Contract Mix in Franchising as an Efficient Monitoring Device under Asymmetric Information

Thierry PENARD
CREM, University of Rennes I
Faculté des Sciences Economiques, 7 Place Hoche,
35065, Rennes cedex, France
Email: thierry.penard@univ-rennes1.fr

Emmanuel RAYNAUD
INRA SADAPT & Centre ATOM (University of Paris I)
INRA SADAPT, 16 rue Claude Bernard,
75231, Paris cedex 05, France
Email: eraynaud@inapg.inra.fr

Stéphane SAUSSIER
ADIS, University of Paris XI & Centre ATOM, University of Paris I,
Maison des Sciences Economiques, 106-112 Bd de l'Hôpital,
75013, Paris, France
Email: saussier@univ-paris1.fr

March 2005

ABSTRACT:
A strong regularity that emerges from empirical studies on franchising is the coexistence of franchised and company-owned units within the same chain. This paper supports the idea that mixed chains or dual distribution correspond to efficient organization when both the abilities and the behaviors of the managers are costly to observe. We first show that partial monitoring - *i.e*. when the franchisor monitors only a subset of his outlets- is an optimal strategy. Then we prove that the coexistence of franchised and company-owned units economizes on monitoring costs and can be analyzed as an efficient – non transitory – organizational choice.

KEYWORDS: dual distribution, monitoring, asymmetric information.

JEL classification: D23., L14., L22.

1. INTRODUCTION

Franchising has received significant attention in the empirical and theoretical literature on contracting. Most of these papers have focused either on the determinants of contractual provisions, especially royalty rates (Mathewson and Winter, 1985, Lafontaine 1992, Lafontaine and Shaw 1999, Scott 1995, Brickley, 2002) or the tradeoff between franchising and company ownership (Brickley and Dark, 1987, Lafontaine, 1992, Brickley, 1999).

A puzzling empirical regularity, mentioned in many empirical studies, is the coexistence of franchised and company-owned units within chains (sometimes defined as "plural form", "contract mix" or "dual distribution"). Most of the literature deals with the issue of dual distribution as a *transitory phenomena*. For example, it is sometimes argued that franchising is more profitable than company ownership. However, at the beginning of their business, firms operate some units directly either to signal their type to potential franchisees (Gallini and Lutz 1992, and Lafontaine, 1993, for an empirical assessment) or to credibly commit to protecting the value of their brand name (Scott 1995). Consequently the extent of company ownership is likely to decrease with chain maturity. Other authors assert that the rationale for franchising should disappear with chain maturity (Oxenfeldt and Kelly 1969). Franchising only gives rise to temporary access to certain scarce resources, either capital (Caves and Murphy 1976), managerial talent (Norton 1988), or local information (Minkler 1990) that eases their expansion. As firms become established, they should reduce their reliance on franchising (ownership redirection).

These two extreme patterns of franchising evolution (franchising redirection, ownership redirection) have not been empirically validated. Empirical studies using panel data have shown a stability of franchise mix as the chain becomes mature (see for instance Lafontaine and Shaw, 2005, for US and Canadians chains). This suggests that dual

distribution may be an *efficient and persistent* organizational form. Several explanations have been put forward. Bradach (1997) proposed an informal explanation, emphasizing the complementarities between the two contractual arrangements in order to maintain quality and homogeneity of the business concept throughout the units while, at the same time, promoting innovation. He argued that "*chain organizations are more than the sum of their parts: by having both company and franchise arrangements together, a chain can leverage some of the strengths and overcome some of the weakness associated with each arrangement*" (1997, p. 279). Lewin-Solomon (1999) formalized part of Bradach's explanation by justifying the existence of dual distribution as a commitment device used by a franchisor to give franchisees incentives to innovate. In most chains, franchisees pay a royalty rate that is a fraction of their revenues (and not their profit). With this royalty structure, chains will favor innovations that are good for revenues, sometimes at the expense of the franchisees' profits. By owning some units in the chain, the franchisor's interests are more aligned with those of his franchisees. A different line of explanation is provided by Bai and Tao (2000) who constructed a multitask model à la Holmström and Milgrom (1991) to show that dual distribution is an optimal way to induce effort towards both brand name development and sales. Finally, Sorenson and Sorensen (2001) explained franchise mix as an efficient learning structure, resulting from a trade-off between exploration (franchising) and exploitation (company-owned units). Throughout these explanations, dual distribution appears to be an efficient arrangement.

This article provides another rationale for dual distribution. It has long been argued that moral hazard and related monitoring costs are important drivers for the organization of chains. However, they have been mostly used to explain either contract design (Lafontaine, 1992, Mathewson and Winter, 1985, Rubin, 1978) or the make or buy decision (Brickley and Dark, 1987). The premise of this paper is that moral hazard and monitoring costs are also important drivers to explain the prevalence of make **and** buy. Indeed, dual distribution

appears as a monitoring cost-saving strategy to mitigate managers' moral hazard.

To establish this result, we consider a game-theoretic model where a chain has to monitor the managers of its outlets, who are heterogeneous in their abilities to run an outlet. Since the chain imperfectly observes both the managers' ability and behaviour, it is exposed to two kinds of moral hazards, depending on the status of the manager. In a company-owned unit, the manager receives a fixed wage and has a strong incentive to shirk (*i.e*. to provide low effort). By contrast, in a franchised unit, the manager (the franchisee) is the residual claimant and is incited to provide intense efforts. However, the problem of franchisee moral hazard still exists under the form of under-declaration of actual sales. A high ability franchisee can declare a sales level below what his true ability actually allows him to reach. Here, we stand in sharp contrast to the previous literature on franchising that implicitly assumes that individual sales are observable without costs by the chain (see, however, Mathewson and Winter, 1985, and Gal-Or, 1995, for exceptions). In our model, chains must incur a strictly positive cost to assess the true sales of individual units. In this sense, this cost can also be viewed as the expenses of the chain to make the sales or revenues verifiable[1]. Bradach's (1997) work on the restaurant industry partly supports this assumption by claiming that, whereas headquarter management information systems exist for company units, this is not the case for the franchisee: more specifically for the franchised units, "*the only financial information that the chain received was a revenue number each month from which the royalty*

[1] Most of the literature on franchising, more generally on contract theory, disentangles observable variables from non-verifiable ones by assuming that the former are observed without costs whereas the latter are too costly to observe. We could argue that, in the real world, verifiability is a decision for agents based on a costs / benefits analysis. Some variables are costly to observe and verify but this does not necessarily deter parties from contracting on them. Recent literature has begun to explicitly introduce the notion of verifiability as a decision based on the cost and incentives to produce evidence for a third-party enforcer (see for instance Bull, 2001, Bull and Watson, 2004).

was calculated" (1997, p. 288). Combs and al. (2004) also stress this cheating hazard as one of the ways a franchisee can hurt the chain[2].

In our model, the problem of managers' moral hazard (either under-declaration of sales or shirking), can be deterred by controlling the units. However, monitoring the units is costly and this cost increases with the number of controlled units (or with the intensity of monitoring for an individual unit). So it might be sub-optimal to monitor all the units (or to monitor all of them with the same intensity) as it has been showed in the literature on auditing that usually deals with tax evasion, utility regulation or antitrust enforcement (Baron and Besanko 1984, Mookherjee and Png 1989, Khalil 1997). In this literature, the model of Besanko and Spulber (1989) can be fruitful for modeling the organizational design of chains under asymmetric information. The authors characterized the optimal monitoring policy of an antitrust authority that imperfectly observes the production costs of the firms and their pricing behaviors. They proved that it can be optimal to tolerate minor collusion or price agreement to some extent[3]. The antitrust agency simply puts into balance the cost of auditing and the welfare gain of this audit that corresponds to the deterrence of collusive behaviors.

