gamma_H$;

(a) If the mutual SRO maximizes the expected income of low-wealth agents, and if parameter values are such that both the AIR for high-wealth agents and the CIR can be met (precisely, if $w_1 + (\frac{w_2 - w_1}{x_H})\pi_2(x_L + s(\gamma_H - \gamma_L)) - \alpha - \pi_1(\frac{w_2 - w_1}{x_H})c) > 0$), then the SRO optimally sets $x = x_H > x_L$ and misreporting occurs in equilibrium.

(b) If the SRO maximizes the expected income of high-wealth agents, then the penalty for misreporting x is either x_L *or* x_H. The SRO sets $x = x_H$ (allowing misreporting) if the following conditions are met:

i. $\pi_2(1\text{-}s)\, x_L \leq [\pi_1 + \pi_2(1\text{-}s)]c$ (12)

ii. parameter values are such that both the AIR for high-wealth agents and the CIR can be met

iii. parameter values are such that $x_H\, P_{M,2} < x_L\, P_{M,1}$, or

$$c[(\gamma_H - \gamma_L)\pi_1 - (1-s)x_L\pi_2] > (1-s)(\gamma_H - \gamma_L)\pi_2 x_L \quad (**)$$

If these three conditions are not met, then high-wealth agents strictly prefer $x = x_L < x_H$. Hence, a mutual SRO that maximizes the expected income of high-wealth agents will prefer the no misreporting equilibrium unless all three conditions are met.

Proof: See Appendix. □

Proposition 2 implies that the general proposition that an enforcement agent will choose maximal fines (which allow for minimal enforcement expenditures for a given expected penalty) need not always apply to SROs with heterogeneous traders. Low-wealth agents will always prefer a maximal level of *x*, which induces misreporting, whereas high-wealth individuals may or may not prefer to have misreporting in equilibrium. To see the intuition of why high-wealth individuals may not prefer higher *x*, note that increasing *x* and reducing *p* has two effects, which are reflected in the two terms in (12). First, conditional on $x > x_L$ (so that misreporting occurs), expected customer income falls (and expected income for low-wealth agents is higher) as *p* declines (holding *px* fixed). This is because higher x does not result in higher payments from low-wealth agents once *x* equals x_L. This is the term on the left-hand side of (12). This effect means that *px* must rise as *x* falls in order to leave customers with incomes of at least *a*. That is, customers must be offered compensation for lower expected penalties, so that *px* must increase. This in turn means that high-wealth agents are effectively transferring income to low-wealth ones. On the other hand, lower *p* and higher *x* reduces inspection costs. This is the term on the right-hand side of (12). When the enforcement costs are low, the second effect is small, and high-wealth individuals prefer $x = x_L$, which eliminates the transfer to low-wealth agents.

As in previous literature, we find that heterogeneity of agents leads to potentially important disagreements with respect to policy. To some extent, that disagreement becomes less important as heterogeneity declines. Specifically, as s goes to 1, the term on the left-hand side of (12) goes to 0, which means that all agents prefer some misreporting in equilibrium. Intuitively, if there are few low-wealth agents, then misreporting is not very costly to high-wealth agents, and the costs savings from reducing the frequency of inspections becomes relatively more important. At the same time, when s is close to 1, it is more plausible that the SRO will attempt to maximize the income of the majority of agents, who are high-wealth. Conversely, as s goes to 0, condition (12) will not hold and high-wealth agents will prefer $x = x_L$, but the SRO is likely to ignore these agents, as low-wealth agents will constitute the majority of SRO members.[25]

3. *For-Profit SROs*

As was the case when agents are homogeneous, the for-profit SRO seeks to maximize its profits (transaction fee minus investigation costs), subject to the participation constraints of agents and customers. In addition, the assumption that $w_1 < \alpha$ means that the SRO must choose parameters consistent with the agents' truth-telling constraints (for at least some of the agents). Once again, the SRO seeks to maximize

$$\max_{t,\,p,\,x} \quad t - \pi_1 c p$$

subject to

AIR (13) $\pi_2 z^*(w_2) + \pi_1 z^*(w_1) \leq \pi_2 w_2 + \pi_1 w_1$

AIC*-k (14) $w_2 - z^*(w_2) \geq p_1[\max\{w_2 - w_1 - x, -\gamma_k\}] + (1-p_1)(w_2 - w_1)$ (k=L,H)

CIR (15a) $w_1 + \pi_2 p\,[sx + (1-s)(w_2 - w_1 + \gamma_L)] - t \geq \alpha$
(if 14 only holds for high-wealth agents)

(15b) $w_1 + \pi_2 px - t \geq \alpha$
(if 14 holds for all agents)

The fee schedule *Z*(W)* reflects the customer's optimizing behavior detailed in Section IV.A.1. Note that we include two potential AICs, corresponding to the case where the AIC is

[25] As discussed in connection with Figure 6 below, if s is sufficiently small so that many all agents misreport the state, conditions necessary for the AIR and CIR to hold are not likely to be met if misreporting occurs.

binding on high-wealth agents only, and where it is binding on all agents; in the pooling equilibrium the AIC cannot be binding on only the low-wealth agents, and as such, misreporting can exist in equilibrium.

Proposition 3: If parameter values are such that both equations (13) and (15a) hold

[i.e., $\frac{w_2 - w_1}{x_H} > \frac{\alpha - w_1}{\pi_2[x_L + s(\gamma_H - \gamma_L)] - [\pi_1 + \pi_2(1-s)]c}$],

then the for-profit SRO will choose $x = x_H$, so that misreporting will occur in equilibrium if both

(16) $\pi_2(1-s)x_L \leq [\pi_1 + (1\text{-}s)\pi_2]\, c$, and

$$c[(\gamma_H - \gamma_L)\pi_1 - (1-s)x_L\pi_2] > (1-s)(\gamma_H - \gamma_L)\pi_2 x_L \quad (**)$$

Otherwise, there will be no misreporting in equilibrium.

Proof: See Appendix. □

Focusing on the set of parameters for which equations (13) and (15a) hold, Proposition 3 is similar to Proposition 2 in that it shows that misreporting can occur in equilibrium as long as the extent of misreporting is limited. One conclusion from this analysis is that the likelihood of misreporting is lower with a for-profit SRO in the following sense: For any set of parameters for which the mutual SRO will prevent all misreporting, the for-profit SRO will also prevent all misreporting. However, there are parameter values for which the for-profit SRO will prevent misreporting, while the mutual SRO will allow misreporting. This follows from the fact that (if ** holds) $\pi_2(1-s)x_L > [\pi_1 + (1\text{-}s)\pi_2]c$ is a sufficient condition for no misreporting in the for-profit case, but only a necessary (and not sufficient) condition for the mutual exchange.

At first glance, one might view the extent of misreporting in our model as uninteresting, in that customers get the same *ex-ante* utility whether or not misreporting occurs. However, if some misreporting does occur in equilibrium, it means that customers are *ex-post* randomly made better- or worse-off by virtue of the misreporting. Conceivably, this randomness is the kind of outcome that trade practice rules are designed to eliminate.

The for-profit SRO's determination of p, x and t when agents are heterogeneous is similar to its calculation when agents are homogeneous. In both cases, the SRO sets p to extract surplus from agents (in the heterogeneous case, the high-wealth agent), and then sets t to extract surplus from customers. When conditions (16) and ** are satisfied and the SRO chooses to allow some misreporting, the SRO sets $x = x_H$ and chooses p to just satisfy the high-wealth agents' AIR, or

$$P_{F,2} = \frac{w_2 - w_1}{w_2 - w_1 + \gamma_H}$$

and sets t to satisfy the CIR , or

$$t = w_1 + P_{F,2}\pi_2(w_2 - w_1 + \gamma_L + s(\gamma_H - \gamma_L)) - \alpha.$$

As long as t is greater than the enforcement costs of $P_{F,2}\ c\ \pi_1$ and condition (16) is satisfied, this combination of P_F and t will maximize SRO profit. It can be readily shown that $t > P_{F,2}\ c\ \pi_1$ is equivalent to $P_{F,2} > P_{M,2}$.

If condition (16) is not satisfied, then the SRO sets $x = x_L$ and sets P_F equal to

$$P_{F,1} = \frac{w_2 - w_1}{w_2 - w_1 + \gamma_L}$$

to prevent misreporting and satisfy all agents' AIR, and sets

$$t = w_1 + \pi_2 P_{F,1}(w_2 - w_1 + \gamma_L) - \alpha = w_1 + \pi_2(w_2 - w_1) - \alpha$$

To see the intuition for t in this case, note that the total gains from the customer's trade is $\pi_2\ w_2 + \pi_1\ w_1 - \alpha - P_{F,1}\ c\ \pi_1 = \pi_2\ (w_2 - w_1) + w_1 - \alpha - P_{F,1}\ c\ \pi_1$. Hence, this combination of $P_{F,1}$ and t extracts all of the agent and customer surplus, and therefore maximizes SRO profit as long as these gains are positive.

