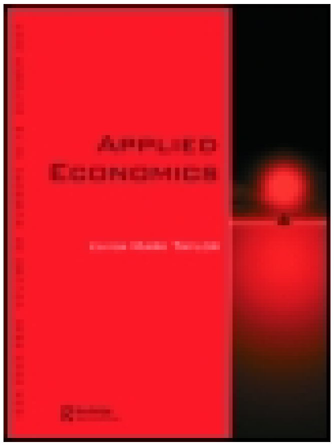

Applied Economics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/raec20

Does self-regulation work? Experimental evidence of the reputational incentives of Self-Regulatory Organizations

José Luis Lima[a] & Javier Núñez[b]

[a] Economics Department and Intelis, University of Chile, Santiago, Chile

[b] Economics Department, University of Chile, Santiago, Chile

Published online: 26 May 2015.

To cite this article: José Luis Lima & Javier Núñez (2015): Does self-regulation work? Experimental evidence of the reputational incentives of Self-Regulatory Organizations, Applied Economics, DOI: 10.1080/00036846.2015.1030567

To link to this article: http://dx.doi.org/10.1080/00036846.2015.1030567

Applied Economics, 2015
http://dx.doi.org/10.1080/00036846.2015.1030567

Routledge
Taylor & Francis Group

Does self-regulation work? Experimental evidence of the reputational incentives of Self-Regulatory Organizations

José Luis Lima[a] and Javier Núñez[b,*]

[a]*Economics Department and Intelis, University of Chile, Santiago, Chile*
[b]*Economics Department, University of Chile, Santiago, Chile*

Self-regulation (SR) is a common way of enforcing quality in markets (such as banking, financial services and several professions) and in a variety of public and private organizations. We provide experimental evidence of the reputational incentives of self-regulatory organizations (SROs) to publicly disclose versus cover-up fraud in an incomplete information environment. We find that observed behaviour is generally consistent with Bayesian equilibrium when subjects are informed about the relative likelihood of fraud detection by a 'vigilant' versus a 'lax' SRO type. In particular, a fraud disclosure equilibrium is supported when subjects are informed that the 'vigilant' SRO is more likely to detect fraud; otherwise, a cover-up equilibrium is supported. However, when subjects are not informed about the relative likelihood of fraud detection by the SRO types (as expected in real SR situations), no equilibrium is strongly supported. Our results suggest that in practice, the reputation-based incentives for effective SR may be inherently ambiguous and weak.

Keywords: self-regulation; Self-Regulatory Organizations; credence goods; quality regulation

JEL Classification: C90; D18; D82; L15; L84

I. Introduction

Self-Regulation (SR) exists in several industries worldwide with the purpose of preventing fraud and malpractice.[1] SR is found for example in banking and financial services, enforcement of pollution and emissions standards and in many collegiate professions such as the legal profession, medicine, accounting and auditing.[2] SR is also in place in diverse institutions and organizations, such as political

*Corresponding author. E-mail: jnunez@fen.uchile.cl

[1] Several market forces may mitigate fraud and malpractice. See, for example Dulleck and Kerschbamer (2006), Emons (1997), Wolinsky (1993) and Taylor (1995).

[2] Some examples of SR in banking and financial services are found in Yue and Ingram (2012), Omarova (2011). For a description of SR in collegiate professions, see, for example Stephen and Love (1999), Wallace *et al.* (2000) and Casterella

parties, religious organizations and public agencies. In spite of its widespread prevalence in markets and nonmarket institutions, SR has received limited attention from researchers, and therefore, little is still known about the incentives under which SR may effectively work.[3] This is a relevant underexplored issue since SR has often been questioned as a credible or reliable way of preventing fraud and malpractice.[4] In this context, this article examines from an experimental perspective the reputation-based incentives of self-regulatory organizations (SROs) for publicly disclosing the fraud perpetrated by their members. We focus on reputation-based incentives because, as the literature suggests, SR is often found in markets and institutions that deliver credence goods (whose quality is hardly observable by experience or search),[5] and thus the main incentive for effective SR arises from the (alleged) private benefits of building a reputation for enforcing high-quality standards.[6]

We claim that an experimental approach to study and inform about the reputational incentives of SR is important because empirical (or field) research on SROs behaviour may be constrained by the fact that quality is often nonobservable, and therefore, fraud may often go undetected by researchers, more so in credence-good contexts.[7] Experimental evidence on SR is also important because both theoretical and applied research suggest that SROs often face conflicting reputation-based incentives for enforcing quality and disclosing fraud to the public (Núñez, 2001, 2007; Carson, 2003; DeMarzo *et al.*, 2007).

More specifically, in Núñez (2001, 2007) (which we follow in particular), the conflicting reputational incentives of SR arise because disclosure of fraud would yield a reputational gain to the SRO if it is interpreted by the public as a sign of a 'vigilant' SRO, but it may also signal a 'lax' (low-vigilance) SRO in which there is widespread underlying fraud as a result. It turns out that both interpretations have theoretical (Bayesian) rationale depending on the public's beliefs regarding the relative likelihood of fraud detection by the 'vigilant' versus the 'lax' SRO type: the optimistic interpretation about an SRO's type requires the public to believe that the 'vigilant' SRO is more likely to detect fraud than the lax SRO type; otherwise, the pessimistic interpretation prevails. However, it is not clear whether a vigilant or a lax SRO should empirically be more or less likely to detect fraud because the underlying level of fraud is endogenously decreasing in the degree of vigilance. Consequently, a vigilant SRO may in principle be more or less likely to detect fraud than a lax SRO type, depending on the functional form of the fraud detection probability (a function of both fraud and vigilance levels). This theoretical ambiguity arising from the two opposite yet plausible interpretations of fraud disclosure and cover-up by the public is the key to understanding the strength of reputational incentives for effective SR. In this context, in this article we study experimentally whether (and under what conditions) SROs either disclose or cover-up evidence of fraud and how consumers interpret and react to both events.

For this purpose, we perform an experiment in an incomplete information/credence-good context that resembles the signalling strategic environment outlined above. We develop a simplified version of the models of Núñez (2001, 2007), which is more suitable for experimental testing, and from which explicit testable implications are derived. The simplified model (and the experiment) has two players: the 'Public' and an SRO that can be of two possible types: a 'High-vigilance/low-fraud' type or a 'Low-vigilance/high-fraud' type. These SRO types are intended to capture in reduced form the equilibrium levels of vigilance and fraud arising from the

et al. (2009) for legal, accounting and auditing professions, respectively. Examples of SR in enforcement of emissions and pollution standards are found in Gamper-Rabindran and Finger (2013) and Lenox and Nash (2003).

[3] However, some authors have studied the role of SR in enhancing market power. See, for example DeMarzo *et al.* (2005) and Shake and Sutton (1981).

[4] See, for example Yue and Ingram (2012), Wallace *et al.* (2000), Carson (2003), Lenox and Nash (2003) and Núñez (2001, 2007).

[5] Darby and Karni (1973) define 'credence' goods and services as those whose qualities are expensive to observe or judge even after purchase. For example, Stephen and Love (1999) observe that legal services are credence goods for most clients, who are usually less informed about the nature of legal problems than lawyers.

[6] See, for example Yue and Ingram (2012), Wallace *et al.* (2000) and Lenox and Nash (2003), who observe that effective enforcing of high-quality standards requires monitoring the quality delivered by SRO members and imposing effective sanctions on those members that breach self-regulation.

[7] See, for example Knechel *et al.* (2013) for a discussion of challenges and limitations of the existing empirical work about the effectiveness of self-regulation in enhancing quality in auditing.

strategic interaction between the SRO and its members in a context where SROs differ in their vigilance cost or technology, as in Núñez (2001, 2007). After privately observing its assigned type, the SRO either exogenously detect or does not detect fraud, with a type-dependent exogenous probability. If fraud is detected, the SRO decides whether to disclose it to the public or cover it up. The Public must express its opinion regarding the SRO's type after observing either disclosure or the absence thereof. The Public gets a positive payoff if its opinion matches the true SRO type, and gets zero otherwise. The SRO gets a higher payoff when the Public believes (correctly or not) that the SRO is the high-vigilance-low-fraud type. We show that in this model, a disclosure pooling Bayesian Equilibrium exists only if the high-vigilance SRO type has a higher probability of detecting fraud. Otherwise, the only equilibrium involves a cover-up. These findings provide the key testable implications for our experiment regarding the SRO's incentives to disclose or cover-up evidence of fraud and the Public's reactions to these events.