Following this theoretical literature, we examine the optimal monitoring policy when a chain cannot observe both the ability and behavior of its units' managers. The chain can monitor every store or only a subset of stores. When it monitors an outlet, we assume that the control is perfect, *i.e*. it perfectly learns the true performance or the effort level (Gal-Or's model, 1995, has a similar assumption). If a franchised outlet is not monitored, the franchisee can under-declare his sales and save on royalties. If the non-monitored unit is owned by the

[2] The other hazards are respectively (i) not adhering to quality standards, and (ii) releasing the chain's proprietary information.

[3] The optimal policy involves an industry audit when the observed market price is too high (above a threshold price) and no audit when the market price is reasonable (under the threshold price).

chain, the manager may provide a low effort. In both cases, opportunistic managers extract information rents. In such a theoretical framework, the optimal monitoring policy should be conditional on the declared sales (the only freely available information for the franchisor). At the optimum, the chain's owner balances the costs of monitoring – the proportion of units that will be controlled or the monitoring intensity for individual stores - with the amount of rents issued to the uncontrolled units. We first analyze the optimal monitoring policy for "pure" chains (wholly franchised or company-owned chains). Like Besanko and Spulber, we find that complete monitoring is never optimal: it is in the interest of the franchisor to relax inspections on the units that declare high sales. Given this monitoring policy, we also demonstrate that dual distribution is more profitable than a pure franchised system or a pure company-owned system even if all the units are perfectly identical in matter of revenues. A plural form enables the franchisor to monitor his stores more efficiently and to minimize the rents that he is forced to concede to his most talented managers, by reducing the information asymmetry on the abilities and performance of the managers: franchisees' sales declarations convey useful information on the competency of employee managers and symmetrically the observed sales of company-owned outlets provide information on the competency of the franchisees. Such a strategy limits the extent or scope of damageable behaviors or equivalently reduces both the potential gain to under-declaring sales for a franchisee and the potential gain to shirking on sale efforts for a salaried manager. In addition, we also prove that royalty rates and the extent of company-owned units are positively linked and appear to be complementary in the overall organizational design of the chain.

While most of the franchising literature explains the coexistence of franchised and company-owned units by a heterogeneity of outlets such as distance from headquarters (Brickley and Dark, 1987)[4], percentage of repeated business (Brickley, 1999), and location specific factors

[4] Brickley and Dark (1987) used geographical distance as a proxy for direct monitoring costs and showed that

(Chakrabarty and al., 2002), our model develops an explanation based on homogeneous outlets (but with heterogeneous managers). In this respect, our paper is related to Gallini and Lutz's (1992) contribution that analyzed dual distribution as a signaling strategy, and to Bai and Tao's (2000) paper that focused on moral hazard in a multitasking perspective, both considering homogeneous units. However, by focusing on monitoring moral hazard, our rationale for dual distribution is different than in the two previous papers. Our model is also related to Mathewson and Winter (1985) and Gal-Or (1995) who proposed a setting with similar opportunistic behaviors at the downstream level. Both of them relied on the assumption that managers may not truthfully reveal a hidden parameter (namely, the real level of local demand). However, Mathewson and Winter (1985) are mainly concerned with the efficiency of sharing contracts in the context of a single franchisor – franchisee pair and does not deal with dual distribution. More interestingly, Gal-Or (1995) also found that it is sometimes in the interest of the chain to monitor only a subset of outlets. In her paper, the extent of monitoring, *i.e*. monitoring some or all the outlets, is equivalent to the extent of vertical integration since she assumes that monitoring a unit will allow the chain to dictate the desired level of effort. In that sense, her paper also deals with dual distribution. However, her result heavily depends on the outlets' heterogeneity (different level of local demand and local competition) whereas partial monitoring in our case is obtained with homogeneous outlets.

The paper proceeds as follows. Section 2 presents a set of stylized facts about the organizational design of franchise systems. Section 3 presents the theoretical model. Section 4 expounds the main results. Section 5 concludes.

---

outlets far from the franchisor's headquarters tend to be franchised. When monitoring the behavior of managers is costly, chains rely more on the incentive provided by a franchise contract to reduce managerial incentives to shirk (see also, Norton, 1988, Lafontaine, 1992, Bercovitz, 2001, and Lafontaine and Slade, 2002, for a survey).

2. STYLIZED FACTS: THE PERSISTENCE OF DUAL DISTRIBUTION

Several papers have emphasized the fact that many chains used both franchised and company-owned units (Bradach, 1997). Interestingly, this pattern of organization has been found for different countries. For instance, 70% of chains in France have a dual distribution form (Pénard, Raynaud and Saussier, 2003) whereas Furquim de Azevedo and Dos Santos Silva (2001) noted the same proportion for Brazilian chains. In their study of the restaurant industry in the US, both Bradach (1997) and Lewin-Solomon (1999) showed that dual distribution was the dominant form governing chains. However, the extent of company-ownership can vary from chain to chain. This diversity is interesting in itself but beyond the scope of the present paper (see Lafontaine and Shaw, 2005, for more on this issue).

As pointed out in the introduction, several explanations have been given for dual distribution either as a transitory or stable phenomenon. The persistence of dual distribution has, for quite some time, been an empirical puzzle. By studying the evolution and determinants of dual distribution in North America, Lafontaine and Shaw (2005) have largely contributed to the enlightenment of this issue. Strikingly, they found that the proportion of company stores tends to decrease in the first years of franchise and then stabilizes. This stability is interpreted as evidence that chains target a given proportion of company and franchised stores, called the managerial control target. Moreover, this target is rather chain specific, even if it is influenced by the activity sector and the experience acquired before franchising. Lafontaine and Shaw also showed that the value of a brand name increases the proportion of company-owned units in chains. They analyzed it as a way to protect the brand name value from franchisees' opportunistic behavior. By limiting the number of franchised stores, a chain can better protect its brand name investments. Similar empirical patterns have

been observed in other countries (see Furquim de Azevedo and dos Santos Silva, 2001, for Brazil or Pénard, Raynaud, Saussier, 2003, for France). For example, figure 1 represents the evolution of company ownership with respect to franchising experience in the French case[5]. This figure is interesting because it complements previous empirical regularities observed in the US and Canadian market.

[INSERT FIGURE 1]

The shape is rather analog of the one exhibited by Lafontaine and Shaw (2005). After an initial decrease, the extent of company ownership tends to stabilize.

In the next section, we provide a theoretical framework to explain why dual distribution is such a widespread and persistent organizational form in the franchised chain. Our argument is that dual distribution is more efficient than a pure form (either fully franchised or fully company-owned chains) in a context where monitoring is costly and the franchisor cannot observe the behavior and abilities of the managers.

3 THE MODEL

The framework

We consider a multi-store chain in which store managers are heterogeneous in their abilities (ability to generate high sales or to achieve a high level of productivity) and these abilities are imperfectly observed by the chain's owner. However we assume that managers have been initially screened and that poor talented managers have been eliminated. Thus, the chain only accepts efficient managers. However, this doesn't mean that all recruited

[5] Figure 1 is based on the franchise yearbooks published by the French Chamber of Commerce (ACFCI). We have compiled data concerning 745 different franchisors on the 1994-2001 periods and plotted the relation

managers are identical. Some are relatively more talented than the others. Formally, managers' ability is represented by the parameter θ, distributed between $\underline{\theta}$ and $\overline{\theta}$, according to a pdf *f* and cdf. The chain knows that its managers do not all have the same level of competency and is aware of the distribution of this competency among the managers, but it cannot directly observe the exact competency of its individual managers[6]. However, the chain does not really encounter a problem of adverse selection because it is able to screen the applicants and eliminate all managers who do not satisfy a minimum ability $\underline{\theta}$. We assume that above the level $\underline{\theta}$, the franchisor is certain to obtain a positive profit for his store (see assumption 2). This assumption enables us to focus on the moral hazard issues within the chain.