For parameter values for which gains from trade exist, the for-profit SRO enforces the trade-practice rules more rigorously for any given fine. The logic is the same as in the homogeneous agent case. For any given x, the for-profit SRO wants to choose a detection probability at which the AIR is just binding. That is necessarily at least as high as the detection probability that maximizes agent profits (which is the goal of the mutual SRO).

When the SRO chooses *x* such that misreporting occurs, the percentage of low-wealth agents (*s*) affects outcomes in a different way for the mutual and for-profit SRO. As indicated in Lemma 7, the mutual SRO has to increase *p* as *s* falls, in order to induce customer participation. In contrast, the for-profit SRO reduces *t* when *s* falls in order to induce customer participation, but *p* is unaffected. In the former case, high-wealth agents are worse off as *s* rises, while in the latter, the SRO's owners are made worse off.

Figure 6 portrays the likelihood of detection in the four cases.[26] Several aspects of the equilibria with heterogeneous agents can be observed in the figure. First, as in the homogeneous case, the detection probability is at least as high with a for-profit SRO as with a mutual SRO for any given level of fines. This follows directly from the fact that the for-profit SRO sets *p* so that the agent's AIR is just satisfied. Second, the equilibrium in which misreporting occurs requires *s* to be above some minimum. In terms of Figure 6, if *s* is below the level at which $P_{F,2}$ is equal to $P_{M,2}$, then high-wealth agents would lose money (and hence not participate) in the misreporting equilibrium (*s* less than 0.63 in the figure). Third, there is a range of values for *s* for which the misreporting equilibrium exists, but the high-wealth agents (in the mutual case), and the owners in the for-profit case prefer the no-misreporting equilibrium (*s* between 0.63 and 0.89). However, for values of *s* in this range, low-wealth agents prefer the misreporting equilibrium.

B. Equilibrium with the potential for non-negative returns to truth-telling agents

The analysis in Section III imposes the restriction that customers must choose fee schedules such that the agent's earnings in both states be at least 0. This restriction seems consistent with actual practice. That is, in actual practice, it does not appear that agents are required to give customers a payment in excess of the agent's actual receipts when the agent honestly reports that receipts were "low." On the other hand, it may be that agents lose money when they report a low receipt, in that agents face positive trading costs which are not recovered in the low state. In any case, in this section we wish to consider how the equilibrium changes when we remove this restriction; that is, if the no-loss constraint in (3) is replaced with the constraint that the payment to the customer in any state can never be greater than the actual receipt of the agent plus the agent's wealth (the no-bankruptcy constraint). This change means

[26] The parameters used in Figure 6 are $\alpha = 1.5$, $\gamma_H = 2$, $\gamma_L = 1.25$, $\pi_1 = .5$, $c = .25$, $w_2 = 2.2$, $w_1 = 1$.

that the constraint in equation (2) is no longer redundant; that is, agents must earn non-negative profits in expectation, but can lose money in any state, even when they give an honest report. Allowing for this possibility allows the agents/SRO to earn more than in the case evaluated in Section III, but the principal conclusions remain unchanged.

To analyze this case, we return to the assumption that agents are homogeneous with respect to γ and then revisit the behavior of customers and SROs, replacing condition (3) with

(3') $\quad w_i - z(w_i) - x_{ij} \geq -\gamma,\ i = 1,2$

Agent wealth, γ, plays a somewhat different role in this model. If $\gamma > 0$, consumers can offer a contract to the agent that yields the agent negative revenues when he announces state 1. Specifically, analogously to Section III, the customer's profit-maximizing choice of $z(w_1)$ is the highest $z(w_1)$ consistent with (3'), which is $z(w_1) = w_1 + \gamma$ in this model.[27] The agent's profits are then equal to $-\gamma$ when he announces the true state is state 1. As above, this not only maximizes the customer's payment when state 1 occurs, but also allows for higher $z(w_2)$. Given $z(w_1) = w_1 + \gamma$, the customer will then choose a $z(w_2)$ so that the incentive compatibility constraint is binding, which means that $w_2 - z(w_2) = p[\max\{w_2 - z(w_1) - x, -\gamma\}] + (1-p)(w_2 - z(w_1))$. Hence, for $x \leq w_2 - w_1$, the binding AIC is $z(w_2) = z(w_1) + px$.

Figure 7 portrays the consumer decision in this alternative model. The main difference between the decision here and that portrayed in Figure 1 is that changes in γ shift the constraint in equation (3'), rather than indirectly changing the AIC. That is, in this model, increasing γ shifts the vertical line representing equation (3') to the right, allowing higher $z(w_1)$, holding p and x fixed. In this model, the customer's optimum either occurs at the intersection of the AIC and the NBC (as portrayed in Figure 7), or at the intersection of the AIC and the AIR, depending on the value of γ. In particular, as discussed below, if γ is sufficiently large, then the latter intersection will be the consumer's optimum in equilibrium.

Given this behavior by customers, the mutual SRO once again chooses values for p and x to maximize the agent's income. As our earlier analysis implies, the SRO will choose a maximal

[27] This generalizes the conclusion in DFH, who only consider the case of $\gamma = 0$.

value for x, which in this case is $x = w_2 - w_1$. The logic here again is that if x were less than $w_2 - w_1$, x could be increased and p decreased without changing the agent or the customer's income, but saving the customer enforcement costs. The reason x is lower in this model than in the model in Section III is that agents pay γ whenever they announce state 1 has occurred, which reduces the maximum penalty the SRO can assess when the agent is discovered misreporting the true state. This means that the $z(w_2)$ that results in the incentive compatibility constraint binding is

$$z(w_2) = p(w_1 + x) + (1\text{-}p)(w_1) = w_1 + \gamma + p(w_2 - w_1).$$

As long as $\gamma \leq \alpha - w_1$ then the SRO will choose p so that the CIR is binding,[28] or

$$\pi_2 z(w_2) + \pi_1 z(w_1) - t = \alpha.$$

Solving these two equations for p, and recalling that $t = \pi_1 p c$ for the mutual SRO yields

$$p_1 = p'_M = \frac{\alpha - w_1 - \gamma}{\pi_2(w_2 - w_1) - \pi_1 c}$$

Note that as long as $\gamma \leq \alpha - w_1$, P'_M is decreasing in γ, as was the case for model presented in Section III. Here, higher agent wealth allows the customer to create greater incentives for truth-telling via the fee schedule, which in turn allows the SRO to achieve the same degree of deterrence with a lower probability of detection.

For $\gamma \geq \alpha - w_1$, P'_M will equal zero. In contrast to the SRO's decision in the earlier model, in this model the CIR can be satisfied even when $p = 0$. Customers receive $z(w_1) = \gamma + w_1$ when the agent claims the state is state 1. When $p = 0$, the AIC implies $z(w_2) = z(w_1) = \gamma + w_1$. Hence if $\gamma + w_1 \geq \alpha$, this payment is greater than the customer's reservation value, and the SRO will indeed set $P'_M = 0$. Once γ is greater than $\alpha - w_1$, higher γ transfers income from the agent to the customer. However, for $\gamma \geq \pi_2 (w_2 - w_1) \geq \alpha - w_1$, the AIR is binding, and hence higher γ has no additional effect on fees.

For the for-profit SRO, the optimization problem is once again to set p and x to extract all agent surplus, and then set t to extract all of the customer's surplus. For the reasons described

[28] As in the earlier model, we also assume $\pi_2 (w_2 - w_1) - \pi_1 c \geq \alpha - (w_1 + \gamma)$.

above, the SRO will set x at its maximum consist with $x < w_2 - z(w_1) + \gamma$, or $x = w_2 - w_1$. As in Section III, increasing p allows the SRO to increase t. Increasing p is profitable as long as $\partial t/\partial p > \pi_1 c$ (i.e, the change in t resulting from the higher p exceeds the marginal cost of raising p). When $z(w_1) = w_1 + \gamma$ and $x = w_2 - w_1$ so that $z(w_2) = w_1 + \gamma + p(w_2 - w_1)$, the CIR becomes

(8') $p(w_2 - w_1) = (\alpha + t - \gamma - w_1/\pi_2$

This implies that $\partial t/\partial p = \pi_2 (w_2 - w_1)$, and since $\pi_2 (w_2 - w_1) > \pi_1 c$ (which is the range of values for which $P'_M > 0$), the for-profit SRO will continue to increase p until condition (2) is binding.

Once p is sufficiently high that (2) is binding, any further increase in p will require $z^*(w_1)$ to fall with p in order to induce the agent to participate. At that point, higher p only transfers income between states for both the agent and the customer; it does not allow the SRO to increase revenue. As such, since $c > 0$, increases in p are no longer profitable. At the point that the CIR is binding, $\pi_2 (w_1 + \gamma + p(w_2 - w_1)) + \pi_1 (w_1 + \gamma) = \pi_2 w_2 + \pi_1 w_1$, or

$$p_1 = p_F = \frac{\pi_2 (w_2 - w_1) - \gamma}{\pi_2 (w_2 - w_1)}$$

and

$$t = w_1 + \pi_2 (w_2 - w_1) - \alpha$$

As with the mutual SRO, the probability of detection is a function of γ. The probability of detection upon a report of a poor outcome by the agent is $P'_F = 1$ if $\gamma = 0$, while P'_F is between 0 and 1 for $\gamma \,\varepsilon\, (0, \pi_2 (w_2 - w_1))$. As was the case for the mutual SRO, we have $\partial P'_F/\partial \gamma < 0$. The intuition for $\partial P'_F/\partial \gamma < 0$ is again that higher agent wealth increases the payment the customer receives and reduces the payment to the agent in state 1. This allows the SRO to induce honest reporting with a lower frequency of inspection. Analogously with P'_M, P'_F goes to 0 for γ sufficiently large. And, as was the case for the mutual SRO, beyond some point (here $\gamma = \pi_2[w_2 - w_1])$ higher agent wealth transfers income from the SRO owners to the customers.