We argue that in this framework, effective SR would require two necessary conditions. First, the SRO and the Public must behave in (an approximately) Bayesian way, that is, their actions must be consistent with the Public's beliefs about the SRO's type updated from Bayes' rule whenever possible.[8] Second, the SRO and the Public must share the belief that the 'vigilant' SRO is more likely to detect fraud than the 'lax' SRO type. To study these two conditions separately, we perform the experiment in four principal treatments, depending on whether the high-vigilance SRO type has a higher or a lower underlying probability of detecting fraud than the low-vigilance type, and on whether this information is actually revealed or not to the subjects at the beginning of the experiment. Revealing the underlying fraud detection probabilities of each SRO type to the subjects allows studying whether observed behaviour and updated beliefs are consistent with the theoretical Bayesian predictions. On the other hand, not revealing to the subjects the underlying probabilities of fraud detection allows studying if there is an 'obvious way to play' the SR game under conditions of uncertainty about the fraud detection probabilities (perhaps the more realistic scenario) and whether either a fraud disclosure or a cover-up equilibrium is supported in this context.

II. A Simple Model of SR

In this section, we develop a reference model that follows the SR models of incomplete information presented in Núñez (2001, 2007). In these models, an SRO has private information regarding its vigilance cost, and the SRO member's optimal level of fraud x^* (who would be closed-door punished by the SRO if detected) is inversely determined by the SRO's level of vigilance y, such that $x^*(y)$. The SRO type is important to the Public because although fraud is not observed, they know that, in equilibrium, a low-cost SRO will be more vigilant, and therefore, the underlying equilibrium level of fraud will be lower. The probability of fraud detection by the SRO is $P(x, y)$, which is increasing in both fraud x and vigilance y. If fraud is detected, the SRO must decide whether to disclose it to the Public or cover it up. The Public is unable to observe the fraud level, yet they can build rational (Bayesian) beliefs about the SRO's type based on the disclosure (or nondisclosure) of fraud. This in turn may provide a reputational incentive for the SRO to disclose fraud, if detected, and signal a 'vigilant' (low cost) type to the Public. However, the crux of the models in Núñez (2001, 2007) is that the probability of fraud detection $P(x^*(y), y)$ does not necessarily increase with the SRO's degree of vigilance, as the degree of fraud also varies inversely with the degree of vigilance. Hence, a vigilant SRO may in principle have a *higher* or a *lower* probability of detecting fraud than a lax SRO, depending on plausible parameters of the model. This implies that the reputational incentives to disclosing fraud may be positive or negative, depending on whether the *total* derivative of $P(x^*(y), y)$ is increasing or decreasing with respect to vigilance y. In particular, a Bayesian pooling

[8] For other experimental analysis of Perfect Bayesian Equilibrium in signalling games, see, for example, Brandts and Holt (1992), Banks *et al.* (1994) and Blume *et al.* (2004). These concepts have been used in the experimental analysis of strategic situations such as political lobbying (Potters and Van Winden, 2000), stock-selling under asymmetric information (Cadsby *et al.,* 1998), agency problems with ratchet effects (Chaudhuri, 1998; Cooper *et al.,* 1999) and limit-pricing in entry games (Cooper *et al.*, 1997a, b; Cooper, 2004).

equilibrium with fraud disclosure and a positive level of vigilance can only exist when $P(x^*(y), y)$ increases with vigilance. Otherwise, a cover-up equilibrium with no vigilance and the maximum level of fraud exists. The intuition is that the Public would confidently update their beliefs in favour of the vigilant SRO type only when fraud detection and disclosure is more likely to occur for the high-vigilance type, which may or may not be the case.

Our experimental design follows a simplified version of the models in Núñez (2001, 2007) that focuses on the SRO exposure incentives and is more suitable for their experimental assessment. Vigilance and the level of fraud are now exogenous, but there are two types of SROs: a 'high vigilance – low fraud' SRO, or type 'H', and a 'low vigilance – high fraud' SRO, or type 'L', to resemble, respectively, the low and high vigilance cost SRO types described earlier. Therefore, these SRO types are intended to reflect the different equilibrium levels of vigilance and fraud discussed in Núñez (2001, 2007). In our simplified model, both SRO types exist with equal probability. After privately observing its own type, each SRO type detects fraud with exogenous probability P_H and P_L, respectively, which are common knowledge. The sign of $P_H - P_L$ is the first control variable of our experiment, which determines whether fraud is more likely detected by the vigilant or the lax SRO type. If fraud is detected, the SRO decides to either disclose it to the Public ($d_i = 1$) or cover it up ($d_i = 0$), where $d_i \in \{0, 1\}$, $i = H, L$ denotes disclosure probabilities in a pure-strategy setting. If fraud is not detected, no disclosure occurs.

The Public only observes fraud disclosure or no disclosure (not the SRO's type), after which they update its beliefs regarding the SRO's type using Bayes' rule whenever possible. These updated beliefs sustain the Public's opinion about the SRO type, which can be 'SRO is *H*', 'SRO is *L*' or 'Not sure'. The Public's payoffs were set to provide an incentive to reveal an opinion about the SRO's type: if their opinion coincides with the true type of the SRO, they get $U > 0$, and zero otherwise. If 'Not sure', they get a lottery with expected payoff 0.5 *U*. This is justified by the argument that the option 'not sure' suggests that consumers (as the Public) would believe that the *H* and *L* SROs are equally likely, and therefore, their subjective expected payoff of choosing either SRO type (or any randomization between *H* and *L*) is approached by 0.5 *U*. Therefore, the payoff structure implies that the Public would benefit from building a correct opinion about the SRO's true type (which proxies its hypothetical decision of whether to purchase or consume the SRO's goods or services or not).

Table 1. Payoffs

PUBLIC	**Payoff**
If opinion matches the true SRO type	$U > 0$
If opinion does not match the true SRO type	0
If '*Not sure*': lottery	0.5 U
SRO (types H and L)	**Payoff**
If Public's opinion is '*SRO is H*'	$W > 0$
If Public's opinion is '*SRO is L*'	0
If Public is '*Not sure*': lottery	0.5 W

Both SRO types earn a reputational gain $W > 0$ if the Public form an opinion 'SRO is *H*', and get zero (no reputational gain) if the Public believes 'SRO is *L*'. If the Public is 'Not sure', the SRO gets a lottery with expected payoff 0.5 *W*. Table 1 summarizes the payoff structure of this incomplete-information SR game.

Figure 1 shows the game outlined above in extended form with payoffs $U = W = 100$ and where *a, b, c* and *d* are the Public's updated beliefs regarding the SRO's type associated with every decision node in the information set following no fraud disclosure and *e* and *f* for the nodes of the information set following fraud disclosure. Note that 'no disclosure' is an event that may arise either from explicit cover-up by the SRO (as in nodes at *a* and *d*) or from the SRO not detecting any fraud at all (as in nodes at *b* and *c*).

As Fig. 1 and Table 1 show, the payoffs are the same for both SRO types. As a consequence, only pooling Perfect Bayesian Equilibria (PBE) are possible in this game, as the incentives to either disclose fraud or cover it up will be the same for both SRO types. Also, the beliefs on decision nodes in information sets off the equilibrium path of play cannot be refined using dominance or dominance-in-equilibrium refinements, as in Cho and Kreps (1987) 'intuitive criterion'.[9]

[9] Learning models can be used to refine equilibria in signalling games (see, for example, Brandts and Holt, 1993; Anderson and Camerer, 2000; Cooper, 2004). However, this is not possible in our credence-good context since the public cannot observe in the experiment if past opinions were right or wrong.