We also assume that all the units belonging to the same chain are identical[7] and we

---

between the years of franchise experience and the percentage of company units.

[6] One can object to the use of discriminatory contracts to incite the managers to reveal their true type, but in the real world, separating contracts are rarely implemented by chains, either for regulatory reasons or because of the complexity in managing too many different contracts. Another argument that could be used for ruling out discriminatory contracts is that, in general, a manager doesn't know exactly what his ability to run a store is before experiencing it on-the-job ability concept). In this case, it would be inefficient to propose self-selection contracts or to organize auctions for the attribution of stores (and the right of residual claimants on these stores). Finally, we can consider that the number of highly talented managers $\overline{\theta}$ is well inferior to the number of outlets to be run. Hence, the chain is forced to recruit relatively less talented managers, with abilities distributed between $\underline{\theta}$ and $\overline{\theta}$.

[7] We could alternatively assume that managers are identical but they run units located in markets that differ in the intensity of demand (with θ the state of the demand encountered by a given store and distributed between $\underline{\theta}$ and $\overline{\theta}$, according to a pdf *f* and cdf F) like in Mathewson-Winter (1985). The results would have been exactly the same (in terms of monitoring policy and dual distribution).

define $V_F(\theta)$ as the expected revenue of a store run by a manager of type θ, with $V_F'(\theta) > 0$ and $V_F''(\theta) < 0$.[8] However, for a store operated by a manager θ, the actual revenue can be less than $V_F(\theta)$, because sales depend not only on ability but also on the efforts provided by the manager. For simplicity, these efforts can take a value between 0 and 1. Let δe be the cost of efforts. Since managers' efforts have a positive impact on the sales, we consider that the actual revenue of a store run by a manager of type θ, exerting an effort e is defined by $eV_F(\theta)$. A manager with ability θ who exerts no effort will achieve no sales whereas the same manager exerting the highest level of effort (e = 1) will generate the maximum revenues $V_F(\theta)$. Moreover, the current operating cost of a store, net of the manager's earnings, is supposed to be constant (independent of sales) and equal to C, whatever the status of the store.

Now let consider the payoffs of the managers and the chain by distinguishing between the franchised and company-owned units.

The franchisee behavior

For the franchised units, we assume that all franchise contracts in a chain are identical and involve the payment of sale-based royalties[9]. This is in conformance with the practices of franchisors: franchisee contract customization is rarely observed in reality (see Bhattacharyya and Lafontaine, 1995, Lafontaine, 1992)[10]. The royalty rate is set by the franchisor and is

[8] Independently of the competency of the managers, $V_F(.)$ should be positively influenced by the value and the reputation of the chain (brand name effect).

[9] We suppose that the franchisor does not claim initial fees from his franchisees.

[10] Several reasons for this absence of contract customization have been provided: high transaction costs of designing and managing several contracts, the potential for franchisor opportunism (McAfee and Schwartz, 1994), low benefits from customization (Bhattacharyya and Lafontaine, 1995).

equal to β. Thus, the profit of a type-θ franchisee is given by $(1-\beta)eV_F(\theta)-C-\delta e$, while the franchisor receives a revenue of $\beta eV_F(\theta)$. With this payoff, individual franchisees can potentially have two types of opportunistic behavior. On the one hand, they can shirk and provide a low effort. This is the main type of opportunistic behavior in the most prevalent models of franchising, namely models based on franchisee moral hazard[11]. On the other hand, a franchisee can also under-declare the true value of sales and save on the amount of royalties paid to the chain (a behavior similar to tax evasion). To focus on the under-declaration of sales as the main opportunistic behavior of franchisees, we rule out franchisees' potential shirking by assuming:

Assumption 1: $(1-\beta)V_F(\underline{\theta}) > \delta$ (1)

For any franchisee, the marginal disutility of efforts is always inferior to his marginal benefits. Under this assumption it is always in the franchisee's best interest to exert the maximum level of effort (e = 1).[12] For franchisees, moral hazard consists only in *under-declaring* sales. This form of opportunism could be costly to detect for the franchisor since a low declaration of sales can come from a low ability franchisee who behaves honestly or from a high ability franchisee who behaves opportunistically. We will further define the cost of monitoring franchisees' sales and how franchisees react if they are likely to be controlled.

The salaried manager behavior

[11] In these models, the severity of the moral hazard problem with the franchisee increases with the royalty rates (Lafontaine, 1992, Lafontaine and Slade, 2002 for a survey).

[12] We ruled out this type of opportunistic behavior not because we consider it as empirically unimportant but because we want to focus on the organizational consequences of the under-declaration hazards. If we relax this assumption, the chain must provide incentives both to mitigate shirking and under-declaration. A way to reinforce this assumption would be to consider that franchisees' effort also affect costs.

Company-owned units are run by salaried managers. Let w denote his wage; this wage is supposed to be fixed (no bonus), regardless of the level of effort provided. For the manager, his utility is $w - \delta e$. If the manager's abilities are θ and his effort is e, then the chain's expected profit is $eV_F(\theta) - w - C$.

Contrary to the franchisee, a salaried manager has strong incentives to provide a minimum level of effort (e = 0). Thus, for salaried managers, moral hazard consists in *shirking*. If the chain is unable to perfectly observe the efforts of its salaried managers, it can use the sales of those units to infer the intensity of efforts. When the chain observes low sales in an outlet, it is impossible to know whether it is a result of low effort (opportunism) or the low ability of its employee. However, even if the salaried managers are tempted to reduce their efforts towards zero, the chain can easily enforce a minimum but positive level of effort. In fact, it is always in the interest of salaried managers to choose a minimum level of effort that leads to sales higher than $V_F(\underline{\theta}) > 0$, because if the chain observes sales lower than $V_F(\underline{\theta}) > 0$, it knows with certainty that the manager has been dishonest and can punish him without any in-depth-control. Furthermore, to focus on the moral hazard, we assume that the chain earns a positive profit even with the less talented managers.

Assumption 2: $V_F(\underline{\theta}) - w - C > 0$

Implicitly, we assume that managers have been initially screened and that poor talented applicants have been eliminated. Thanks to the recruited managers' abilities, the franchisor is certain to make profits in all his outlets since he is always able to enforce a minimum level of sales $V_F(\underline{\theta})$.

Auditing, monitoring and moral hazard

To overcome the two forms of moral hazard from franchisees and salaried managers,

the chain can monitor them, by auditing their accounts and/or by visiting them. We assume that the control is "perfect", *i.e.* an audit enables the franchisor to discover with certainty if the manager has cheated (under-declaration of sales or shirking on efforts, see Gal-Or ,1995, for a similar assumption). However, this control is costly to implement. We assume that the cost of monitoring an outlet is fixed and denoted by K if the store is company-owned and Q if the store is franchised. We can presume that Q<K since sales are probably easier to control than efforts. Moreover, Q may be close to zero for many types of activities (product sales), althought for some activities (services), it is complex and costly to control sales (for example in a fast-food franchise, the franchisee can declare that a certain amount of hamburgers were disposed of and in reality, were sold). However, for the main results of our model to be valid, it is sufficient here to have Q>0 even if Q is close to zero. With a positive cost of monitoring, the chain can be reluctant to monitor all its franchisees and it can be in the chain's best interest to relax monitoring on some franchisees, as we will see later.

Now, what is an efficient monitoring policy for the franchisor? It is always optimal for the chain to design a monitoring policy that takes into consideration the available information coming from its outlets, rather than to opt for a myopic or random policy. Here, the existing information is the declared sales by franchisees or the observed sales from company-owned units. Through these sales, it is as if the manager announces his abilities or talent.