As in the analysis in Section III, $P'_F \geq P'_M$, if $\pi_2 (w_2 - w_1) - \pi_1 c > \alpha - w_1$ (recalling that, if the inequality fails, there is no gain to trading when $\gamma = 0$).

Proposition 4: If $\pi_2 (w_2 - w_1) - \pi_1 c > \alpha - w_1$ then $P'_F \geq P'_M$, and $P'_F > P'_M$ for $\gamma < \pi_2 (w_2 - w_1)$.

Proof: See Appendix.

A second and related difference between mutual and for-profit SRO is that, if the SRO could choose its γ, the γ chosen by the for-profit SRO would be larger. In other words, *ceteris paribus*, a for-profit SR exchange strictly prefers wealthier agents (Futures Commission Merchants or FCMs in futures markets) than does a mutual SR exchange. There are two possible interpretations of this finding. (i) First, as noted above, there has been a shift from mutual to for-profit SROs over the past decade. One potential reason for that shift is that exogenous changes have occurred in technology or competition that have increased γ or α (increased competition can be thought of as an increase in α). (ii) Second, if one treats γ as a choice variable for the SRO, then the implication is that the for-profit SRO will choose a larger γ. Whether this is interpreted in terms of increased agent wealth or increased agent productivity is something we turn to below.

C. Observability of enforcement parameters and the role of government

Like the previous CSV literature, our analysis assumes that the enforcement parameters were observable to all agents. This is an important feature of the model, in that the choice of these parameters by the SRO affects the fee schedule set by the customers.

One might be concerned that the assumptions that p and x are known to customers may deviate from reality in a significant way if (i) it is plausible that p and x may not be observable in reality, and (ii) SROs could have incentives to deviate from the p and x derived above, once the fee schedules are set.

Whereas it would seem that the levels of fines chosen by an SRO are fairly transparent, it also seems plausible that customers may not be able to observe the true likelihood of detection chosen by the SRO. In the case of a mutual SRO, reducing the detection probability p_1 below P_M can both increase agent income and reduce SRO costs. That is, once the fee schedule is set and customers start trading, the SRO can communicate its intention to lower p to the agents, with the results that the agents will all misreport (and earn more) and the SRO will save on enforcement expenditures. By contrast, once the fee schedule is set, the incentive for the for-profit SRO to reduce *p* only comes about through the potential to save enforcement costs. Unlike the mutual SRO, the for-profit SRO does not increase its revenues by reducing *p*. If anything, this difference in the net benefit from reducing *p* reinforces the earlier conclusion that a for-profit SRO will have stricter enforcement policies.

Of course, several factors also mitigate an SRO's incentives to reduce p_1. First, for both kinds of SROs, customers may rationally anticipate the behavior of the SRO. Hence, if it is indeed profitable to cut *p* once fee schedules are set and customers decide to trade, then the fee schedules will reflect that potential. In turn, this response by customers will reduce agent and/or SRO profit. Thus, SROs will attempt to establish a reputation, whereby it is costly to them to deviate from the probabilities derived in the static environment described above.

Second, for mutual SROs, saving costs by reducing *p* may not be to the agents' benefit, even in the short run (although the benefit of higher expected agent revenue from lower p is unambiguous). This is because the mutual, non-profit organization form can be viewed as a commitment to a certain level of enforcement expenditures. That is, with the mutual form, transaction fees are set equal to enforcement expenditures, so that members may receive no direct benefit from a reduction in enforcement costs.[29] And, in the case of for-profit SROs, the enforcement budget is *ex-post* observable (e.g., by reading the company's annual report) and thus deviations will harm the SRO's reputation.

Formal considerations of the potential for such opportunistic behavior on the part of SROs are beyond the scope of this paper. These considerations do have some implications for the role of regulatory authorities towards enforcement of rules against misreporting (such as trade practice rules in the futures context), however. Specifically, DFH suggest that the threat of duplicative government enforcement can induce mutual SROs to increase the probability of detection. This can increase customer welfare (and in a model in which α is heterogeneous, increase social welfare). In the case of for-profit SROs, that policy is less likely to be welfare-enhancing, because increases in *p* beyond P_F do not increase consumer welfare. The potential

[29] Perhaps the concern expressed by various regulatory agencies that for-profit SROs will reduce enforcement expenditures reflects the differential effect of cost savings from reduced enforcement expenditures for the two kinds of SROs.

for opportunism suggests that government policy might better be aimed at insuring that SROs provide the promised level of enforcement. As a practical matter, this intervention might consist of reviewing the SRO's budgeted allocation of resources for enforcement, and monitoring the level of actual expenditure, to insure against opportunistic behavior. In addition, since profit maximizing (absent opportunism) increases enforcement expenditures, policies that lead SRO closer to the objective of maximizing profits result in greater enforcement..

V. Conclusion

This paper analyzes how an SRO might make use of penalties and incentives to influence contracts between financial investors and their agents. Of particular interest is how the ownership structure of the SRO (for-profit *vs*. mutual) influences the enforcement strategy of the SRO. Broadly speaking, we find that a for-profit SRO uses more "sticks" and fewer "carrots" to provide incentives for agents to report honestly to their clients. That is, the for-profit SRO spends more on enforcement than a mutual SRO. We also find that trade practice violations are more likely (that is, occur for additional parameter values) when the SRO is a mutual exchange.

We have tried to capture the essence of alternative ownership structures in a tractable environment by limiting the strategy space. Nevertheless, we think that the basic effects we find should carry over to richer environments. For example, because we have a static environment, we limit the analysis to monetary fines. In a dynamic environment, the SRO would have the ability to suspend the trading privileges of an agent that violates a trade practice rule. In that case, the agent's wealth parameter γ would reflect future trading profits, and could differ between a for-profit and a mutual SRO. In particular, since agents earn less when the SRO is for-profit (and, hence, have less to lose from a trading ban), a for-profit SRO – which we find to be a stricter enforcer than its mutual counterpart – will have to rely on an *even* greater likelihood of inspection to insure against misreporting.

The role of agent heterogeneity is also simplified in our model, as we focus on only one aspect of heterogeneity (wealth). Previous work (e.g., Pirrong, 2000; Hart and Moore, 1998) has emphasized the role of heterogeneity in the choice of ownership structure. To the extent that heterogeneity influences the choice of structure, the analysis of the enforcement of trade practice regulations would have to incorporate the endogeneity of the form of ownership structure.

References.

Aggarwal, Reena and Sandeep Dahiya, 2006, "Demutualization and Public Offerings of Financial Exchanges," *Journal of Applied Corporate Finance*, 18 (3), pp. 96-106.

Boyd, John H. and Bruce D. Smith, 1994, "How Good Are Standard Debt Contracts: Stochastic versus Nonstochastic Monitoring in a Costly State Verification Environment," *Journal of Business*, 67 (4), pp. 539-61.

CFTC, 2005. *Self-Regulation and Self-Regulatory Organizations in the Futures Industry* (Federal Register Notice, 11/25/2005), U.S. Commodity Futures Trading Commission.

Cummins, Douglas and Sofia Johan, 2006, "Market Surveillance Around the World," Working Paper, Rensselaer Polytechnic Institute and Tilburg University, June.

DeMarzo, Peter M., Michael J. Fishman, and Kathleen M. Hagerty, 2005, "Self-Regulation and Government Oversight," *Review of Economic Studies*, 72 (3), pp. 687-706.

Elliott, Jennifer, 2002, "Demutualization of Securities Exchanges: A Regulatory Perspective," *IMF Working Paper 02/119*, International Monetary Fund, Washington, DC.

Gale, Douglas and Martin Hellwig, 1985, "Incentive-Compatible Debt Contracts: The One-Period Problem," *Review of Economic Studies*, 52 (4), pp. 647-63.

Hart, Oliver and John Moore, 1996, "The Governance of Exchanges: Members' Ccooperatives versus Outside Ownership," *Oxford Review of Economic Policy*, 12 (4), pp. 53-69.

Hart, Oliver and John Moore, 1998, "Cooperatives vs. Outside Ownership," NBER Working Paper 6421, National Bureau of Economic Research, February.

IMF, 2005. *Financial Sector Assessment: A Handbook*. International Monetary Fund, Washington, DC.

IOSCO Technical Committee, 2006, *Regulatory Issues Arising from Exchange Evolution*. International Organization of Securities Commissions, Madrid, Spain.