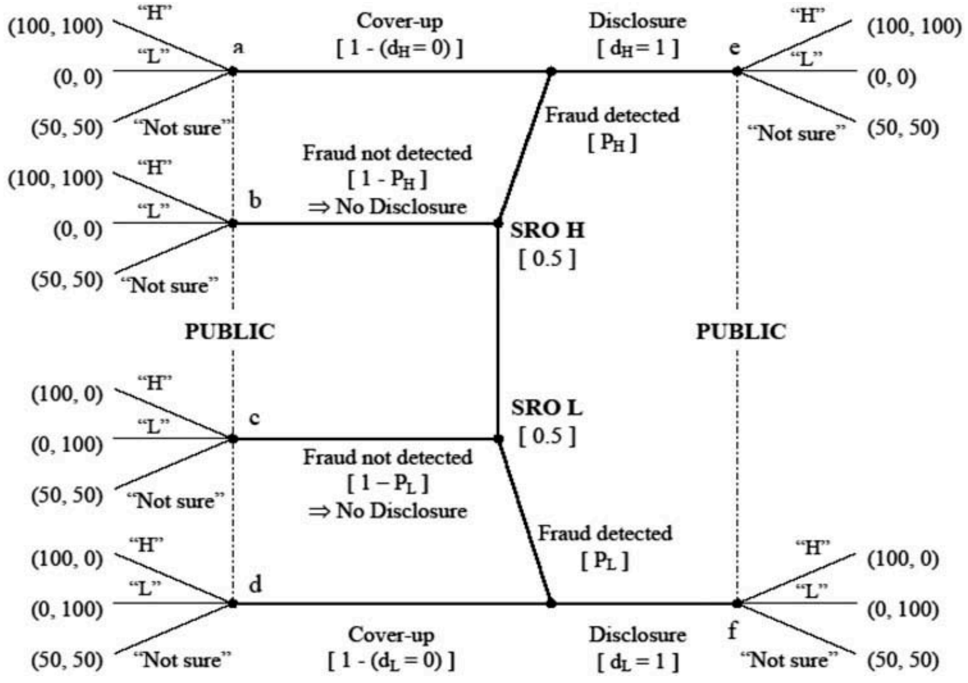

Fig. 1. SRO versus Public signalling game in extensive form

From Fig. 1, after observing disclosure, the Bayesian beliefs e and f associated with H and L SRO types are, respectively,

$$P(H \mid \text{disclosure}) = e = \frac{0.5(P_H d_H)}{0.5(P_H d_H) + 0.5(P_L d_L)}$$

$$P(L \mid \text{disclosure}) = f = \frac{0.5(P_L d_L)}{0.5(P_H d_H) + 0.5(P_L d_L)}$$

If fraud is not disclosed (either because of cover-up or no fraud detection by the SRO), the Public updates its Bayesian beliefs regarding H and L SRO types from

$$P(H \mid \text{no} - \text{disclosure}) = a + b = \frac{0.5P_H(1 - d_H) + 0.5(1 - P_H)}{0.5P_H(1 - d_H) + 0.5(1 - P_H) + 0.5P_L(1 - d_L) + 0.5(1 - P_L)}$$
$$= \frac{0.5(1 - P_H d_H)}{0.5(1 - P_H d_H) + \ 0.5(1 - P_L d_L)}$$

$$P(L \mid \text{no} - \text{disclosure}) = c + d = \frac{0.5P_L(1 - d_L) + 0.5(1 - P_L)}{0.5P_H(1 - d_H) + 0.5(1 - P_H) + 0.5P_L(1 - d_L) + 0.5(1 - P_L)}$$
$$= \frac{0.5(1 - P_L d_L)}{0.5(1 - P_H d_H) + \ 0.5(1 - P_L d_L)}$$

Multiple pooling PBE when $P_H > P_L$

A pooling disclosure equilibrium implies that beliefs e and f on the information set following disclosure can be computed using Bayes' rule from the equilibrium strategies of both SRO types ($d*_H = d*_L = 1$) as

$$e = P(H \mid \text{disclosure}) = \frac{0.5P_H}{0.5P_H + 0.5P_L}$$

$$f = P(L \mid \text{disclosure}) = \frac{0.5P_L}{0.5P_H + 0.5P_L}$$

Given these Bayesian beliefs, opinion 'H' is optimal for the Public if $e \ > f$, which is guaranteed if $P_H > P_L$. On the other hand, note that the information set following no disclosure is reached with possible probability (despite there being a pooling equilibrium), because observing no disclosure may still occur as a consequence of the SRO not detecting any fraud. Accordingly, in a pooling disclosure equilibrium, beliefs *a* and *d* must satisfy $a = d = 0$, and beliefs *b* and *c* can be computed from Bayes' rule as

$$b = \frac{0.5\ (1 - P_H)}{0.5\,(1 - P_H)\ +\ 0.5\,(1 - P_L)}$$

$$c = \frac{0.5\ (1 - P_L)}{0.5\,(1 - P_H)\ +\ 0.5\,(1 - P_L)}$$

Given $P_H > P_L$, after observing no disclosure, opinion '*L*' must be optimal for the Public. Finally, given the opinions held by the Public after observing disclosure or no disclosure, both SRO types consider it optimal to disclose fraud, if detected (i.e. $d^*_H = d^*_L = 1$).

Result 1. If $P_H - P_L > 0$, there is a pooling PBE where both SRO types disclose fraud (if detected), and where the Public holds opinion '*H*' after observing disclosure and opinion '*L*' if no disclosure is observed.

The intuition is that disclosure will optimally be chosen by both SRO types if fraud detection is an event more likely to occur for the 'vigilant' (*H*) SRO type.

However, when $P_H - P_L > 0$, there is another pooling equilibrium that involves a cover-up by the SRO. In fact, this equilibrium implies ($d^*_H = d^*_L = 0$), and therefore, the Public's posterior beliefs about the SRO's type coincide with their prior beliefs, that is, $P(H|\text{no disclosure}) = P(L|\text{no disclosure}) = 0.5$. Hence, opinions '*H*', '*L*' or '*Not sure*' yield an expected payoff of 50 and are equally optimal for the Public. The information set after observing disclosure is off the equilibrium path of play, and therefore, beliefs *e* and *f* are unconstrained by Bayes' rule. However, any possible strict cover-up equilibrium requires that both SRO types should get an expected payoff lower than 50 if they deviate from equilibrium by disclosing fraud (if detected). This requires disclosure to be answered with opinion '*L*' by the Public, which requires $e < f$, or $P_H d_H < P_L d_L$. This condition would be satisfied even if $P_H > P_L$, if inequality $d_H < d_L$ is satisfied with a sufficiently large difference. Although these out-of-equilibrium beliefs may seem peculiar, they cannot be ruled out using standard dominance or dominance-in-equilibrium refinements of PBE.

Result 2. If $P_H > P_L$, there is a pooling PBE if out-of-equilibrium beliefs *e* and *f* satisfy $P_H d_H < P_L d_L$. This equilibrium involves fraud cover-up by the two SRO types, with the Public holding posterior Bayesian beliefs equal to the prior beliefs about the SRO types and opinions '*H*', '*L*' and '*Not sure*' being equally optimal for the Public.

Results 1 and 2 indicate that both disclosure and cover-up equilibria are possible when the high-vigilance '*H*' SRO type is more likely to detect fraud ($P_H > P_L$). However, the nondisclosure equilibrium in Result 2 seems less plausible, as it requires non-obvious out-of-equilibrium beliefs, and therefore, the disclosure equilibrium in Result 1 stands out as the plausible testable equilibrium when $P_H > P_L$.

Single pooling PBE when $P_H < P_L$

From the previous discussion, a cover-up equilibrium would exist if out-of-equilibrium beliefs *e* and *f* satisfy $P_H d_H < P_L d_L$. If $P_H < P_L$, this condition is satisfied, for example, should the Public believe that an unexpected deviation from a cover-up equilibrium were equally likely to occur in both SRO types ($d_H = d_L > 0$). Note that if $P_H < P_L$, no disclosure pooling equilibrium exists because, if disclosure occurs, Bayes' rule would dictate that $e < f$, and therefore, opinion '*L*' would be the Public's best response. Accordingly, no SRO type would optimally choose to disclose fraud, as cover-up would yield a best response '*H*' from the Public and therefore a higher expected payoff for the SRO.

Result 3. If $P_H < P_L$, the unique PBE involves fraud cover-up by both SRO types. After observing no disclosure, the Public hold Bayesian beliefs equal to the prior beliefs regarding the SRO types. Accordingly, opinions '*H*', '*L*' and '*Not sure*' (or any randomization of them) are equally optimal for the Public.