Thus, when a type θ franchisee declares a revenue *R* to the franchisor, it is equivalent to announce that a type θ_R with $\theta_R = V_F^{-1}(R)$ where $V_F^{-1}(.)$ is the inverse revenue function. If the franchisee behaves honestly (declaration of his actual sales), then $\theta_R = \theta$, and if he behaves opportunistically (under-declaration of his sales, *i.e.* $R < V_F(\theta)$), then $\theta_R < \theta$. In any case, the franchisor's revenues are given by $\beta V_F(\theta_R)$.

Likewise, when a type θ salaried manager generates a revenue *R*, it is equivalent to a type θ_E, with $\theta_E = V_F^{-1}(R)$. Thus, if the salaried manager behaves honestly (e=1), we have $\theta_E = \theta$ and $\theta_E < \theta$ if he provides a lower effort.

Hence, the optimal monitoring policy will be conditional on the declared sales θ_R and observed sales θ_E. Let $\chi(\theta_R)$ denote the probability of controlling a franchisee who declares an amount of sales $V_F(\theta_R)$ (*i.e.* who declares competence θ_R) and let $\varpi(\theta_E)$ represent the probability of controlling a salaried manager who achieves an amount of sales $V_F(\theta_E)$ *(i.e.* who declares competence θ_E). If $\chi(\theta_R)$=1, then the franchisee knows that he will be systematically controlled and if $\chi(\theta_R)$=0 he knows that he will never be controlled. In the intermediary case, the franchisee will be inspected with a probability or frequency $\chi(\theta_R)$ [13].

The monitoring policy must also plan for punishment schemes if opportunistic behavior is detected. In company-owned units, if the chain detects an effort below 1, it has the possibility of punishing the manager. These punishments can take different forms ranging from a fine to dismissal. For instance, the termination of a labor contract represents a credible punishment if the manager is paid above the market wage, at an efficiency wage level (Shapiro and Stiglitz, 1984), or if he has invested in specific human capital, whether partly or entirely financed by him[14]. Let *b* represent the penalty for an opportunistic salaried manager. Similarly, if the franchisor detects under-declaration of revenues, he can inflict financial penalties or terminate the franchise contract. Let *a* represent the penalty inflicted on an opportunistic franchisee[15]. In both cases, penalties are supposed to be fixed and independent

[13] An alternative interpretation is to consider that over period T, the franchisee will receive $\chi(\theta_R)$ T visits.

[14] See Krueger (1991) for empirical evidence suggesting that the efficiency wage is effective for employees in company-owned units in the fast food industry.

[15] The termination or the non renewal of the franchise contract appears to be a severe punishment if the

of the damage[16]. Moreover, we suppose that

Assumption 3: $a \geq \beta\left[V_F(\overline{\theta}) - V_F(\underline{\theta})\right]$

Assumption 4: $b \geq \dfrac{\delta\left(V_F(\overline{\theta}) - V_F(\underline{\theta})\right)}{V_F(\overline{\theta})}$.

A3 and A4 indicate that the penalties are sufficiently high to deter opportunistic behavior if a manager is certain to be controlled ($\gamma(\theta_R) = 1$ or $\varpi(\theta_E) = 1$). Theoretically, given the severity of these punishments, any manager can be discouraged from cheating.

Figure 1 displays the timing of the sequential game. In step (1), the franchisor organizes the chain (opting either for a pure system or a plural form) and announces his monitoring policy (that can be different for the franchised units and for the company units). In step (2), given the monitoring policy, the franchisee declares a level of sales $V_F(\theta_R)$ and the salaried manager chooses a level of effort inducing a level of sales $V_F(\theta_E)$. In the last step, the chain implements its monitoring program.

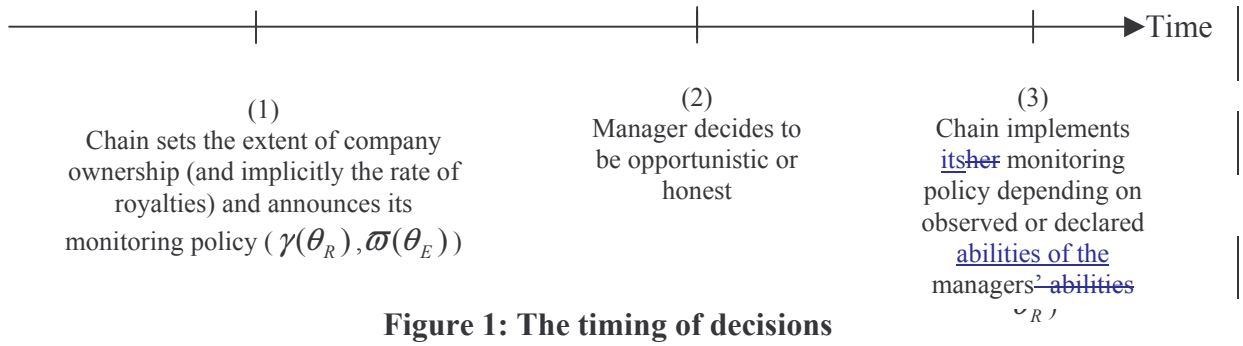


Figure 1: The timing of decisions

---

franchisee holds specific assets or if he earns rents (Kaufman and Lafontaine, 1994). See Klein (1995) and Lafontaine and Raynaud (2002) for insights on self-enforcement in franchising.

[16] a and b can also stand for the maximal liability of managers. If agents are risk neutral, then the optimal monitoring policy involves a penalty at its highest level (i.e. up to the liability: *a* for the franchisee, *b* for the salaried manager), in order to reduce the frequency of controls (that are costly by nature) – see Becker (1968).

In the next section, we solve the strategic game and determine the optimal monitoring policy under the different organizational forms of the chain (*i.e.* pure chain or dual distribution).

4 THE EXTENT OF MONITORING AND OPTIMALITY OF DUAL DISTRIBUTION

4.1 Pure Systems and Complete Monitoring

A pure franchised system

Consider first the case of a pure franchised chain. If the franchisor wants to deter a type θ from declaring θ_R, then he must announce a probability of control (*i.e.* an intensity of control), equal to $\chi(\theta_R)$ such that it is in the interest of the type θ manager to reveal his true type

$$(1-\beta)V_F(\theta) - C - \delta \geq V_F(\theta) - \beta V_F(\theta_R) - C - \delta - \gamma(\theta_R)a \qquad (1)$$

The expression on the left corresponds to the expected payoff of a franchisee who declares his actual sales and the expression on the right reflects his payoff when he behaves opportunistically (declaring $V(\theta_R)$ instead of $V(\theta)$). Using the latter strategy, he can capture a higher part of the total revenues but he runs the risk of being detected and punished by the franchisor ($\gamma(\theta_R)a$). After rearrangement, the intensity of monitoring $\chi(\theta_R)$ should be at least equal to

$$\chi(\theta_R) \geq \frac{\beta(V_F(\theta) - V_F(\theta_R))}{a} \qquad (2)$$

If the franchisor wishes to deter all franchisees from behaving opportunistically, then he must announce a monitoring policy defined for any θ_R by

$$\chi(\theta_R) = \frac{\beta(V_F(\bar{\theta}) - V_F(\theta_R))}{a} \qquad (3)$$

Since the highest talented franchisees (type $\overline{\theta}$) have the highest incentive to under-declare their sales, it is sufficient to deter them from cheating and to prevent all the other types of franchisees θ (with $\theta<\overline{\theta}$) from cheating. This can be easily proven by expressing the expected payoff for cheating for a type θ

$$V_F(\theta) - \beta V_F(\theta_R) - C - \delta - \left(\frac{\beta\left(V_F(\overline{\theta}) - V_F(\theta_R)\right)}{a} \right) a \qquad (4)$$

After simplification, (4) is equal to $V_F(\theta) - \beta V_F(\overline{\theta}) - C - \delta$, a gain well below what an honest franchisee could expect (*i.e.* $V_F(\theta) - \beta V_F(\theta) - C - \delta$ since V_F is an increasing function of θ).