Johnson, Philip M. and Thomas L. Hazen, 1999, *Commodities Regulation* (3rd ed.) Aspen Law and Business, Gaithersburg, MD.

Karmel, Roberta S., 2002, "Turning Seats Into Shares: Implications of Demutualization for the Regulation of Stock Exchanges," *Hastings Law Journal*, 53 (2), pp. 367-430.

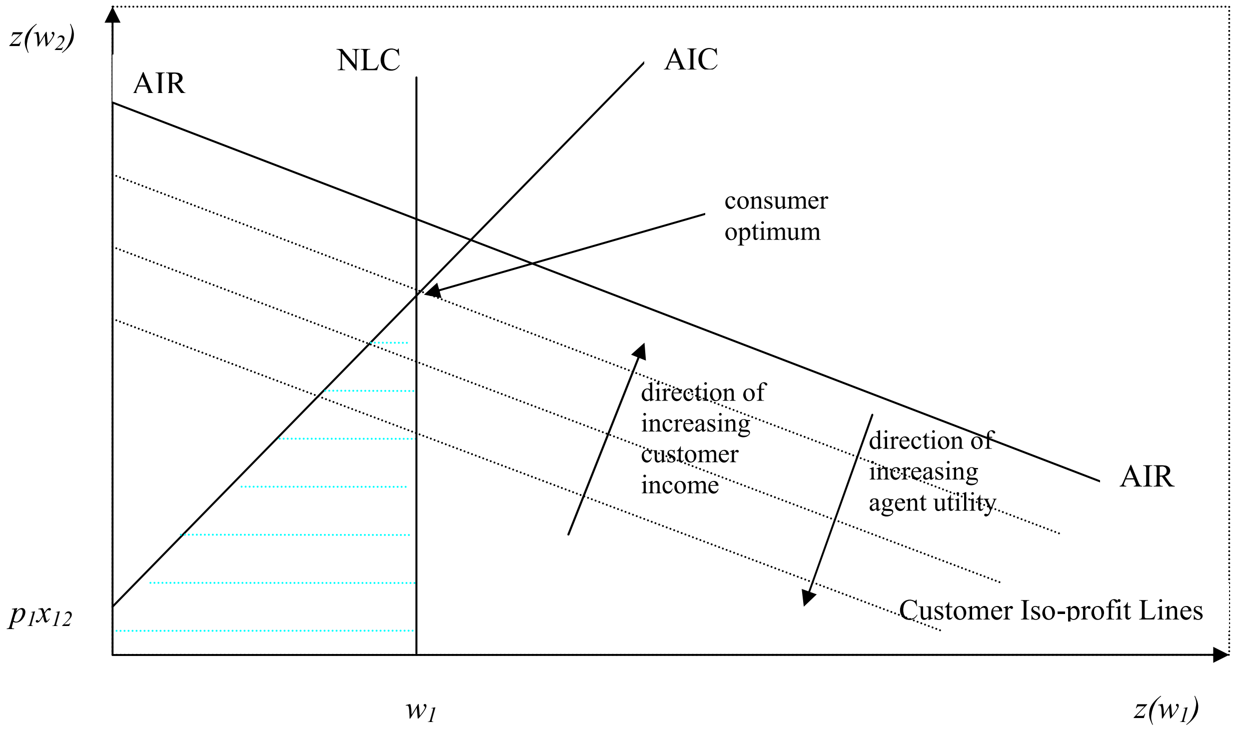
Pirrong, Craig, 1995, "The Self-Regulation of Commodity Exchanges: The Case of Market Manipulation," *Journal of Law and Economics*, 38 (1), pp. 141-206.

Pirrong, Craig, 2000, "A Theory of Financial Exchange Organization," *Journal of Law and Economics*, 43 (2), pp. 437-71.

SEC, 2004. U.S. Securities and Exchange Commission Concept Release Concerning Self-Regulation, *Federal Register* v69 n235, 71256-82 (17 CFR PART 240, Release No. 34-50700; File No. S7-40-04, RIN 3235-AJ36). Available online: http://www.sec.gov/rules/concept/34-50700.pdf

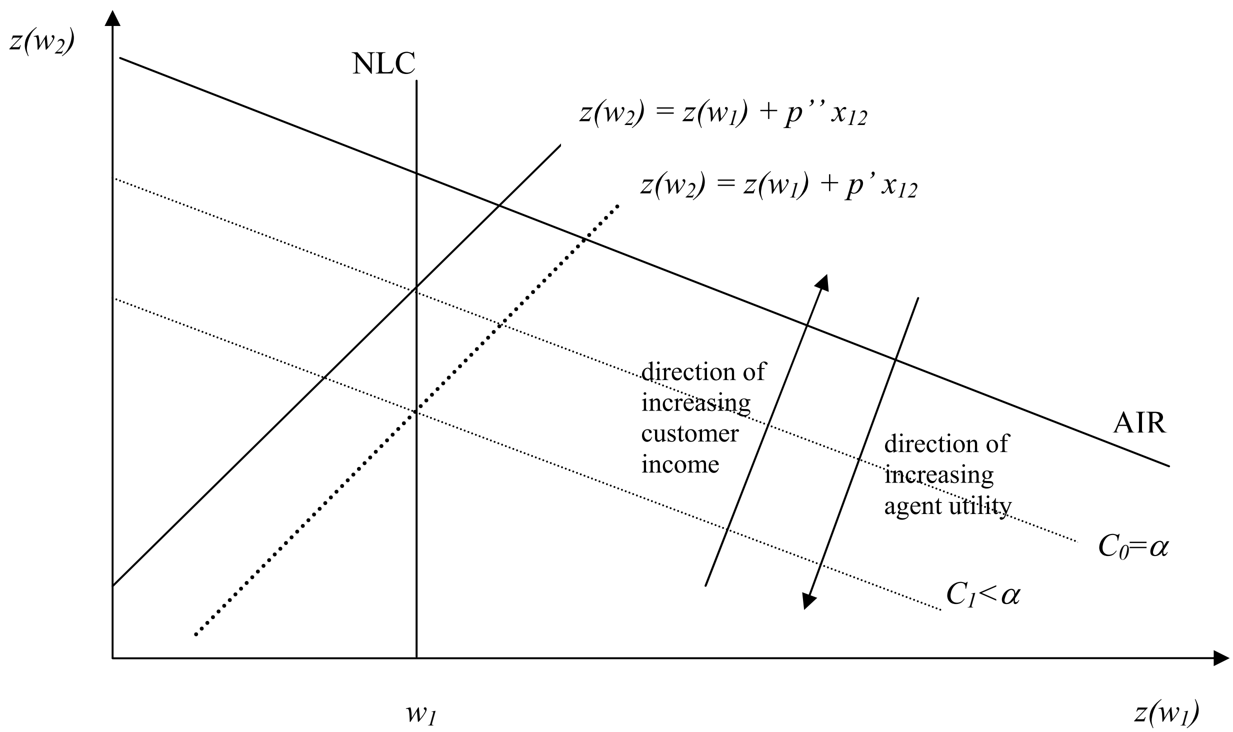
Townsend, Robert, 1979, Optimal Contracts and Competitive Markets with Costly State Verification, *Journal of Economic Theory* v21, n2: 265-93