The intuition of this result is that after observing no disclosure, the Public's posterior beliefs are equal to their prior beliefs, and therefore 'H' and 'L' are perceived as equally likely. Under these circumstances, no Bayesian learning is possible for the

Table 2. Theoretical proportions of play for each action, by sign of $P_H - P_L$ and type of equilibrium (%)

	$P_H < P_L$	$P_H > P_L$	
	Cover-up pooling equilibrium	Cover-up pooling equilibrium	Disclosure pooling equilibrium
SRO (both types)			
Disclose fraud, if detected	0	0	100
PUBLIC			
If fraud is disclosed, *'H'*	0	0	100
If fraud is disclosed, *'L'*	100	100	0
If fraud is disclosed, *'Not Sure'*	0	0	0
If no disclosure, *'H'*	0–100	0–100	0
If no disclosure, *'L'*	0–100	0–100	100
If no disclosure, *'Not Sure'*	0–100	0–100	0

Public. This resembles situations where consumers, unable to observe or infer differences in quality between firms, simply choose one supplier at random. The stylized model developed here cannot provide further insight as to how subjects would in practice choose between *'H', 'L'* and '*not sure*' in the experiment.

Table 2 summarizes the equilibrium proportions of play for each possible action for the cases $P_H < P_L$ and $P_H > P_L$, which are the testable hypothesis to be contrasted with the experimental evidence in Section IV.

III. Experimental Design and Procedures

Our experimental design involves four principal treatments, comprising sessions where the underlying fraud detection probabilities are $P_H - P_L < 0$ and $P_H - P_L > 0$, combined with sessions where this information is or is not revealed to the subjects before the experiment. The purpose of having the sign of $P_H - P_L$ not revealed to the subjects is twofold. First, in this case, it can be hypothesized that Bayesian-like subjects would use their intuitive or subjective prior beliefs about the sign of $P_H < P_L$, that is, their intuition of whether they would expect fraud detection to be more or less likely for the H or the L SRO type under 'reasonable' circumstances. Holding these prior subjective beliefs, Bayesian-like SRO and Public would presumably behave according to the theoretical model, and an 'obvious way to play' could arise. Second, as the experiment unfolds, one would expect some degree of learning of the underlying sign of $P_H - P_L$ by the subjects, and hence some partial convergence towards the predicted behaviour according to the underlying sign of $P_H - P_L$.

The four treatments and the corresponding number of sessions, players and rounds are shown in Table 3. We run most of the sessions using a 'meaningful context', in which the underlying game is presented to the subjects as a stylized SR game, and fewer sessions using a 'neutral context' where the same underlying strategic game is presented to the subjects in an abstract way without reference to a stylized SR context, as explained in footnote 12. The 'neutral context' is therefore used as a control to test for possible 'context effects'. The use of a meaningful context may facilitate the comprehension of the strategic complexities involved in this signalling game, speeding up the learning process and acting as a partial substitute for experience (weak context effect) or may even affect the equilibrium selection of subjects (strong context effect, as in Cooper and Kagel, 2003). A meaningful context could be a useful alternative for studying specific signalling phenomena with policy implications (as our experiment) because of the possibility of strong context effects (Cooper *et al.,* 1999). Section IV reports the results of the sessions with meaningful context. The results of the sessions using a neutral context are presented in Tables AI and AII in Appendix.

The experimental sessions were conducted on a University of Chile campus.[10] Each of the 17

[10] Five paid pilots were run before the current design to test and correct the experimental procedures.

Table 3. Experimental treatments and design

	Fraud detection probabilities								
Treatment	P_H	P_L	Context	Sessions	Players per session	Total players	SRO-Public pairs per session	Rounds per session	Total SRO-Public pairs
$P_H - P_L > 0$, revealed	0.8	0.2	Meaningful	2	12	24	6	42	504
			Neutral	1	12	12	6	42	252
$P_H - P_L > 0$, not revealed	0.8	0.2	Meaningful	3	12	36	6	42	756
			Neutral	2	12	24	6	42	504
$P_H - P_L < 0$, revealed	0.2	0.8	Meaningful	4	12	48	6	42	1008
			Neutral	1	12	12	6	36	216
$P_H - P_L < 0$, not revealed	0.2	0.8	Meaningful	2	12	24	6	36 and 42	468
			Neutral	2	12	24	6	42	504
Total				17		204			4212

experimental sessions used 12 subjects recruited among Business and Economics, Auditing and Accounting, Architecture, Geography and Design undergraduate students without relevant experience or background in experiments or Game Theory. In an introductory, brief instructions were read out loud and each subject received a written copy. Then, participants were required to fill in a questionnaire to ensure their understanding of the experiment's rules and payoffs. The correct answers were provided and further questions of the subjects were answered. Participants were then randomly assigned to isolated seats and given time to study a summary of the experiment's instructions. Participants had a registry to record the role and type they had in each block and the decisions they made and received from other participants in each round. A copy of the instructions set (translated into English) is available upon request.

In the meaningful-context sessions, we used a 'Public' versus 'organization' wording context, where an organization has members that provide either 'good service' or 'poor quality service' to the Public.[11] The organization has a level of vigilance that can be 'High' or 'Low'. The reaction of SRO members to vigilance is such that with a high level of vigilance, few members provide poor quality service, but with low vigilance, most of them provide poor quality service. These possibilities define two equally likely types of organizations: a 'high vigilance – good service' type (H) and a 'low vigilance – poor service' type (L).[12]

In sessions in which $P_H - P_L$ was revealed, we included in the instructions (and summary) the probability of 'poor service' detection for each SRO type, according to the sign of $P_H - P_L$. In the treatments in which $P_H - P_L$ was not revealed, subjects were simply told that both SRO types could sometimes detect members providing poor service, but the number of times it occurred could vary among the SRO types.

The sessions comprised 42 rounds (except for two early sessions comprising 36 rounds). The number of rounds was not revealed to avoid last-period behaviour. At the beginning of the session, six participants were randomly assigned the role 'Public' and six the role 'organization', and the latter were equally split into organizations type *H* or *L*. These roles and types remained fixed within a block of six rounds where each Public subject was paired with a different organization. Every six rounds, the roles and types were randomly and privately reassigned for another block of six rounds and so on until the end of the session. We limited the possibility of a lengthy repeated sequence of the same role in participants

[11] We used 'poor quality service' and 'poor service' to moderate the wording context and avoid the strong connotations of using 'fraud'.

[12] In the neutral-context sessions, we used a Player A – Player B wording. Player A draws (with reposition) balls from a bowl with white and red balls, facing two equally likely situations: (1) draw few balls from a bowl with many red balls, ('Draw few- many Red balls' type equivalent to 'L' type in the meaningful context) or (2) draw many balls from a bowl with few red balls, a 'Draw many – few Red balls' type ('H' type in the meaningful context). The rest of the experiment structure is similar to the meaningful context.

by imposing a probability of 0.05 of having the same role for more than two consecutive blocks. We also favoured configurations where participants experienced the role of both types of organization earlier. We conducted the experimental sessions manually; hence, we determined in advance the assignment of roles and types and the poor service detection for each organization. Eight paid graduate students were trained and employed as laboratory assistants, and three of them worked in each session.

At the beginning of each round, the organizations were given a written card containing the organization's type and an indication of whether it did or did not detect poor service in that round. Organizations disclosed poor service to the public by marking an empty box stating, 'I have detected members delivering poor service'. Only when poor service was detected did the organization decide whether it would mark the box indicating disclosure or not. False disclosure was explicitly penalized according to instructions (occurring less than 1% of cases). After receiving a message showing either disclosure or no disclosure of poor service (but not the organization's type), the public had to choose from three possible opinions: 'I think it is *H*', 'I think it is *L*', 'I am not sure'. The organization was able to observe the public's opinion, but the public was not informed whether its opinion was right or wrong, in accordance with the credence-good assumption. Wealth effects were controlled by randomly choosing and remunerating one round per block of six rounds.