A pure company-owned system

For salaried managers, the monitoring policy is different. If the owner wants to deter a manager θ from cheating, he has to control him with a probability of $\varpi(\theta_E)$, defined by

$$w - \delta \geq w - \delta e - \varpi(\theta_E) b \qquad \text{with } e = \frac{V_F(\theta)}{V_F(\theta_E)} \qquad (5)$$

After rearrangement, the intensity of monitoring $\varpi(\theta_E)$ should be at least equal to

$$\frac{\delta\left(V_F(\theta) - V_F(\theta_E)\right)}{b V_F(\theta)} \qquad (6)$$

If the owner wants to deter all salaried managers from behaving opportunistically, then he must implement a monitoring policy $\varpi(\theta_E)$ characterized for any θ_E by

$$\varpi(\theta_E) = \frac{\delta\left(V_F(\overline{\theta}) - V_F(\theta_E)\right)}{b V_F(\overline{\theta})} \qquad (7)$$

With such a monitoring intensity, the highest talented managers (type $\overline{\theta}$) are deterred from shirking, but so are all the less talented types with $\theta<\overline{\theta}$. Indeed, from (7) the expected payoff of an opportunistic salaried manager is equal to

$$w-\delta\frac{V_F(\theta_E)}{V_F(\theta)}-\left(\frac{\delta\left(V_F(\bar{\theta})-V_F(\theta_E)\right)}{bV_F(\bar{\theta})}\right)b \qquad (8)$$

This expression is equivalent to $w-\delta-\delta\left(\frac{V_F(\theta_E)}{V_F(\theta)}-\frac{V_F(\theta_E)}{V_F(\bar{\theta})}\right)$, a payment inferior to $w-\delta$, (the earnings of the honest manager), since $V_F(\theta)<V_F(\bar{\theta})$.

Now we will examine whether monitoring all the stores is an optimal policy. We start with the case of a vertically integrated chain.

4.2 Pure Company-owned Chain and Partial Monitoring

Following Besanko and Spulber (1989), we consider a two-part monitoring program defined by

$$\text{if} \quad \begin{array}{lll} \theta_E>\hat{\theta}, & \text{then} & \varpi(\theta_E)=0 \\ \theta_E\le\hat{\theta}, & \text{then} & \varpi(\theta_E)=\delta\dfrac{\left(V_F(\hat{\theta})-V_F(\theta_E)\right)}{bV_F(\hat{\theta})} \end{array} \qquad (9)$$

A complete control would consist in setting $\hat{\theta}=\bar{\theta}$ and a partial control is characterized by $\hat{\theta}<\bar{\theta}$. In the latter case, the franchisor does not control the salaried managers who achieve sales superior to $V_F(\hat{\theta})$, but managers who generate revenues below $V_F(\hat{\theta})$ are controlled with a probability that tends to increase as the level of observed sales declines. Thus, the chain concentrates its monitoring activity on the less profitable units (those generating sales below $V_F(\hat{\theta})$)[17].

Given such a monitoring policy, managers with talent $\theta\le\hat{\theta}$ are deterred from shirking and will always choose $\theta_E=\theta$ (*i.e.* an effort e=1). Managers with talent $\theta>\hat{\theta}$ are aware that they

[17] Let us recall that all the units are homogeneous and differ only in the abilities and efforts of their managers. Two units run by managers with identical talent should generate identical revenues.

are never controlled as long as they achieve an amount of sales superior or equal to $V_F(\hat{\theta})$. Therefore it is in their best interest to choose the exact level of effort $e = \frac{V_F(\hat{\theta})}{V_F(\theta)}$ that enables them to reach the amount of revenues $V_F(\hat{\theta})$[18]. With partial monitoring, it is as if the owner offers a rent to the more talented managers, (allowing them to exert a lower effort). The distribution of the performances of the stores of a chain tends to be upward bounded with a concentration of stores displaying the same sales ($V_F(\hat{\theta})$).

Given the previous monitoring program, the next proposition shows that a partial monitoring (*i.e.* if $\hat{\theta} < \overline{\theta}$) is an optimal strategy.

Proposition 1. *In a pure company-owned chain, it is never optimal to control all the units.*

Proof: The optimal monitoring policy is determined by maximizing the chain's profit with respect to the monitoring threshold $\hat{\theta}$

$$G = \int_{\underline{\theta}}^{\hat{\theta}} (V_F(\theta) - w - C - \varpi(\theta)K) f(\theta) d\theta + \int_{\hat{\theta}}^{\overline{\theta}} (V_F(\hat{\theta}) - w - C) f(\theta) d\theta$$

The derivative with respect to $\hat{\theta}$ is given by

$$\frac{\partial G}{\partial \hat{\theta}} = \left[V_F(\hat{\theta}) - w - C - \varpi(\hat{\theta})K - V_F(\hat{\theta}) + w + C\right] f(\hat{\theta}) + V'_F(\hat{\theta}) \int_{\hat{\theta}}^{\overline{\theta}} f(\theta) d\theta + \frac{\partial \varpi}{\partial \hat{\theta}} K \int_{\underline{\theta}}^{\hat{\theta}} f(\theta) d\theta$$

---

[18] We can easily prove that a salaried manager with a type $\theta > \hat{\theta}$, can never expect a higher payoff by exerting an effort below $e = \frac{V_F(\hat{\theta})}{V_F(\theta)}$, since we have $w - \delta \frac{V_F(\theta_E)}{V_F(\theta)} - \left(\frac{\delta (V_F(\hat{\theta}) - V_F(\theta_E))}{b V_F(\hat{\theta})} \right) b < w - \delta \frac{V_F(\hat{\theta})}{V_F(\theta)}$ where the expression on the left represents his expected benefit from cheating by choosing a level of effort $\theta_E < \hat{\theta}$ and the expression on the right the expected benefit if he chooses $\theta_E = \hat{\theta}$

After rearrangement:

$$\frac{\partial G}{\partial \hat{\theta}} = -\int_{\underline{\theta}}^{\hat{\theta}} \frac{\delta V_F^{'}(\hat{\theta}) V_F(\theta) K}{b\left(V_F(\hat{\theta})\right)^2} f(\theta) d\theta + V_F^{'}(\hat{\theta}) \int_{\hat{\theta}}^{\bar{\theta}} f(\theta) d\theta$$

The first expression has a negative sign and corresponds to the *monitoring cost effect*: a higher $\hat{\theta}$ increases the cost of supervising the salaried managers. The second expression has a positive sign and represents the *incentive effect*: *i.e*. the positive impact of more intense controls on salaried managers, leading to higher efforts, sales and profits.

Full monitoring involves setting a threshold $\hat{\theta} = \bar{\theta}$ (all the company-owned units are controlled regardless of the level of sales declared). However, such a policy is not optimal since

$$\left.\frac{\partial G}{\partial \hat{\theta}}\right|_{\hat{\theta}=\bar{\theta}} = -\int_{\underline{\theta}}^{\bar{\theta}} \frac{\delta V_F^{'}(\bar{\theta}) V_F(\theta) K}{b\left(V_F(\bar{\theta})\right)^2} f(\theta) d\theta < 0$$

The chain could obtain higher profits by relaxing its control on the most efficient units: from the situation $\hat{\theta} = \bar{\theta}$, reducing $\hat{\theta}$ slightly enables the chain to significantly lower its monitoring costs (first order effect), without reducing the incentive for its managers to behave honestly (second order effect).