Figure 1: The Customer's Problem



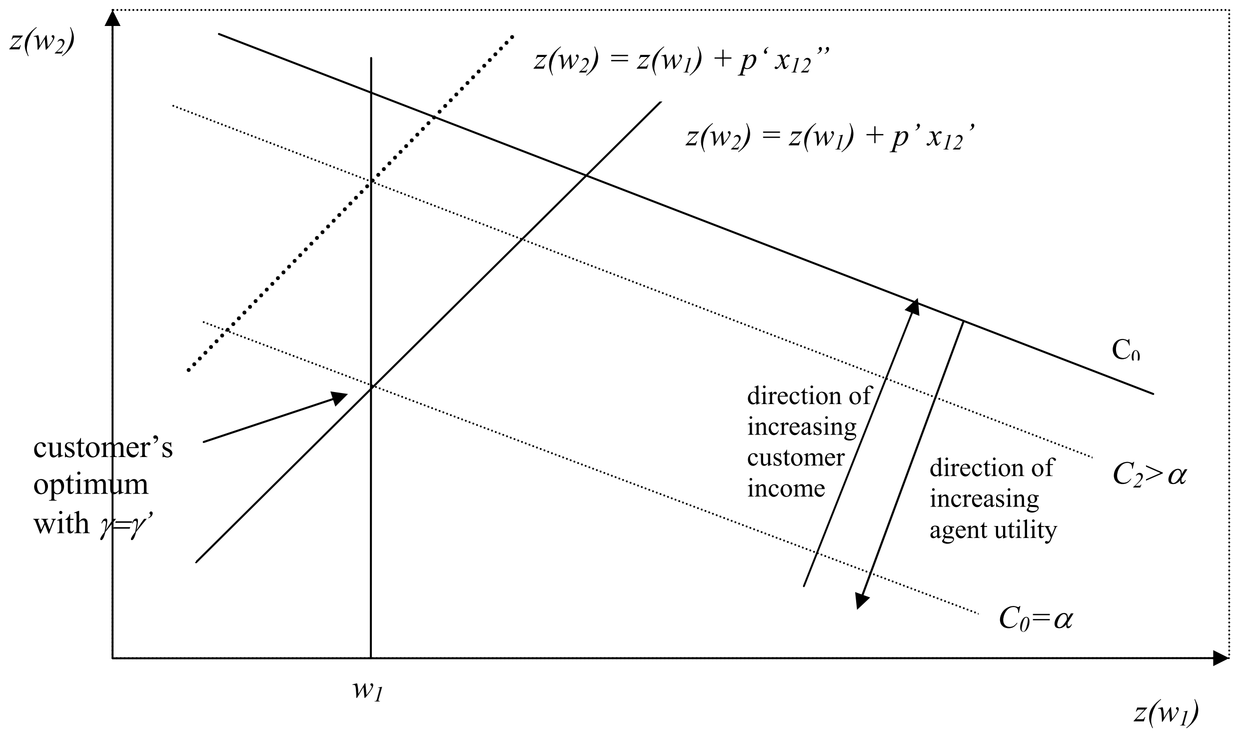
Notes: Figure 1 provides a graphical illustration of the constraints that the customer faces when choosing the agent's compensation schedule $\{w_1\text{-}z(w_1), w_2\text{-}z(w_2)\}$. The agent's incentive compatibility constraint (1) is depicted by the 45-degree AIC line $z(w_2) = z(w_1) + p_1 x_{12}$. To ensure truth-telling, $z(w_2)$ must lie on or below this line. The agent's individual rationality constraint (2) is depicted by the downward-sloping AIR line $z(w_2) = w_2 - \pi_1 w_1/\pi_2 - \pi_1 z(w_1)/\pi_2$. To get participation, $z(w_2)$ must lie on or below this line. Finally, the no-loss condition (3) is depicted for $i = 1$ by the vertical line NLC at $z(w_2) = w_1$. To respect limited liability, $z(w_2)$ must lie to the left of this line. The striped fived-sided area depicts the combinations of $z(w_1)$ and $z(w_2)$ that simultaneously meet all three constraints (AIC, AIR and NLC).

Figure 2: The Mutual SRO's Choice of Investigation Probability



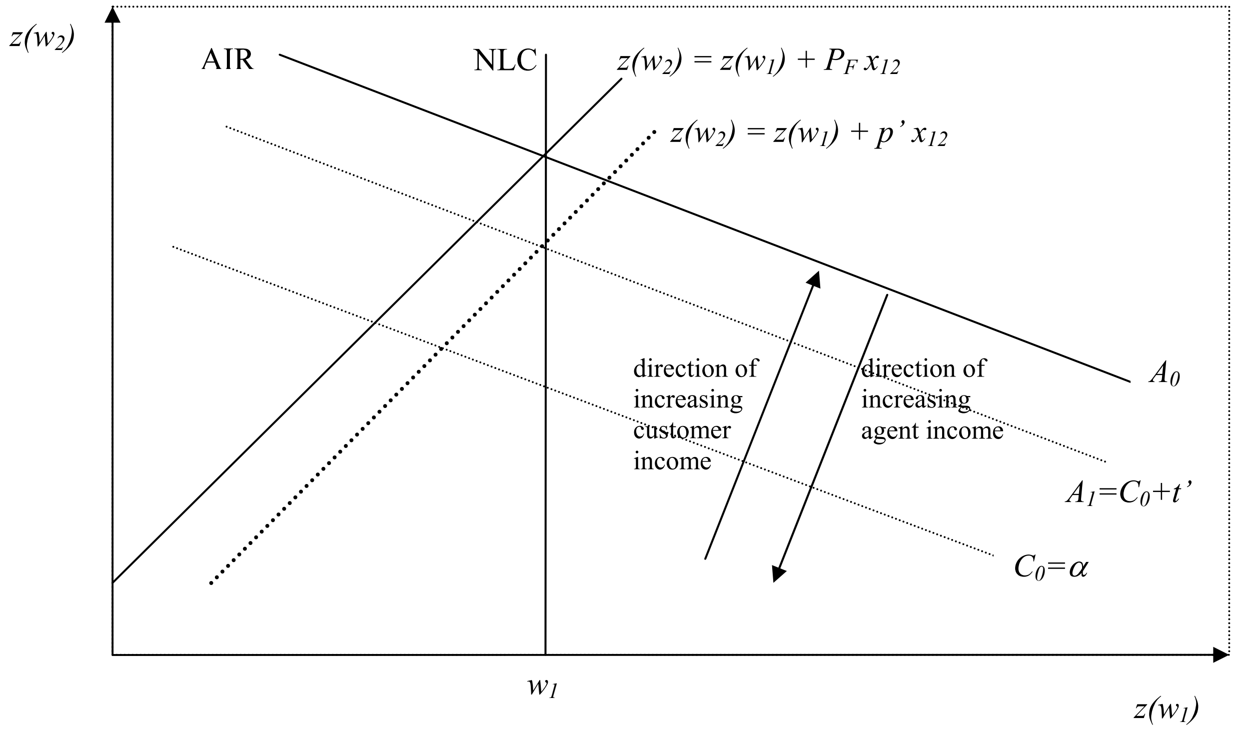
Notes: Figure 2 illustrates how the mutual SRO's choice of investigation probability, p_1, affects the agent's and the customer's respective expected utilities. *Ceteris paribus*, by increasing p_1 from p' to p'', the SRO shifts the AIC upward and increases the customer's expected income from $C_1 < \alpha$ to $C_0 = \alpha$.

Figure 3: Effect of Agent Wealth on Mutual SRO Policy

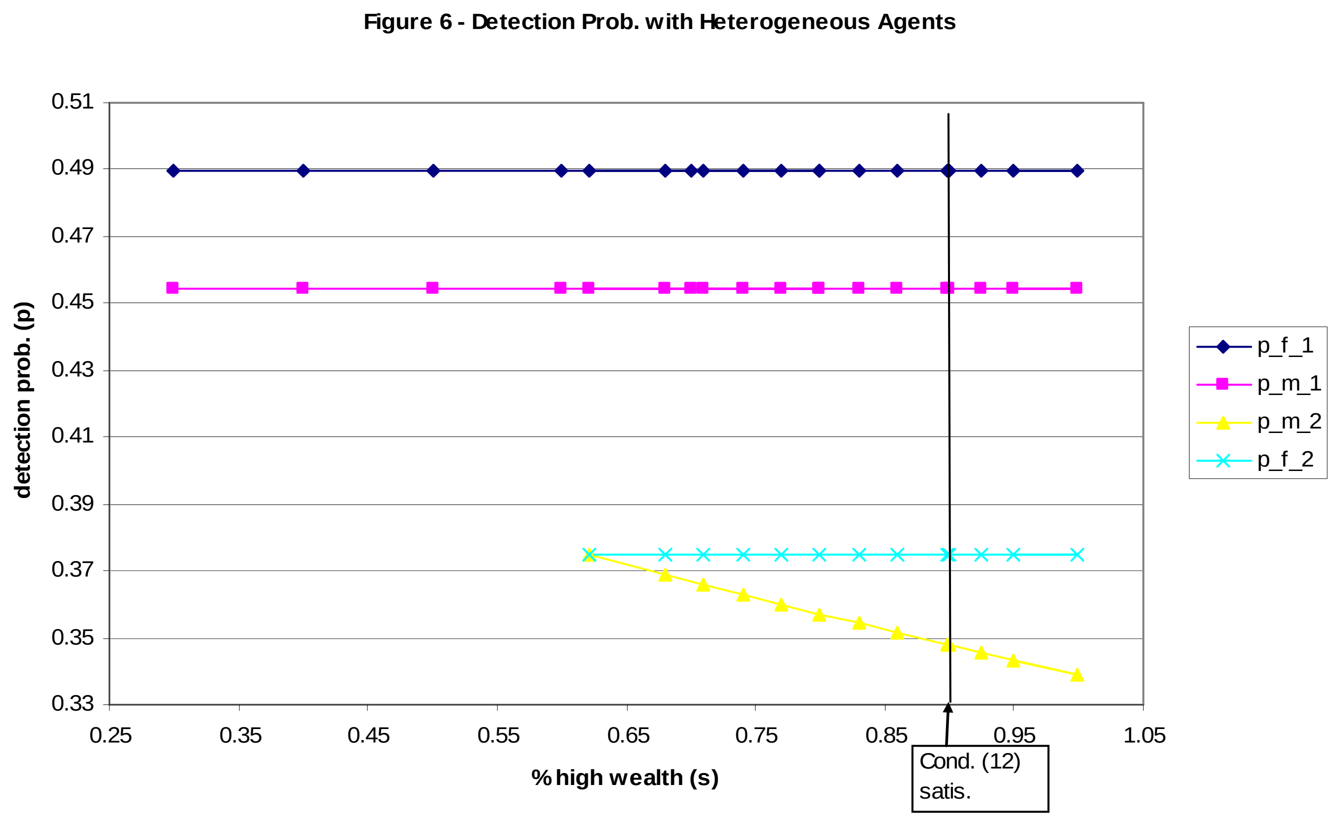


Notes: Figure 3 illustrates the relation between customer income and the exogenous agent-wealth parameter, γ. If γ increases, then the optimal penalty increases and the customer *ceteris paribus* could attain a higher expected income (moving from C_0 to C_2) and the agent would move to a lower income (holding p_1 constant at p''). Accordingly, as long as $w_1 < \alpha$, the SRO lowers p_1 as γ increases (from p'' to p' in this depiction), which reduces the customer's expected income back to α. By reducing the expected investigation costs $p_1\pi_2c$, this reduction in p_1 increases the combined agent-customer income net of the fees ($t = p_1\pi_2c$) and, thus, raises the agent's income.