We set $U = W = 2400$ Chilean Pesos (CH$, equivalent to US$ 5.9) for each round of sessions where $P_H - P_L$ was revealed, and $U = W = 2600$ CH$ (US$ 6.4) for sessions where $P_H - P_L$ was not revealed, to compensate for the higher complexity of the latter treatment. Participants earned Ch$ 10,600 on average per session, a significant amount for a developing country.[13] A typical session lasted about 1 hour and 45 minutes. The corresponding payment per hour was higher than those offered in jobs available for undergraduate students, at about Ch$ 2500–3000 per hour at the time of the experiment.[14]

IV. Experimental Results

Evidence from the meaningful-context sessions

Table 4 shows the theoretical predictions of the model developed in Section II on the columns on the left hand side, and the experimental evidence for all the meaningful-context treatments on columns on the right hand side. Columns (1) and (2) of the experimental evidence show that when $P_H - P_L$ was revealed to the subjects, the observed behaviour (statistically) resembles the expected frequencies of play implied by the model, namely a 'cover-up equilibrium' when $P_H - P_L < 0$, and a 'disclosure equilibrium' when $P_H - P_L > 0$, and so do the differences in the behaviour of the SRO and the public according to the revealed signs of $P_H - P_L$. In fact, disclosure by the SRO is statistically more frequent when $P_H - P_L > 0$ than otherwise (0.88 versus 0.35), and both disclosure proportions are statistically different from 0.5 (as a measure of random behaviour) in the directions suggested by the model. One somewhat puzzling fact is the higher than expected rate of disclosure when $P_H - P_L < 0$, of 0.35. This could be due to some degree of experimentation by the subjects and/or the possibility that part of the subjects did not consistently behave according to the model's predictions.[15] Nevertheless, the statistically significant difference in disclosure rates in the $P_H - P_L < 0$ versus the $P_H - P_L > 0$ cases indicates that the revealed sign of $P_H - P_L$ is one relevant factor shaping the exposure versus cover-up decision, as the model indicates.[16]

In addition, after observing disclosure, opinion '*H*' is prevalent when $P_H - P_L > 0$, while opinion '*L*' is prevalent when $P_H - P_L < 0$ (0.84 versus 0.21 and

[13] For a correct international comparison, we used the highest OECD PPP conversion factor for Chile in the last 5 years (Ch$ 407 per USD$ in 2011). This factor yields an average payment of USD$ 26.

[14] For example, an undergraduate Economics and Business teaching assistant at University of Chile was paid at the time of the experiment CH$ 100,000 for a 4-month academic semester, about CH$ 3000 per hour (at 2 hours per week).

[15] An explanation based on a limited understanding of the experiment by some subjects seems less likely, since there was an individual questionnaire and a questions and answers session before the experiment to ensure that subjects understood the rules of the experiment.

[16] Note also the infrequent choice of 'not sure' of 0.16. However, recall that in a nondisclosure equilibrium as indicated in Result 3 in Section II, '*H*', '*L*' and 'Not sure' (or any randomization of them) are all equally optimal for the public (and therefore all equilibrium behaviour).

Table 4. Theoretical predictions versus experimental evidence, by treatment (meaningful context sessions). Proportions of cases, all rounds

	Theoretical predictions			Experimental evidence					
	$P_H < P_L$	$P_H > P_L$		$P_H < P_L$	$P_H > P_L$	$P_H < P_L$	$P_H > P_L$	Difference[2]	
	Cover-Up Equilibrium	Cover-Up Equilibrium	Disclosure Equilibrium	Revealed (1)	Revealed (2)	Not Revealed (3)	Not Revealed (4)	(2)–(1)	(4)–(3)
SRO									
Disclosure[1]	0	0	1	0.35**	0.88**	0.50	0.73**	0.53**	0.23**
PUBLIC									
If Disclosure observed:									
‘*H*’	0	0	1	0.21	0.84	0.45	0.65	0.63**	0.19**
‘*L*’	1	1	0	0.69	0.10	0.47	0.25	−0.58**	−0.22**
‘Not Sure’	0	0	0	0.10	0.06	0.08	0.10	−0.04†	0.02
Diff. ‘*H*’ – ‘*L*’[2]	−1	−1	1	−0.48 **	0.74 **	−0.02	0.40 **		
If no Disclosure									
‘*H*’	[0,1]	[0,1]	0	0.58	0.12	0.54	0.37	−0.46**	−0.17**
‘*L*’	[0,1]	[0,1]	1	0.26	0.75	0.29	0.39	0.49**	0.10**
‘Not Sure’	[0,1]	[0,1]	0	0.16	0.13	0.17	0.24	0.03	0.07*
Diff. ‘*H*’ – ‘*L*’[2]	[−1,1]	[−1,1]	−1	0.32**	−0.63**	0.25**	−0.02		

Notes: [1]Proportions are statistically different from 0.5 at 10% (†), 5% (*) or 1% (**) level, using a *Z*-test.
[2]Estimated differences are statistically significant at 10% (†), 5% (*) or 1% (**) level, using Fisher's exact test.

0.69 versus 0.1, respectively), and the differences (−0.48 and 0.74) are statistically significant, with signs coherent with the theoretical predictions. Selection of the '*Not sure*' option is minimal, as expected (0.6 in the revealed case). Also, opinion '*L*' after observing no disclosure is statistically more predominant than '*H*' (0.75 versus 0.12) when $P_H - P_L > 0$, as expected in a disclosure equilibrium.

Note that the experimental evidence supports the disclosure equilibrium when $P_H - P_L > 0$, therefore helping to rule out the cover-up equilibrium in Result 2 that is based on rather implausible out-of-equilibrium beliefs.

In conclusion, these results suggest that when $P_H - P_L$ is revealed to the subjects, observed behaviour is generally coherent with Bayesian behaviour and the theoretical implications of the model depending on the revealed sign of $P_H - P_L$.

The situation is different when $P_H - P_L$ is not revealed to the subjects. As column (3) of Table 4 shows, the behaviour of the SRO and the Public when $P_H - P_L < 0$ does not support the expected cover-up equilibrium. On the other hand, when $P_H - P_L > 0$ as in column (4), the proportion of SRO disclosure (0.73) and the difference in the proportions of opinions '*H*' versus '*L*' by the public after observing disclosure (0.4) are statistically significant and consistent with the predicted disclosure equilibrium. However, the difference in the proportions of opinions '*H*' versus '*L*' after observing no disclosure is not significant (−0.02) and therefore inconsistent with a disclosure equilibrium. Thus, when $P_H - P_L$ is not revealed to the subjects, the evidence provides no robust support for the predicted behaviour.

To better understand the differences in behaviour between the treatments throughout the sessions, Tables 5 and 6 show the observed frequencies of play for rounds 1–6, 7–21 and 22–42 for all the treatments. Table 5 shows that when $P_H - P_L$ was revealed, observed behaviour is similar between the different stages of the sessions and consistent with the theoretical predictions (depending on the sign of $P_H - P_L$) even from the outset (as in columns (1) and (4) for rounds 1–6). Moreover, the results also suggest a further gradual evolution of observed behaviour towards the theoretical predictions.

The situation is different in the treatments in which $P_H - P_L$ is not revealed to the subjects (Table 6). The first rounds of these treatments provide an experimental context to analyse how subjects play the game in a context of uncertainty about the sign of $P_H - P_L$, a situation in which subjects would presumably use their prior beliefs about the sign of $P_H - P_L$ to choose their actions. As expected, the evidence shows no significant differences in behaviour in the early rounds 1–6 (columns (1) and (4)) between the treatments with different underlying signs of $P_H - P_L$ (with the weak exception of the frequency of opinion '*L*' after observing no disclosure). However, the evidence of these early rounds does not suggest a clear consensual pattern of how subjects play the game under these conditions of uncertainty. The fact that the proportion of disclosure by the SRO is statistically higher than 0.5 (0.75) would suggest some support for a disclosure equilibrium. However, although opinion '*H*' is more frequent than opinion '*L*' after observing disclosure (0.5 versus 0.36–0.47), the differences are not statistically significant. Note also from columns (1) and (4) that the frequency of opinion '*H*' is also higher than opinion '*L*' when no disclosure is observed (0.48 and 0.56 versus 0.43 and 0.23), which is in effect inconsistent with a disclosure equilibrium. We interpret this mixed evidence as an indication that many implicit beliefs regarding the relative likelihood of fraud detection by the two SRO types seem to coexist across the subjects in the first rounds, and therefore no 'obvious way to play the game' emerges when subjects have no knowledge of the sign of $P_H - P_L$.