Besides, we find that a laissez-faire policy ($\hat{\theta} = \underline{\theta}$) is not optimal

$$\left.\frac{\partial G}{\partial \hat{\theta}}\right|_{\hat{\theta}=\underline{\theta}} = V_F^{'}(\underline{\theta}) > 0$$

By increasing $\hat{\theta}$, the chain can stimulate the efforts of all its managers (first order effect) and the expected benefits largely compensate for the additional cost of monitoring. □

Proposition 1 can be explained by the fact that partial deterrence allows for a significant reduction in monitoring costs. Even if the revenues are less than expected in the stores run by the more talented managers, this is more than counterbalanced by the savings on

monitoring costs. Note that the "informational rent" captured by the high-ability managers is an increasing function of monitoring costs and a decreasing function of penalty severity (as $\hat{\theta}$ decreases in *K* and increases in *b*). Thus, the proportion of controlled units (units that can receive a visit with a positive probability) should be close to 100% when monitoring costs decline towards zero or when the punishments become extremely severe. However, by strengthening the penalties, we have two opposite effects on the intensity of control at the individual level. The units that were initially subject to a probabilistic control are now less intensively monitored (Beckerian effect resulting from the trade-off between the probability of control and the amount of penalty), whereas the units that were initially out of control are now likely to be monitored (even if the probability is infinitesimal for the high talented managers). However, even in these extreme conditions (severe punishments, monitoring cost nearing zero), it is always in the best interest of the owner to give up auditing the most profitable units since it serves no purpose.

4.3 Pure franchised Chain and Partial Monitoring

Now let us consider the monitoring program for a pure franchised system as follows:

$$\begin{array}{lll} \theta_R > \hat{\hat{\theta}}, & then & \gamma(\theta_R) = 0 \\ if & & \\ \theta_R \le \hat{\hat{\theta}}, & then & \gamma(\theta_R) = \beta \dfrac{\left(V_F(\hat{\hat{\theta}}) - V_F(\theta_R)\right)}{a} \end{array} \qquad (10)$$

Full monitoring involves setting $\hat{\hat{\theta}} = \overline{\theta}$. When $\hat{\hat{\theta}} < \overline{\theta}$, the franchisor will choose partial monitoring and offers a rent to some of his more talented franchisees. Formally, all the franchisees with a type $\theta > \hat{\hat{\theta}}$ can under-declare their revenues ($V(\hat{\hat{\theta}})$ instead of $V(\theta)$) and thus pay lower royalties than normally required ($\beta V_F(\hat{\hat{\theta}})$ instead of $\beta V_F(\theta)$). As long as they

pay royalties superior or equal to $\beta V_F(\hat{\hat{\theta}})$, they are not controlled. So it is profitable for them to declare revenues equal to $V(\hat{\hat{\theta}})$[19]. Conversely, all franchisees with a type $\theta \leq \hat{\hat{\theta}}$ are audited and have strong incentives to be honest (to declare their actual revenues).

Proposition 2. *In a pure franchised chain, it is never optimal to audit all the stores.*

Proof: The optimal intensity of monitoring $\hat{\hat{\theta}}$ is given by maximizing the franchisor's profit with respect to the managerial threshold monitoring $\hat{\hat{\theta}}$

$$Max_{\left\{\hat{\hat{\theta}}\right\}} G = \int_{\underline{\theta}}^{\hat{\hat{\theta}}} (\beta V_F(\theta) - \gamma(\theta)Q) f(\theta) d\theta + \int_{\hat{\hat{\theta}}}^{\overline{\theta}} \beta V_F(\hat{\hat{\theta}}) f(\theta) d\theta$$

The derivative of the profit with respect to $\hat{\hat{\theta}}$ is as follows

$$\frac{\partial G}{\partial \hat{\hat{\theta}}} = -\gamma(\hat{\hat{\theta}})Qf(\hat{\hat{\theta}}) - \beta \frac{V'_F(\hat{\hat{\theta}})}{a} Q \int_{\underline{\theta}}^{\hat{\hat{\theta}}} f(\theta) d\theta + \beta V'_F(\hat{\hat{\theta}}) \int_{\hat{\hat{\theta}}}^{\overline{\theta}} f(\theta) d\theta$$

The two first expressions represent the negative monitoring cost effect when the franchisor decides to strengthen controls on his franchisees (to increase $\hat{\hat{\theta}}$). The third expression corresponds to the incentive effect (more intense controls limit opportunistic behaviors and incite franchisees to declare their actual sales).

---

[19] For franchisees with type $\theta > \hat{\hat{\theta}}$, the benefit from declaring revenues $V_F(\theta_R)$ below $V_F(\hat{\hat{\theta}})$ are given by:

$V_F(\theta) - \beta V_F(\theta_R) - C - \delta - \left(\frac{\beta\left(V_F(\hat{\hat{\theta}}) - V_F(\theta_R) \right)}{a} \right) a = V_F(\theta) - \beta V_F(\hat{\hat{\theta}}) - C - \delta$, so they are perfectly indifferent when declaring between $V_F(\hat{\hat{\theta}})$ and lower revenues.

Complete monitoring consists in setting $\hat{\hat{\theta}} = \overline{\theta}$ (all the franchisees will be audited regardless of their declared sales). However, such a policy is not optimal, since the franchisor can increase his profit by relaxing monitoring efforts on the most efficient units. Indeed, marginal profit decreases with $\hat{\hat{\theta}}$ at $\hat{\hat{\theta}} = \overline{\theta}$:

$$\left.\frac{\partial G}{\partial \hat{\hat{\theta}}}\right|_{\hat{\hat{\theta}} = \overline{\theta}} = -\beta \frac{V_F^{'}(\overline{\theta})}{a} Q < 0$$

We can also see that a "laissez-faire" policy $\hat{\hat{\theta}} = \underline{\theta}$ is sub-optimal

$$\left.\frac{\partial G}{\partial \hat{\hat{\theta}}}\right|_{\hat{\hat{\theta}} = \underline{\theta}} = \gamma(\underline{\theta}) Q f(\underline{\theta}) + \beta V_F^{'}(\underline{\theta}) > 0 \quad \square$$

The rationale for partial monitoring in a pure franchised system is the same as in a company-owned system. The existence of monitoring costs incites the chain to refrain from auditing the most profitable franchised units. Thus, a franchisee with a type $\theta > \hat{\hat{\theta}}$ benefits from an "informational rent" equal to $\beta\left(V_F(\theta) - V_F(\hat{\hat{\theta}})\right)$. The greater the franchisee's abilities, the higher the rent he may receive. Offering a rent to the most talented franchisees is a low-cost method for the franchisor to deter them from engaging in more harmful cheating (the more detrimental action being to declare the lowest level of sales $V_F(\underline{\theta})$). We notice that partial monitoring leads the franchisor to audit only honest franchisees, but it is absolutely necessary to comply with this if the chain wants to prevent the high-talented franchisees from passing themselves off as low-talented franchisees (by declaring low revenues).

Until now, we have demonstrated that complete monitoring is never optimal in a pure

system. Let us turn to the issue of dual distribution. Is a mix chain more or less efficient than a pure system in containing moral hazards from managers?

4.4 Dual Distribution and Efficiency

A Measure of the Company Ownership

As we consider that managers have the opportunity to apply for the most profitable status whenever a chain offers several positions, the more talented managers should choose to become franchisees and the less talented should opt for a salaried manager status; the earnings of a franchisee increase with his talent, whereas the gains of an honest manager are independent of his talent.