Figure 4: The For-Profit SRO's Problem

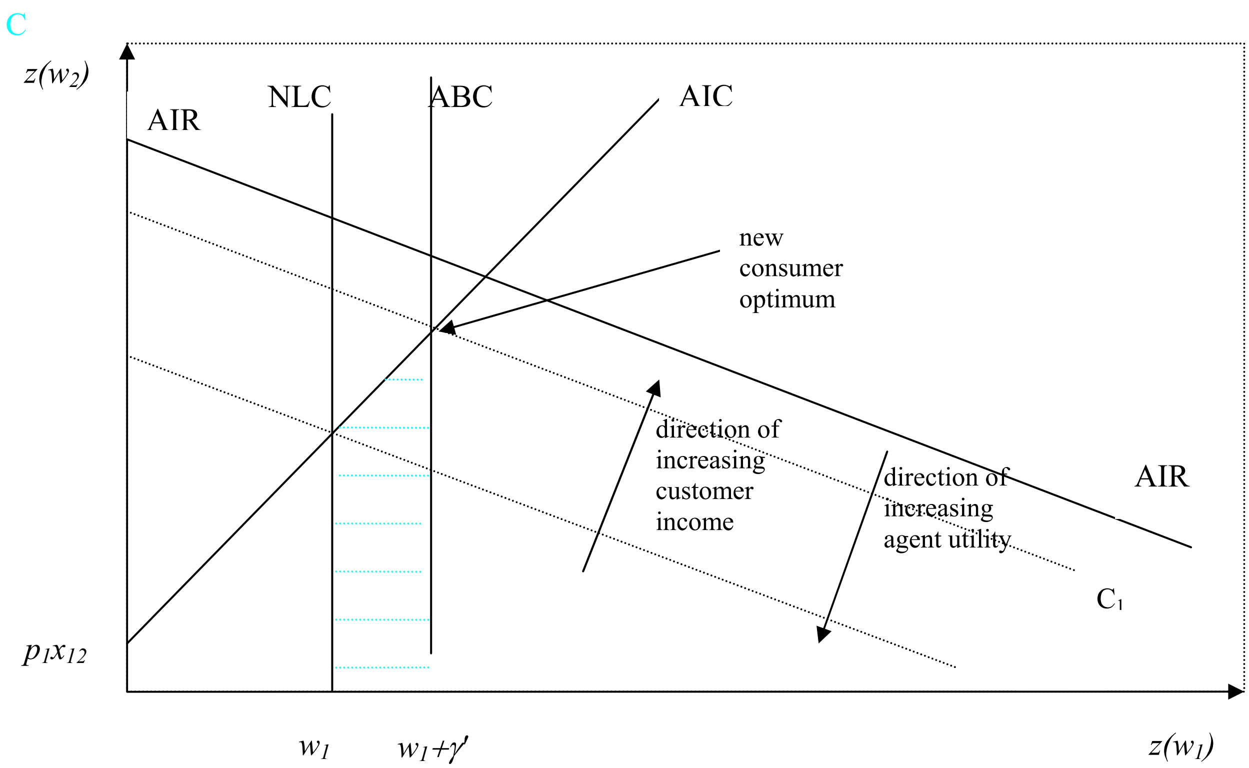


Notes: Figure 4 illustrates the for-profit SRO's optimization. C_0 is the income level associated with the customer's participation constraint, i.e., with the CIR condition (4). A_0 is the income level associated with the agent's participation constraint, i.e., with the AIR condition (2). As in Figure 3, $z^*(w_2) = z^*(w_1) + p_1 (w_2 - w_1)$ is the agent's incentive compatibility constraint (1), and the vertical NLC line at w_1 reflects the no-loss constraint (3). For some arbitrary value of $p_1 < P_F$, say p', the agent would expect income $A_1 > A_0$, and t would be set equal to $A_1 - C_0$. As p_1 is increased towards P_F, the agent's income falls towards A_0, and t can be increased without violating condition (4). Thus, the equilibrium fee t is the vertical distance between A_0 and C_0, with $p_1 = P_F$.



Parameters are $\alpha = 1.5$, $\gamma_H = 2$, $\gamma_L = 1.25$, $\pi_1 = .5$, $c = .25$, $w_2 = 2.2$, $w_1 = 1$

Figure 7: Importance of Agent-Liability Limits



Notes: Figure 7 illustrates the relation between customer income and the exogenous agent-wealth parameter, γ, in the event that agent income can be negative in some states of the world. In Figure 1, the agent's income is restricted to non-negative values in all states of the world, i.e., the no-loss condition (3) required that $z(w_1) \leq w_1$. This constraint is represented by the vertical NLC line above. In Figure 7, this no-loss constraint is replaced by the weaker requirement that the agent can lose money in some states of the world, but only to the extent that his wealth γ is sufficient to cover the loss: $z(w_1) \leq w_1+\gamma$. This weaker no-bankruptcy condition is represented by the vertical ABC line above. *Ceteris paribus*, relaxing the NLC allows the customer to attain a higher expected income (moving from C_0 to C_1) and forces the agent to a lower income (holding p_1 constant).

Appendix: Proofs.

Lemma 2: If $\pi_2 (w_2 - w_1 + \gamma) - \pi_1 c \geq \alpha - w_1$, the mutual SRO will choose

$$p_1 = p_M = \frac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma) - \pi_1 c}$$

$p_2 = 0$, $x_{12} = w_2 - w_1 + \gamma$ and $x_{ij} = 0$ for all other i,j.

Proof: To see that $p_2 = 0$ and $x_{2j} = 0$ ($j=1,2$), note that condition (3) requires that $z(w_1) < \alpha$, so condition (4) requires $z(w_2) > \alpha > z(w_1)$. Because $z(w_2) > z(w_1)$, the revelation principle implies that we can restrict our attention to contracts that induce truthful reporting by agents. As such, there is no reason for an SRO to set $p_2 > 0$, since the agent will never report that state 2 occurred if state 1 actually occurred. Consequently, x_{2j} is irrelevant, and we therefore set it equal to 0.

To show that $x_{12} \equiv x = w_2 - w_1 + \gamma$, first note that if $x \leq w_2 - w_1 + \gamma$ then the agent's income is $\pi_2 (w_2 - w_1 - px)$. Suppose $x = x' < w_2 - w_1 + \gamma$, and let p' be the value of p such that the CIR is satisfied when $x = x'$. Let $p'x' = k$. It follows that agent income is the same for any p as long as $px = k$. Customer income is $\pi_2 (w_1 + px) + \pi_1 w_1 - \pi_1 pc$, and therefore decreasing in p for $x = k/p$. Hence, setting $x = w_2 - w_1 + \gamma$ (rather than x') has no direct effect on agent's income, but relaxes the CIR, allowing the agent to profitably increase his income without violating the CIR.

To determine p, note that for given $z^*(w_1)$ and $z^*(w_2, \gamma)$ functions, the agent's wealth is decreasing in px. Therefore, the SRO will choose the minimum px that is consistent with condition (4'). Since we know that $x = w_2 - w_1 + \gamma$, the mutual SRO will choose the minimum p that satisfies

(7) $\quad z^*(w_2) \geq (\alpha - \pi_1 (w_1 - c\, p))/\pi_2$

or, given the customer's optimal fee schedule,

$$w_1 + p_1(w_2 - w_1 + \gamma) \geq (\alpha - \pi_1 (w_1 - c\, p_1))/\pi_2$$

This implies that if $\pi_2 (w_2 - w_1 + \gamma) - \pi_1 c \leq \alpha - w_1$, then there is no $p \leq 1$ which allows the SRO to provide a sufficient return to allow customers to expect an income $\geq \alpha$. As long as $\pi_2 (w_2 - w_1 + \gamma) - \pi_1 c \geq \alpha - w_1$, we have:

$$p = p_M = \frac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma) - \pi_1 c} \quad \square$$

Lemma 3: If $\pi_2 (w_2 - w_1) - \pi_1 c (w_2 - w_1) / (w_2 - w_1 + \gamma) \geq \alpha - w_1$, the for-profit SRO will choose

$$p_1 = p_F = \frac{w_2 - w_1}{w_2 - w_1 + \gamma}$$

$$t = w_1 + \pi_2 P_F x - \alpha = w_1 + \pi_2 (w_2 - w_1) - \alpha$$

$p_2 = 0$, $x_{12} = w_2 - w_1 + \gamma$ and $x_{ij} = 0$ for all other i,j.