This situation evolves as the game unfolds. The coefficients and signs in the last column of Table 6 indicate that in the second half of the sessions (rounds 22–42), there is some partial convergence of observed behaviour towards the equilibria predicted by the model, suggesting some gradual and partial learning of the underlying sign of $P_H - P_L$. The convergence is, however, different depending on the underlying sign of $P_H - P_L$. Column (3) of Table 6 indicates that, even in the second half of the sessions, observed behaviour is not supportive of the expected cover-up equilibrium when $P_H - P_L < 0$: disclosure occurs nearly in half of the situations (0.49), and the proportions of opinions '*H*' and '*L*' are not statistically different (0.37 and 0.51, respectively). Learning seems stronger when $P_H - P_L < 0$. Column (6) of Table 6 shows that observed behaviour in the second half of the sessions is fairly consistent with a disclosure equilibrium: the proportion of disclosure is 0.78 and statistically greater than 0.5, opinion '*H*' is statistically more frequent than '*L*'

Table 5. Experimental evidence, by treatment and rounds (meaningful context sessions). Sign of $P_H - P_L$ revealed to subjects (proportions of cases)

	$P_H < P_L$, Revealed			$P_H > P_L$, Revealed			Differences[2]		
	Rounds 1–6 (1)	Rounds 7–21 (2)	Rounds 22–42 (3)	Rounds 1–6 (4)	Rounds 7–21 (5)	Rounds 22–42 (6)	Rounds 1–6 (4)–(1)	Rounds 7–21 (5)–(2)	Rounds 22–42 (6)–(3)
SRO									
Disclosure[1]	0.29**	0.36**	0.36**	0.73**	0.88**	0.93**	*0.44***	*0.52***	*0.57***
PUBLIC									
If Disclosure observed:									
‘*H*’	0.22	0.21	0.20	0.81	0.78	0.87	*0.61***	*0.57***	*0.67***
‘*L*’	0.65	0.64	0.72	0.11	0.12	0.09	*−0.54***	*−0.52***	*−0.63***
‘Not Sure’	0.13	0.15	0.08	0.07	0.09	0.04	*−0.06*	*−0.06*	*−0.04*
Diff. ‘*H*’ – ‘*L*’[2]	*−0.43***	*−0.43***	*−0.52***	*0.70***	*0.66***	*0.78***			
If no Disclosure:									
‘*H*’	0.50	0.55	0.61	0.11	0.13	0.12	*−0.39***	*−0.42***	*−0.49***
‘*L*’	0.36	0.29	0.22	0.69	0.70	0.79	*0.33***	*0.41***	*0.58***
‘Not Sure’	0.13	0.17	0.17	0.20	0.18	0.09	*0.07*	*0.01*	*−0.08*
Diff. ‘H’ – ‘L’[2]	*0.14**	*0.26***	*0.39***	*−0.58***	*−0.57***	*−0.67***			

Notes: [1] Proportions are statistically different from 0.5 at 10% (†), 5% (*) or 1% (**) level, using a *Z*-test.
[2] Estimated differences are statistically significant at 10% (†), 5% (*) or 1% (**) level, using Fisher's exact test.

Table 6. Experimental evidence, by treatment and rounds (meaningful context sessions). Sign of $P_H - P_L$ not revealed to subjects (proportions of cases)

	$P_H < P_L$, Not revealed			$P_H > P_L$, Not revealed			Differences[2]		
	Rounds 1–6 (1)	Rounds 7–21 (2)	Rounds 22–42 (3)	Rounds 1–6 (4)	Rounds 7–21 (5)	Rounds 22–42 (6)	Rounds 1–6 (4)–(1)	Rounds 7–21 (5)–(2)	Rounds 22–42 (6)–(3)
SRO									
Disclosure[1]	0.75**	0.39†	0.49	0.75**	0.61*	0.78**	0.00	0.22**	0.29**
PUBLIC									
If Disclosure observed:									
‘*H*’	0.50	0.58	0.37	0.50	0.63	0.69	0.00	0.05	0.32**
‘*L*’	0.47	0.39	0.51	0.36	0.22	0.23	0.11	−0.17†	−0.28**
‘Not Sure’	0.03	0.03	0.12	0.14	0.15	0.08	0.11	0.12†	−0.04
Diff. ‘*H*’ – ‘*L*’[2]	0.03	0.19	−0.14	0.14	0.41**	0.46**			
If no Disclosure:									
‘*H*’	0.48	0.52	0.56	0.56	0.42	0.29	0.09	−0.10	−0.27**
‘*L*’	0.43	0.34	0.24	0.23	0.35	0.46	−0.19†	0.01	0.22**
‘Not Sure’	0.10	0.14	0.21	0.20	0.23	0.25	0.11	0.09†	0.04
Diff. ‘*H*’ – ‘*L*’[2]	0.05	0.18**	0.32**	0.33**	0.07	−0.17**			

Notes: [1] Proportions of disclosure are statistically different from 0.5 at the 10% (†), 5% (*) or 1% (**) level, using a *Z*-test.
[2] Estimated differences are statistically significant at 10% (†), 5% (*) or 1% (**) level, using Fisher's exact test.

after observing disclosure (0.69 versus 0.23) and the opposite occurs when no disclosure is observed (0.29 versus 0.46), all of which yields some support for a disclosure equilibrium.[17]

Further evidence of the patterns shown in Tables 4–6 is provided in Table 7, which reports the results of Probit regressions for the SRO's disclosure versus cover-up decision, and multinomial Probit regressions for the Public's opinions about the SRO's type after observing fraud disclosure and no disclosure.[18] The regressors are a dummy variable associated with the underlying sign of $P_H - P_L$, ($D = 1$ if $P_H - P_L > 0$, $D = 0$ otherwise), the block of rounds (0 = first block of 6 rounds, 6 = last block of 6 rounds) and the interaction of the dummy and block variables, and dummies per sessions as controls.[19] Table 7 shows that in sessions in which $P_H - P_L$ was revealed to the subjects, observed behaviour is generally consistent with predicted behaviour even from the early rounds: the dummy coefficient suggests that from the outset, fraud disclosure was statistically less probable when $P_H - P_L < 0$ than otherwise. In addition, the coefficients of the dummy variable in the multinomial Probits for the sessions where $P_H - P_L$ was revealed indicate that the relative likelihood of choosing opinion '*H*' versus '*L*' was lower when $P_H - P_L < 0$, as expected. Also, the relative likelihood of choosing opinion '*H*' versus '*L*' was higher when $P_H - P_L < 0$ after observing no disclosure, again consistent with the predicted behaviour.

The signs and values of the coefficients for the interactive and block variables in the sessions where $P_H - P_L$ was revealed suggest some further convergence towards the predicted equilibria for the SRO and the Public's behaviour as the sessions unfold, which are statistically significant for three of the four coefficients for the Public's opinions.

However, experimental behaviour was different in the sessions in which $P_H - P_L$ was not known to the subjects. The dummy coefficients in Table 7 for these treatments show no systematic effects of the sign of $P_H - P_L$ in the early rounds, either on the probability of disclosure or on the relative likelihood of choosing opinion '*H*' versus '*L*' after observing disclosure, which is consistent with the evidence reported in Table 6 for rounds 1–6. The effect of the sign of

Table 7. Probit and multinomial probit regressions. All meaningful context sessions

	PROBIT		MULTINOMIAL PROBIT		MULTINOMIAL PROBIT	
	Dependent Variable: Disclosure | Poor quality detected		Dependent Variable: Opinion 'H' | Disclosure (Baseline: Opinion 'L' | Disclosure)		Dependent Variable: Opinion 'H' | No Disclosure (Baseline: Opinion 'L' | No Disclosure)	
	P_H, P_L Revealed	P_H, P_L Not revealed	P_H, P_L Revealed	P_H, P_L Not revealed	P_H, P_L Revealed	P_H, P_L Not revealed
D ($P_H < P_L$) [(1)]	−1.79**	0.02	−2.44**	0.12	0.73*	−1.07**
Block	−0.10	0.09*	0.15†	0.11†	−0.09	−0.19**
Block × $D(P_H < P_L)$	−0.07	−0.14*	−0.24*	−0.21*	0.21**	0.30**
Session's Dummies	Yes	Yes	Yes	Yes	Yes	Yes
No. observations	762	619	403	405	1109	819
Log-likelihood	−405.3	−382.3	−258.6	−329.0	−962.4	−826.9

Notes: [(1)]Dummy variable $D = 1$ if $P_H < P_L$, $D = 0$ otherwise.
Estimated coefficients are statistically significant at 10% (†), 5% (*) or 1% (**) level.