Let us define the manager with talent $\tilde{\theta}$ as the type indifferent to being a franchisee or salaried manager. If such a type exists, then all the managers with talent $\theta>\tilde{\theta}$ should preferably choose a franchisee status and all the managers with talent $\theta\leq\tilde{\theta}$ should choose a salaried position. The type $\tilde{\theta}$ is defined by

$$w-\delta\frac{V_F(\hat{\theta})}{V_F(\tilde{\theta})}=(1-\beta)V_F(\tilde{\theta})-C-\delta \qquad (11)$$

The expression on the left corresponds to the gain of a salaried manager with talent $\tilde{\theta}$, who provides an effort $e=\frac{V_F(\hat{\theta})}{V_F(\tilde{\theta})}<1$, because he is one of the most talented employee managers and knows that he will never be controlled. The expression on the right corresponds to the profit of a type $\tilde{\theta}$ franchisee who behaves honestly and knows for a fact that he will be controlled, since he declares the lowest revenues of all the franchised units. If $\tilde{\theta}$ increases, the franchisees will be less numerous inside the chain, whereas a decrease of $\tilde{\theta}$ indicates a larger number of managers will choose franchisee status. Thus, choosing the

proportion of company-owned units is equivalent to choosing the threshold type $\tilde{\theta}$.

Moreover, modifying the intensity of control in the company-owned stores $\hat{\theta}$ or the level of royalty rates β, has an impact on the extent of company ownership ($\tilde{\theta}$). Indeed, equation (11) shows that $\tilde{\theta}$, $\hat{\theta}$ and β are linked. Therefore, the franchisor has three interdependent strategic variables through which he can improve the chain's efficiency: the monitoring intensity $\hat{\theta}$, the level of royalty rates β and the proportion of company-owned stores $\tilde{\theta}$. If we fix one of the three variables, then the other two appear to be positively linked. This leads to an interesting corollary:

Corollary. *Assuming that the monitoring intensity of salaried managers is invariable, then royalty rates (β) and the percent of company-owned units ($\tilde{\theta}$) tend to be positively linked and can be used as complementary variables by the franchisor.*

This result gives certain theoretical foundations to evidence of a positive relationship between royalty rates and the proportion of company-owned units exhibited in several previous empirical studies (Lafontaine 1992, Lafontaine and Shaw 1999, Pénard, Raynaud, Saussier, 2003). This suggests that these two instruments may be complementary in the governance of chains[20], confirming the findings of Bradach (1997) regarding restaurant chains.

Now, we must define the monitoring policy and the profits of the franchisor, based on the three strategic variables $(\tilde{\theta},\hat{\theta},\beta)$.

Monitoring Policy under Dual Distribution

If $\tilde{\theta}$ is the type indifferent to being a salaried manager or a franchisee, then the monitoring policy $(\varpi(\theta), \gamma(\theta))$ announced by the franchisor should be as follows:

$$
\begin{array}{lll}
 & \theta \in \left[\underline{\theta}, \hat{\theta}\right] & \text{then} \quad \varpi(\theta) = \delta \dfrac{\left(V_F(\hat{\theta}) - V_F(\theta)\right)}{bV_F(\hat{\theta})} \\
\text{if} & \theta \in \left[\hat{\theta}, \tilde{\theta}\right] & \text{then} \quad \varpi(\theta) = 0 \\
 & \theta \in \left[\tilde{\theta}, \hat{\hat{\theta}}\right] & \text{then} \quad \gamma(\theta) = \beta \dfrac{\left(V_F(\hat{\hat{\theta}}) - V_F(\theta)\right)}{a} \\
 & \theta \in \left[\hat{\hat{\theta}}, \overline{\theta}\right] & \text{then} \quad \gamma(\theta) = 0
\end{array}
\tag{12}
$$

The franchisor implements a four-part monitoring program as displayed in Figure 2. Part I refers to the monitoring of the less talented salaried managers and part II to the higher talented salaried managers. Part III applies to the less talented franchisees and part IV to the high ability franchisees.

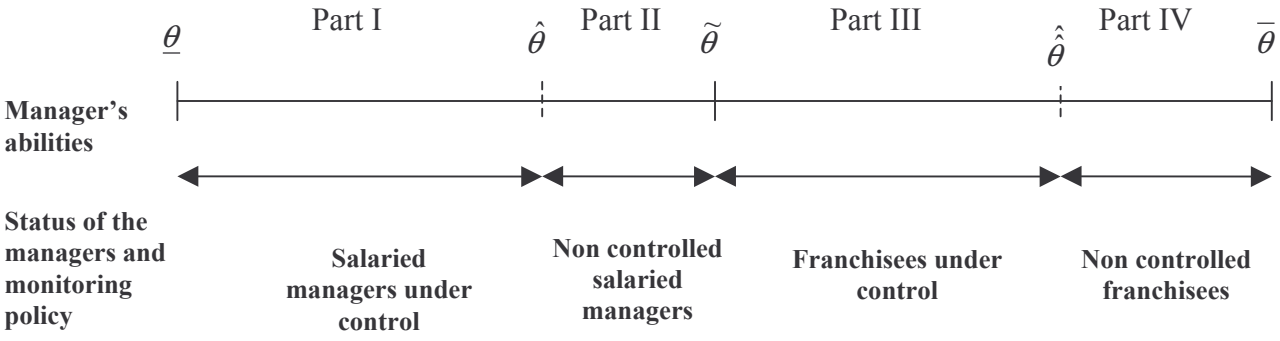


Figure 2. The monitoring program of the franchisor in a mixed chain

Given the distribution of the manager's abilities within the chain, the profit function of the mixed chain is given by

[20] This can be also related to the works of Milgrom and Roberts (1995) regarding strategic complementarities.

$$\max_{\{\tilde{\theta},\hat{\theta},\hat{\hat{\theta}}\}} = \int_{\underline{\theta}}^{\hat{\theta}} \left(V_F(\theta) - w - C - \varpi(\theta)K\right) f(\theta)d\theta + \int_{\hat{\theta}}^{\tilde{\theta}} \left(V_F(\hat{\theta}) - w - C\right) f(\theta)d\theta$$
$$+ \int_{\tilde{\theta}}^{\hat{\hat{\theta}}} \left(\beta V_F(\theta) - \gamma(\theta)Q\right) f(\theta)d\theta + \int_{\hat{\hat{\theta}}}^{\bar{\theta}} \beta V_F(\hat{\hat{\theta}}) f(\theta)d\theta \qquad (13)$$

$$\text{s.t. } w - \delta\frac{V_F(\hat{\theta})}{V_F(\tilde{\theta})} = (1-\beta)V_F(\tilde{\theta}) - C - \delta$$

If an interior solution exists, the proportion of company ownership that maximizes the chain's revenue is given by

$$\frac{\partial G}{\partial \tilde{\theta}} = \left(V_F(\hat{\theta}) - w - C\right) f(\tilde{\theta}) - \left[\beta V_F(\tilde{\theta}) f(\tilde{\theta}) - \gamma(\tilde{\theta})Qf(\tilde{\theta})\right] = 0 \qquad (14)$$

The choice of $\tilde{\theta}$ results from a comparison between the profit obtained from a company-owned unit run by a (non monitored) manager, who is the most highly talented among salaried managers[21], and the royalties perceived from a (monitored) franchisee, who is the less talented among all the franchisees, minus the cost of supervising him (since the franchisor must deter the other franchisees from declaring sales $V_F(\tilde{\theta})$)[22]. The optimal level of franchise mix is such that the gain from an additional franchised unit is equal to the profit from an additional company-owned unit (the franchisor is indifferent to the two contractual arrangements). In the two following propositions, we prove that an interior solution to the optimisation program of the chain's owner always exists.

Proposition 3. *Under asymmetric information, if a chain is completely company-owned, then it is optimal to franchise some units*

Proof: As $\gamma(\underline{\theta}) = 0$ (if we introduce a franchisee at the margin, there is no need for

[21] For this reason, he will not be monitored, the owner preferring to issue him a rent.