Proof: By the same logic as for the mutual SRO, the for-profit SRO sets $x = x_{12} = w_2 - w_1 + \gamma$. To find P_F, note that the CIR condition (4) must be binding. That is, (4) can be written:

(8) $\pi_2 p (w_2 - w_1 + \gamma) = \alpha + t - w_1$

Because $z^*(w_1) = w_1$, the NLC condition (3) is binding in state 1 for any p. Condition (3) will be binding in state 2 if $z^*(w_2) = w_2$, i.e., if $p = (w_2 - w_1)/x$.

If condition (3) is not binding in state 2, then the optimal transaction fee solves the SRO's first order condition with respect to t, so that

(9) $1 - \pi_1 c \, \partial p/\partial t = 0$,

where, $\partial p/\partial t$ is the slope of the CIR constraint (8), holding customer income fixed at α. This slope is equal to $1/((w_2 - w_1+\gamma) \pi_2)>0$, and therefore (9) can be written

(9') $1 - \pi_1 c/((w_2 - w_1+ \gamma) \pi_2) = 0$.

Given that $c < \pi_2 (w_2 - w_1 + \gamma)/ \pi_1$, it follows from (9') that, as long as equation (3) is not binding in state 2, the for-profit SRO should increase the transactions fee t and the detection probability p, in order to keep customers at income level α. As long as (3) is not binding, the SRO can raise p, resulting in a higher $z^*(w_2)$, without violating any other constraint.

Once t is sufficiently large that the resulting $z^*(w_2)$ is high enough that the NLC (3) is binding, then further increases in t are not profitable. That is, once (3) is binding in state 2, the SRO can only increase p (without inducing agent exit) by reducing x. Since the optimal x is $w_2 - w_1 + \gamma$, it will not be profitable to increase p beyond

$$P_F = \frac{w_2 - w_1}{w_2 - w_1 + \gamma}.$$

Given this choice of p, the for-profit SRO chooses t to set customer income equal to α (so that equation (4') binds), or $t = w_1 + \pi_2 P_F (w_2 - w_1 + \gamma) - \alpha = w_1 + \pi_2 (w_2 - w_1) - \alpha$. Finally, the SRO will not operate unless profits are positive, i.e., unless:

$$\pi_2 (w_2 - w_1)- \pi_1 c(w_2 - w_1)/(w_2 - w_1 + \gamma) \geq \alpha - w_1 \quad \square$$

Proposition 1: If $\pi_2\ (w_2 - w_1) - \pi_1\ c > \alpha - w_1$, then the for-profit SRO will spend at least as much on enforcement as the mutual SRO.

Proof: Lemmas 1 and 2 together imply that, if $\gamma < \alpha - w_1$ and $\pi_2\ (w_2 - w_1) - \pi_1\ c > \alpha - w_1$, then the mutual SRO will choose

$$P_M = \frac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma) - \pi_1 c}$$

and the for-profit SRO will choose

$$P_F = \frac{w_2 - w_1}{w_2 - w_1 + \gamma}$$

Hence, $P_F > P_M$ if

$$(w_2 - w_1)[\pi_2(w_2 - w_1) - (\alpha - w_1) - \pi_1 c] + \gamma[\pi_2(w_2 - w_1) - (\alpha - w_1)] > 0$$

This expression is positive if $\pi_2\ (w_2 - w_1) - \pi_1\ c > \alpha - w_1$. Note that t is greater than enforcement costs if and only if

$$\pi_2(w_2 - w_1) - \pi_1 c(w_2 - w_1)/(w_2 - w_1 + \gamma) > \alpha - w_1 \ \square$$

Lemma 4: AIC-H is binding whenever AIC-L is binding; he reverse is not true.

Proof: There are two cases to consider, depending on the SRO's choice for the penalty x.
(i) If the SRO has set $x \leq w_2 - w_1 + \gamma_L$, then γ_L does not enter the AIC, so the Lemma holds trivially.
(ii) If the SRO has instead set $x > w_2 - w_1 + \gamma_L$, then the right-hand side of AIC-L equals $-p\gamma_L + (1-p)\ (w_2 - w_1)$. In this second case, it follows from (10) and from $x \geq w_2 - w_1 + \gamma_L$ that AIC-H is binding whenever AIC-L is binding – whereas the reverse is not true. $\square$

Lemma 6: Both a high-wealth and a low-wealth agent would choose $x \geq w_2 - w_1 + \gamma_L$.

Proof: When $x < w_2 - w_1 + \gamma_L$, the AIC is the same for high and low-wealth agents:

$$w_2 - z(w_2) \geq w_2 - w_1 - px.$$

Hence, agent wealth does not affect the decision on whether to honestly report the state. Since the customer wants to set $z(w_2)$ to just satisfy the AIC, $z(w_2) = w_1 + px$, and the income of an agent of type i is,

$$\pi_2\ (w_2 - w_1 - px) - \pi_1\ \gamma_i$$

This means that both kinds of agents' incomes fall with *px* at the same rate. As in Lemma 2, letting $x = k/p$, it follows that increases in *x* accompanied by decreases in *p* that leave *px* unchanged will also leave the AIC for both kinds of agents unaffected, but relax the CIR, leading to higher agent profits. Therefore, setting $x < w_2 - w_1 + \gamma_L$ is never optimal. The upper bound obtains from limited liability condition for high-wealth agents. □

Lemma 7: If $x = x_L$, then the mutual SRO chooses $P_{M,1} = \dfrac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma_L) - \pi_1 c}$

If $x = x_H$, then $P_{M,2} = \dfrac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma_L) + \pi_2 s(\gamma_H - \gamma_L) - [\pi_1 + \pi_2(1-s)]c} < P_{M,1}$

In the latter case, $\partial p/\partial s < 0$.

Proof: Since agents' income is decreasing in *p* (holding x constant), the mutual SRO wants to choose the minimal *p* consistent with the customer's participation constraint (CIR).

(i) When $x = x_L$, the CIR for all agents is $\pi_1 w_1 + \pi_2 [w_1 + p x_L] - \pi_1 cp \geq \alpha$, or

$$w_1 + \pi_2 p x_L - \pi_1 cp \geq \alpha,$$

Solving for *p* yields $P_{M,1} = \dfrac{\alpha - w_1}{\pi_2 x_L - \pi_1 c}$, i.e., $P_{M,1} = \dfrac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma_L) - \pi_1 c}$.

The last expression is the same as that found in Lemma 2 when agents are homogenous.

(ii) When $x > x_L$, x_L is the maximum fine that can be imposed on a low-wealth agent. Hence, as stated in Corollary 1 and footnote 22, AIC-L will be violated and low-wealth agents will misreport. This violation implies that, for $x > x_L$, the CIR becomes

$$w_1 + \pi_2 [spx + (1-s)px_L] - [\pi_1 + (1-s)\pi_2] pc \geq \alpha$$

where, the probability $\pi_1 + (1-s)\pi_2$ of having state 1 reported reflects misreporting of state 2 by low-wealth agents.

In particular, when $x = x_H$, the CIR simplifies to

$$w_1 + \pi_2 p[x_L + s(\gamma_H - \gamma_L)] - [\pi_1 + (1-s)\pi_2] cp \geq \alpha$$

Solving for *p* yields $P_{M,2} = \dfrac{\alpha - w_1}{\pi_2(w_2 - w_1 + \gamma_L) + \pi_2 s(\gamma_H - \gamma_L) - [\pi_1 + \pi_2(1-s)]c} < P_{M,1}$ □

Proposition 2: Let $x_L \equiv w_2 - w_1 + \gamma_L$ and $x_H \equiv w_2 - w_1 + \gamma_H$;

(a) If the mutual SRO maximizes the expected income of low-wealth agents, and if parameter values are such that both the AIR for high-wealth agents and the CIR can be met (precisely, if $w_1 + (\frac{w_2 - w_1}{x_H})\pi_2(x_L + s(\gamma_H - \gamma_L)) - \alpha - \pi_1(\frac{w_2 - w_1}{x_H})c) > 0$), then the SRO optimally sets $x = x_H > x_L$ and misreporting occurs in equilibrium.

(b) If the SRO maximizes the expected income of high-wealth agents, then the penalty for misreporting x is either x_L *or* x_H. The SRO sets $x = x_H$ (allowing misreporting) if the following conditions are met:

i. $\pi_2(1\text{-}s)\, x_L \leq [\pi_1 + \pi_2(1\text{-}s)]c$ (12)

ii. parameter values are such that equations (2) and (4) can be satisfied

iii. parameter values are such that $x_H P_{M,2} < x_L P_{M,1}$, or

$c[(\gamma_H - \gamma_L)\pi_1 - (1-s)x_L\pi_2] > (1-s)(\gamma_H - \gamma_L)\pi_2 x_L$ (**)

If these conditions are not met, then high-wealth agents strictly prefer $x = x_L < x_H$. Hence, a mutual SRO that maximizes the expected income of high-wealth agents will prefer the no misreporting equilibrium unless all three conditions are met.