[17] One possible interpretation is that situation $P_H - P_L > 0$ may seem to the subjects a more intuitive or plausible situation than $P_H - P_L < 0$, and therefore easier to learn.

[18] The multinomial Probits of Table 7 present estimated coefficients for the effect of a one-unit change in the regressors on the relative log odds of choosing opinion '*H*' versus opinion '*L*', which is defined as the baseline choice. The effects of regressors on the relative risk (the ratio of the probabilities of opinion '*H*' versus opinion '*L*') are obtained by exponentiating the coefficients reported in Table 7.

[19] We also ran Probits without controls and Probits with session, gender and undergraduate program as controls. The coefficients are similar to those reported in Table 7.

$P_H - P_L$ on the likelihood of opinion '*H*' versus '*L*' after observing no disclosure is statistically significant, but with a sign opposite to the expected behaviour.[20]

This reinforces the finding that no obvious way to play emerges under conditions of uncertainty about the relative probabilities of fraud detection of both SRO types.

The sessions where $P_H - P_L$ was not revealed also confirm the partial and slow convergence towards the predicted equilibria suggested by Table 6. In fact, the signs, values and statistical significance of the coefficients of the block and the interactive variables indicate that the SROs and the public behaviour evolves towards the expected behaviour, suggesting some gradual learning by the subjects of the underlying sign of $P_H - P_L$.

Evidence from the neutral-context sessions

Table AI in Appendix reports the aggregate behaviour for the sessions run under a neutral context. Columns (1) and (2) of Table AI show a pattern similar to the evidence in Table 4 for the meaningful context when $P_H - P_L$ was revealed to the subjects, namely that observed behaviour was generally consistent with a cover-up equilibrium when $P_H - P_L < 0$ and a disclosure equilibrium when $P_H - P_L > 0$. However, unlike the meaningful context sessions observed, behaviour in the neutral context sessions was not statistically supportive of any particular equilibrium when $P_H - P_L$ was not revealed to the subjects (columns (3) and (4) of Table AI). This suggests a weaker ongoing learning process in the neutral context sessions. This is corroborated by the Probits and Multinomial Probits in Table AII of Appendix, where the coefficients of the dummies show no effect of the sign of $P_H - P_L$ on disclosure or the Public's opinions after disclosure (as in the meaningful sessions), but the coefficients of the block and interactive variables provide weak evidence of convergence towards predicted equilibria. Thus, the comparison of the neutral versus the meaningful context sessions suggests a weak-context effect that seems to have facilitated a behaviour by the subjects more convergent with equilibrium behaviour. Note, however, that our experimental setting assumed wide differences in detection probabilities (0.2 and 0.8), which facilitates the learning process by the subjects. If the difference of probabilities of detection were smaller, it seems reasonable to believe that subjects would take a longer time to distinguish types and observe some convergence towards equilibrium behaviour.

V. Conclusions

One essential theoretical concern about the reputation-based incentives for effective SR is the ambiguity of how the public would react to fraud disclosure by an SRO and update their beliefs about the SRO's quality (its vigilance level and the quality delivered by its members). This interpretative ambiguity arises because, in principle, fraud disclosure may signal a 'vigilant' attitude by the SRO, but also may signal a 'lax vigilance' attitude, suggesting therefore widespread fraud among its members. The conflict between these opposed yet plausible interpretations of what fraud disclosure reveals about an SRO is the central issue addressed experimentally in this article.

We find that the observed experimental behaviour is generally consistent with Bayesian predictions of how fraud disclosure should be interpreted, provided that subjects are fully informed about the relative likelihoods of fraud detection of the two SRO types: as expected in this (rather unrealistic) context, a 'cover-up equilibrium' is experimentally supported when fraud is known by the subjects to be more likely to be detected by a lax (low vigilance-high fraud) SRO, and a 'disclosure equilibrium' is supported when subjects are informed that a vigilant SRO is more likely to detect fraud. This would suggest that (i) subjects seem generally equipped enough to undertake a Bayesian-consistent updating of their beliefs about SRO quality required for effective SR and (ii) effective SR would occur if the public shared information or beliefs that fraud detection is likelier among 'vigilant' SROs.

However, when subjects are not informed about the underlying fraud detection probabilities by the SRO types, observed SRO and Public behaviour is noisier and not supportive of any equilibrium. The early rounds of these treatments reveal no consensual or 'obvious' way to play the game, suggesting the coexistence among the subjects of opposed interpretations of what disclosure and absence of disclosure reveal about an SRO's underlying quality. Although

[20] However, recall from the model that when $P_H - P_L < 0$, opinions '*L*', '*H*' and '*Not sure*' are equally optimal.

the evidence of the later rounds shows that the subjects in both roles (SRO and Public) eventually begin to learn the underlying relative probabilities of fraud detection by both SRO types and converge towards the expected Bayesian Equilibria, this learning process is partial and slow.

Our results suggest that a first requirement for effective SR, namely that the Public's and SROs' behaviour somehow resemble the Bayesian inference and behaviour required to sustain a consensual interpretation of fraud disclosure and cover-up, could be satisfied. Yet, the bottleneck for effective SR may lie in another hurdle suggested by our results, namely that subjects in the role of the public do not seem to share the belief that fraud is more likely detected and exposed by a 'vigilant' than a 'lax' SRO type, which is required to sustain a consensual interpretation of fraud disclosure as a signal of a 'vigilant' SRO. It seems reasonable to expect that in most real SR situations, the information and beliefs about the relative likelihoods of fraud detection by different SRO types would always be noisy and heterogeneous across individuals. In this context, there would be accordingly many conflicting interpretations and opinions across individuals about what disclosure says about an SRO's expected quality, and therefore the reputational incentives for effective SR in preventing and disclosing fraud would remain inherently ambiguous and weak. This raises the issue of whether and how could the incentives for effective SR be enhanced, for example by means of public parallel regulation of fraud and malpractice (in tandem with SR), and the application of other nonreputation-based incentives such as fines and license revocations whenever fraud is detected, as often suggested by practitioners and the related literature.

Acknowledgements

We are indebted to John Vickers, Meg Meyer and Tim Besley for their valuable comments and insights for previous works. Javier Núñez thanks CILAS at the University of California, San Diego, for the support during the writing stage of this article.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by FONDECYT-Chile [project 1030812].

References

Anderson, C. and Camerer, C. (2000) Experience-weighted attraction learning in sender-receiver signaling games, *Economic Theory*, **16**, 689–718.

Banks, J., Camerer, C. and Porter, D. (1994) An experimental analysis of Nash refinements in signaling games, *Games and Economic Behavior*, **6**, 1–31. doi:10.1006/game.1994.1001

Blume, A., De Jong, D. and Sprinkle, G. (2004) The effect of message space size on learning and outcomes in sender – receiver games, in *Handbook of Results in Experimental Economics*, Plott and Smith (Eds), Elsevier Science. doi:10.1016/S1574-0722(07)00063-7

Brandts, J. and Holt, C. (1992) An experimental test of equilibrium dominance in signaling games, *American Economic Review*, **82**, 1351–65.