Proof: Lemma 6 implies that $x \geq w_2 - w_1 + \gamma_L$. Assuming an equilibrium exists, whether misreporting occurs will depend on whether $x > x_L$. When $x > x_L$, the expected income of a low-wealth agent (who always misreports in state 2 and gets caught with probability p) is

$$\pi_2 [(1-p)(w_2 - w_1) - p\, \gamma_L]$$

which is decreasing in p and independent of x. Since increases in x allow the SRO to decrease p without inducing customer exit, low-wealth individuals always prefer higher x, up to the point that x is no longer binding on high-wealth individuals, i.e., $x = x_H$.

The expected income of a high-wealth agent (who does not misreport) when $x \varepsilon [x_L, x_H)$ is

$$\pi_1 (w_1 - z^*(w_1)) + \pi_2 (w_2 - z^*(w_2)) = \pi_2 (w_2 - w_1 - px)$$

Hence, for $x > x_L$, increasing x will only be profitable for high-wealth individuals if it allows px to fall (relaxes the CIR). The CIR in this case is

CIR (4'') $w_1 + p\, \pi_2 [s\, x + (1-s)(w_2 - w_1 + \gamma_L)] - (\pi_1 + (1-s)\pi_2)\, cp \geq \alpha$

where the transaction fee $t = (\pi_1 + (1-s)\pi_2)\, cp$ because state 2 is always misreported by low-wealth agents.

Letting $px = k$, the change in the CIR from increasing p is

(12) $(1\text{-}s)\, \pi_2\, x_L - [\pi_1 + (1-s)\, \pi_2)]\, c$

If expression (12) is positive, then a reduction in x (along with an increase in p so that px is unchanged) will relax the CIR, which means that the SRO can reduce px (thereby increasing the high-wealth's agent's income) without violating the CIR. Hence, if this derivative is positive, a high-wealth agent will want to set x arbitrarily close to x_L (conditional on $x > x_L$). Moreover, because the high-wealth agent's profits when $x = x_L$ are unambiguously higher (and misreporting is avoided) than his profits from $x = x_L + \varepsilon$ (as ε goes to zero), high-wealth agents will prefer $x = x_L$ if (12) is positive. Conversely, if the expression is negative, a high-wealth agent may want to set $x = x_H$. The condition under which a high-wealth agent prefers $x = x_H$ is that his income when $x = x_H$ (and p is set at $P_{M,2}$) is higher than his income when $x = x_L$ (and p equals is set at $P_{M,1}$). Since the high-wealth agent's income is $\pi_2\, (w_2 - w_1 - px)$, the agent prefers to allow misreporting if $x_H\, P_{M,2} < x_L\, P_{M,1}$, or

$$(\alpha - w_1)[\frac{x_L}{\pi_2 x_L - \pi_1 c} - \frac{x_L + (\gamma_H - \gamma_L)}{\pi_2 x_L + \pi_2 s(\gamma_H - \gamma_L) - (\pi_1 + (1-s)\pi_2)c}] > 0$$

Which is true if

$$c[(\gamma_H - \gamma_L)\pi_1 - (1-s)x_L \pi_2] > (1-s)(\gamma_H - \gamma_L)\pi_2 x_L \quad (**)$$

Finally, in order for any x and p to constitute an equilibrium, it must be the case that both the CIR and AIR can simultaneously be satisfied. The highest p that can satisfy the high-wealth agent's AIR in the equilibrium with misreporting by low-wealth agents solves

$$\pi_1\, w_1 + \pi_2\, [w_1 + p\, x_H] = \pi_1\, w_1 + \pi_2\, w_2 \quad \text{or}$$

$$p = (w_2 - w_1)/x_H$$

and this p can satisfy the CIR if

$$w_1 + \pi_2\, p\, [s\, x + (1\text{-}s)\, x_L] - p\, [\pi_1 + (1\text{-}s)\, \pi_2]\, c - \alpha > 0, \text{ or}$$

$$w_1 + (\frac{w_2 - w_1}{x_H})\pi_2(x_L + s(\gamma_H - \gamma_L)) - \alpha - \pi_1(\frac{w_2 - w_1}{x_H})c) \geq 0 \quad .\square$$

Proposition 3: If parameter values are such that both equations (13) and (15a) hold [i.e., $\frac{w_2 - w_1}{x_H} > \frac{\alpha - w_1}{\pi_2[x_L + s(\gamma_H - \gamma_L)] - [\pi_1 + \pi_2(1-s)]c}$],

then the for-profit SRO will choose $x = x_H$, so that misreporting will occur in equilibrium if both

(16) $\pi_2(1-s)x_L \leq [\pi_1 + (1\text{-}s)\,\pi_2]\,c$, and

$c[(\gamma_H - \gamma_L)\pi_1 - (1-s)x_L\pi_2] > (1-s)(\gamma_H - \gamma_L)\pi_2 x_L$ (**)

Otherwise, there will be no misreporting in equilibrium.

Proof: Misreporting occurs in equilibrium if $x > x_L$. First note that x will be at least x_L. To see this, note that when $x < x_L$, the logic of Lemma 6 holds; it will be profitable to raise x and lower p. That is, the AIC is the same for high and low-wealth agents for $x < x_L$, and can be written as

$z(w_2) \leq w_1 + px$.

Thus, if the SRO sets p and x in order that equation (14) just binds and $x < x_L$, then only the product p times x matters. Similarly, the customer's income is

$\pi_2\,(w_1 + px) + \pi_1 w_1 - t$.

so that only the product px affects customer income.

Hence, if $x < x_L$, and the SRO increases x and lowers p to keep px fixed (so that the constraints continue to bind), it will not affect SRO income, but it will reduce SRO costs. Thus, it will never be optimal to set $x < x_L$.

Next consider $x > x_L$. Suppose that x and p are chosen so that the AIR holds for high-wealth agents (and hence necessarily holds for low-wealth agents). Holding xp constant at k, the AIR will hold at all $x \in [x_L, x_H)$. As before, for any p and x, the SRO will then set t such that (15) is binding on customers. The CIR when $x > x_L$ is

CIR (15a) $[\pi_2\,(s\,px + (1\text{-}s)\,x_L) + w_1] - t \geq \alpha$

So that the change in the CIR from increasing p (decreasing x) is

$\pi_2\,(1\text{-}s)\,x_L$

This represents the amount by which the SRO can increase t as it lowers x and raises p (to keep $px = k$). If this expression is more than the cost of increasing p (which is $[\pi_1 + (1\text{-}s)\,\pi_2]$ c), then the SRO will increase its profits by lowering x (until x is arbitrarily close to x_L). Since the SRO profits are higher at $x = x_L$ than at $x = x_L + \varepsilon$ (in the limit as ε goes to zero), it will set $x = x_L$

if the inequality in (16) does not hold. If the inequality in (16) does hold, then the SRO will increase its profits by raising x (up to x_H). However, since profits at $x = x_L + \varepsilon$ are lower than profits at $x = x_L$ setting $x = x_H$ will be more profitable than $x = x_L$ only if the increase in the transactions fee from avoiding misreporting is less than the increased enforcement costs of avoiding misreporting. (Misreporting may either increase or decrease enforcement costs. If misreporting increases enforcement costs, then the for-profit exchange will always chosen $x = x_L$, so that there is no misreporting.) The increase in transaction fees is less than the enforcement costs saving if

$$t_1 - t_2 = \pi_2 P_{F,1}(x_L) - \pi_2 P_{F,2}[(1\text{-}s)\, x_L + s\, x_H)] < \pi_1 c\, (P_{F,1} - P_{F,2}) - \pi_2 c\, (1\text{-}s)\, P_{F,2}, \text{ or}$$

$$c[(\gamma_H - \gamma_L)\pi_1 - (1-s)x_L\pi_2] > (1-s)(\gamma_H - \gamma_L)\pi_2 x_L \quad (**)$$

Given the behavior of customers and agents, this means that misreporting will occur if and only if both the inequality in expression (16) and condition (**) hold. □

Proposition 4: If $\pi_2 (w_2 - w_1) - \pi_1 c > \alpha - w_1$ then $P'_F \geq P'_M$, and $P'_F > P'_M$ for $\gamma < \alpha - w_1$,

Proof*:* If $\pi_2 (w_2 - w_1) - \pi_1 c > \alpha - w_1 > \gamma$ then

$$p'_M = \frac{\alpha - w_1 - \gamma}{\pi_2(w_2 - w_1) - \pi_1 c}$$

and

$$p'_F = \frac{\pi_2(w_2 - w_1) - \gamma}{\pi_2(w_2 - w_1)}$$

so that $P'_F \geq P'_M$ if $\pi_2(w_2 - w_1)[\pi_2(w_2 - w_1) - (\alpha - w_1 - \gamma) - \pi_1 c] + \gamma\pi_1 c$, which is positive. When $\pi_2 (w_2 - w_1) \geq \gamma \geq \alpha - w_1$, $P'_M = 0$ and $P'_F \geq 0$. □