Brandts, J. and Holt, C. (1993) Adjustment patterns and equilibrium selection in experimental signaling games, *International Journal of Game Theory*, **22**, 279–302. doi:10.1007/BF01240058

Cadsby, C., Murray, F. and Maksimovic, V. (1998) Equilibrium dominance in experimental financial markets, *Review of Financial Studies*, **11**, 189–232. doi:10.1093/rfs/11.1.189

Carson, J. (2003) Conflicts of interest in self-regulation: can demutualized exchanges successfully manage them?, World Bank Policy Research Working Paper Series, No. 3183. http://dx.doi.org/10.1596%2F1813-9450-3183

Casterella, J., Jensen, K. and Knechel, W. (2009) Is self-regulated peer review effective at signaling audit quality?, *The Accounting Review*, **84**, 713–35. doi:10.2308/accr.2009.84.3.713

Chaudhuri, A. (1998) The ratchet principle in a principal agent game with unknown costs: an experimental analysis, *Journal of Economic Behavior and Organization*, **37**, 291–304. doi:10.1016/S0167-2681(98)00095-X

Cho, I. and Kreps, D. (1987) Signaling games and stable equilibria, *The Quarterly Journal of Economics*, **102**, 179–221. doi:10.2307/1885060

Cooper, D. (2004) Learning in entry limit pricing games, in *Handbook of Results in Experimental Economics*, Plott and Smith (Eds), Elsevier Science. doi:10.1016/S1574-0722(07)00064-9

Cooper, D., Garvin, S. and Kagel, J. (1997a) Signalling and adaptive learning in an entry limit pricing game, *The RAND Journal of Economics*, **28**, 662–83. doi:10.2307/2555781

Cooper, D., Garvin, S. and Kagel, J. (1997b) Adaptive learning vs. equilibrium refinements in an entry limit pricing game, *The Economic*

Journal, **107**, 553–75. doi:10.1111/j.1468-0297.1997.tb00027.x

Cooper, D. and Kagel, J. (2003) The impact of meaningful context on strategic play in signaling games, *Journal of Economic Behavior and Organization*, **50**, 311–37. doi:10.1016/S0167-2681(02)00025-2

Cooper, D., Kagel, J., Lo, W. *et al.* (1999) Gaming against managers in incentive systems: experimental results with Chinese students and Chinese managers, *American Economic Review*, **89**, 781–804. doi:10.1257/aer.89.4.781

Darby, M. and Karni, E. (1973) Free competition and the optimal amount of fraud, *The Journal of Law and Economics*, **16**, 67–88. doi:10.1086/466756

DeMarzo, P., Fishman, M. and Hagerty, K. (2005) Self-regulation and government oversight, *Review of Economic Studies*, **72**, 687–706. doi:10.1111/j.1467-937X.2005.00348.x

DeMarzo, P., Fishman, M. and Hagerty, K. (2007) Reputations, investigations and self-regulation, *mimeo*. Available at http://kelley.iu.edu/Finance/Research/seminarseries/files/fishman07.pdf (accessed 12 May 2015).

Dulleck, U. and Kerschbamer, R. (2006) On doctors, mechanics, and computer specialists: the economics of credence goods, *Journal of Economic Literature*, **44**, 5–42. doi:10.1257/002205106776162717

Emons, W. (1997) Credence goods and fraudulent experts, *The RAND Journal of Economics*, **28**, 107–19. doi:10.2307/2555942

Gamper-Rabindran, S. and Finger, S. (2013) Does industry self-regulation reduce pollution? Responsible care in the chemical industry, *Journal of Regulatory Economics*, **43**, 1–30. doi:10.1007/s11149-012-9197-0

Knechel, W., Krishnan, G., Pevzner, M. *et al.* (2013) Audit quality: insights from the academic literature, *Auditing: A Journal of Practice & Theory*, **32**, 385–421. doi:10.2308/ajpt-50350

Lenox, M. and Nash, J. (2003) Industry self-regulation and adverse selection: a comparison across four trade association programs, *Business Strategy and the Environment*, **12**, 343–56. doi:10.1002/bse.380

Núñez, J. (2001) A model of self-regulation, *Economics Letters*, **74**, 91–97. doi:10.1016/S0165-1765(01)00521-3

Núñez, J. (2007) Can self regulation work?: a story of corruption, impunity and cover-up, *Journal of Regulatory Economics*, **31**, 209–33. doi:10.1007/s11149-006-9020-x

Omarova, S. (2011) Wall street as community of fate: toward financial industry self-regulation, *University of Pennsylvania Law Review*, **159**, 411–92.

Potters, J. and Van Winden, F. (2000) Professionals and students in a lobbying experiment, *Journal of Economic Behavior and Organization*, **43**, 499–522. doi:10.1016/S0167-2681(00)00133-5

Shake, A. and Sutton, J. (1981) The self-regulating profession, *The Review of Economic Studies*, **48**, 217–34. doi:10.2307/2296881

Stephen, F. and Love, J. (1999) Regulation of the legal profession, in *Encyclopedia of Law & Economics*, Bouckaert and De Geest (Eds), Edward Elgar and The University of Ghent. Available at http://encyclo.findlaw.com/5860book.pdf (accessed 12 May 2015).

Taylor, C. (1995) The economics of breakdowns, checkups and cures, *Journal of Political Economy*, **103**, 53–74. doi:10.1086/261975

Wallace, J., Ironfield, D. and Orr, J. (2000) Analysis of market circumstances where industry self-regulation is likely to be most and least effective, in *Consultant's Report for the Taskforce on Industry Self-regulation*, Department of the Treasury, Australian Government. Available at http://goo.gl/e2KHyr (accessed 19 March 2015).

Wolinsky, A. (1993) Competition in a market for informed experts' services, *The RAND Journal of Economics*, **24**, 380–98. doi:10.2307/2555964

Yue, L. and Ingram, P. (2012) Industry self-regulation as a solution to the reputation commons problem: the case of the New York clearing house association, in *The Oxford Handbook of Corporate Reputation*, Pollock and Barnett (Eds), Oxford University Press. doi:10.1093/oxfordhb/9780199596706.001.0001

Appendix

Table AI. Theoretical predictions versus experimental evidence, by treatment (neutral context sessions). Proportions of cases, all rounds

	Theoretical predictions			Experimental Evidence					
	$P_H < P_L$	$P_H > P_L$		$P_H < P_L$	$P_H > P_L$	$P_H < P_L$	$P_H > P_L$	Difference(2)	
	Cover-Up Equilibrium	Cover-Up Equilibrium	Disclosure Equilibrium	Revealed (1)	Revealed (2)	Not Revealed (3)	Not Revealed (4)	(2)–(1)	(4)–(3)
SRO									
Disclosure(1)	0	0	1	0.28**	0.97**	0.54	0.47	0.69**	−0.07
PUBLIC									
If Disclosure observed:									
‘*H*’	0	0	1	0.17	0.91	0.49	0.45	0.74**	−0.03
‘*L*’	1	1	0	0.83	0.06	0.43	0.51	−0.77**	0.08
‘Not Sure’	0	0	0	0.00	0.03	0.08	0.04	0.03*	−0.04
Diff. ‘*H*’ – ‘*L*’(2)	−1	−1	1	−0.66**	0.85**	0.06	−0.06		
If no Disclosure:									
‘*H*’	[0,1]	[0,1]	0	0.48	0.19	0.52	0.37	−0.29**	−0.15**
‘*L*’	[0,1]	[0,1]	1	0.40	0.66	0.27	0.42	0.26 **	0.15**
‘Not Sure’	[0,1]	[0,1]	0	0.12	0.15	0.21	0.21	0.03	0.00
Diff. ‘*H*’ – ‘*L*’(2)	[−1,1]	[−1,1]	−1	0.08	−0.47**	0.25**	−0.05		

Notes: (1)Proportions of disclosure are statistically different from 0.5 at the 10% (†), 5% (*) or 1% (**) level, using a *Z*-test.
(2)Estimated differences are statistically significant at 10% (†), 5% (*) or 1% (**) level, using Fisher's exact test.

Table AII. Probit and multinomial probit regressions. All neutral context sessions

	Probit		Multinomial probit		Multinomial probit	
	Dependent variable: disclosure | poor service detected		Dependent variable: opinion '*H*' | disclosure (baseline: opinion '*L*' | disclosure)		Dependent variable: opinion '*H*' | no disclosure (baseline: opinion '*L*' | no disclosure)	
	P_H, P_L Revealed	P_H, P_L Not revealed	P_H, P_L Revealed	P_H, P_L Not revealed	P_H, P_L Revealed	P_H, P_L Not revealed
$D(P_H < P_L)$ [(1)]	−1.05†	0.45	−2.75**	0.63	1.74**	0.43
Block	0.28	0.01	−0.20	−0.03	0.13	−0.02
Block × $D(P_H < P_L)$	−0.48	−0.13	−0.21	−0.22†	0.15	0.10
Session's dummies	Yes	Yes	Yes	Yes	Yes	Yes
No. observations	241	516	163	268	305	740
Log-likelihood	−74.6	−347.9	−58.2	−227.9	−275.6	−740.8

Notes: [(1)]Dummy variable $D = 1$ if $P_H < P_L$, $D = 0$ otherwise.

Estimated coefficients are statistically significant at 10% (†), 5% (*) or 1% (**) level.