[22] Indeed the existing franchisees will attempt to be taken to be the lowest talented franchisee.

monitoring since his competency is perfectly known) and according to equation (14),

$$\left.\frac{\partial G}{\partial \tilde{\theta}}\right|_{\tilde{\theta}=\bar{\theta}} = \left(V_F(\hat{\theta}) - C - w\right) f(\bar{\theta}) - \beta V_F(\bar{\theta}) f(\bar{\theta})$$

This derivative corresponds to the difference of incomes between a company-owned store and a franchised store, both run by the highest talented managers.

As $w - \delta \frac{V_F(\hat{\theta})}{V_F(\bar{\theta})} = (1-\beta) V_F(\bar{\theta}) - C - \delta$, then

$$\left.\frac{\partial G}{\partial \tilde{\theta}}\right|_{\tilde{\theta}=\bar{\theta}} = -\left(V_F(\bar{\theta}) - V_F(\hat{\theta})\right)\left(1 - \frac{\delta}{V_F(\bar{\theta})}\right)$$

From assumption 1, we know that $\delta < V_F(\bar{\theta})$.

So $\left.\frac{\partial G}{\partial \tilde{\theta}}\right|_{\tilde{\theta}=\bar{\theta}} < 0$

A pure company-owned chain is never optimal. □

By allowing a few franchisees to integrate his chain, the franchisor reduces his information asymmetry on managers' competency. Indeed, he knows that the recruited franchisees have a higher degree of competence distribution (limited information asymmetry). As a result, it will not be necessary for the franchisor to perform intense monitoring to discipline them. Moreover, these franchisees will also exert pressure on the opportunistic behaviors of the salaried managers and enable the chain's owner to reduce the rent that he is forced to give them. For these reasons, the chain will save on monitoring costs by transforming some company-owned units into franchised units. This organizational switch also implies setting reasonable royalty rates (because if royalties are too high, the applicants will prefer the status of salaried manager)[23].

[23] The lowering of the royalty rates attracts and leads to the recruiting of high talented franchisees.

Proposition 4. *Under asymmetric information, if a chain is completely franchised, then it is optimal to convert some franchised units into company-owned units*

Proof: According to equation (14),

$$\left.\frac{\partial G}{\partial \tilde{\theta}}\right|_{\tilde{\theta}=\underline{\theta}} = (V_F(\underline{\theta}) - C - w) f(\underline{\theta}) - \left(\beta V_F(\underline{\theta}) f(\underline{\theta}) - \gamma(\underline{\theta}) Q f(\underline{\theta})\right)$$

As previously shown, this derivative corresponds to the difference of incomes between a company-owned store and a franchised store, both run by the lowest talented managers.

As $w - \delta \frac{V_F(\underline{\theta})}{V_F(\underline{\theta})} = (1-\beta) V_F(\underline{\theta}) - C - \delta$, then we obtain

$$\left.\frac{\partial G}{\partial \tilde{\theta}}\right|_{\tilde{\theta}=\underline{\theta}} = \gamma(\underline{\theta}) Q f(\underline{\theta}) > 0 \qquad \square$$

The role of company units is to limit the opportunism of franchisees and to prevent them from declaring low sales. Thus, by converting some franchises into company units, the franchisor can save a certain amount on monitoring costs.

From the two previous propositions, we can directly infer the key proposition of this model: a dual system is more efficient than a pure franchised system or a pure company-owned system.

Proposition 5: when a chain cannot perfectly observe managers' abilities and behaviors, then it is always optimal to hire both franchised and salaried managers.

The franchisor has an ideal level of franchising mix ($\tilde{\theta}$) that enables him to maximize profits. This company ownership target always belongs to the interval $]\underline{\theta}, \bar{\theta}[$ and should

strongly depend on two key variables: royalty rates and the monitoring technology. As mentioned earlier, the ideal proportion of company-owned units should increase with royalty rates (complementarity). As well, the extent of company ownership should increase with the monitoring costs of franchisees and decrease with the severity of punishments that the chain can inflict on opportunistic franchisees. Indeed, a chain will call for more intensive franchising when it has better technology to monitor its franchised units. Conversely, the proportion of company-owned units should decrease with the monitoring costs of salaried managers and increase with the severity of punishments that the chain can inflict on opportunistic franchisees.

4.5. Discussion

Our theoretical framework gives some insight into the predominance and the efficiency of dual distribution in franchising. Our model supports that the rationale for dual distribution strongly relies on the monitoring policy. This is rather original with regard to the previous literature. Part of this originality comes from the assumption that the franchisor cannot freely observe the actual level of a franchisee's sales and that franchisees have an incentive to under-declare their actual revenues in order to minimize the royalties claimed by the franchisor. This type of opportunistic behavior is similar to that studied in Mathewson and Winter (1985) where a franchisee can under-declare his sales. However, in their model, the possibility of franchisee cheating arises because the franchisor cannot identify the realized state of demand without incurring costs. In our model, franchisees are tempted to cheat because the franchisor must incur a cost to control the actual sales in the franchised outlets. However, this doesn't imply that franchisees will indeed under-declare their sales. Their conduct will depend on the risk to be detected and punished. Some may decide to actually

cheat (if they think that the gain is higher than the expected cost) and others may behave honestly. Even if all the franchisees are tempted to cheat, revenue chiseling will be *de minimis* at the equilibrium as long as the monitoring cost is low.

Our model may generate several implications. First, highly talented managers should opt for franchisee status and the less-talented managers should become salaried employees. Since the sale level is an increasing function of ability, franchisees should have higher sales than vertically-integrated units (and higher profits because we assumed that the costs are the same regardless of the organizational form). Shelton's (1967) study, as quoted in Bai and Tao (2000), supports our prediction. However, other papers found the opposite result, namely, sales at company-owned units are greater than sales in franchised units (Martin, 1988, Lafontaine, 1992). While this last result seems to refute our prediction, it can reflect the units' heterogeneity. Chains could decide to own the units with the more "profitable" market or the greater potential. A more accurate empirical test for the predictions of our model should control for heterogeneity at the outlet level or compare company-owned and franchised units with identical characteristics.

Second, if the effort level of the downstream unit is not critical for the individual profits, then the shirking hazard is a less important concern and we should observe more vertical integration in order to concentrate on the mitigation of the under-declaration hazard. This seems to be consistent with empirical results. For instance, Lafontaine (1992) found that the proportion of franchised units tends to rise with the "level" of the franchisee's effort and, conversely, that the extent of vertical integration is more prevalent for chains where the franchisee's effort is relatively less important. Similarly, if the prospect for under-declaration of sales is weaker than the shirking problem or is easily controlled, then we should observe a greater reliance on franchising.

Finally, the monitoring intensity should also depend on the contractual provisions.

For instance, it has been argued in the literature that particular provisions such as tying may lower monitoring costs or improve quality control (Klein and Saft, 1985, Michael, 2000). This can be illustrated by the case where the chain provides one or several inputs to franchisees. If the production function at the store is of fixed-proportion, then the chain can infer the true level of sales from the quantity of inputs sold to a particular outlet (i.e. the chain can infer the true level of output and avoid "too large" a discrepancy between output and declared sales). As the discretional behavior of franchisees is limited, we expect the monitoring intensity of franchised units to be lower with greater reliance on franchising in a chain with a tying provision.

5. CONCLUSION

Most of the existing literature on franchising considers dual distribution as a transitory phenomenon. Our paper suggests that dual distribution might well be an efficient organizational choice when both the abilities and the behaviors of the managers are costly to observe, because it enables the chain's owner to save on the outlets' monitoring costs and to mitigate moral hazard.

We would like to underline that our explanation of dual distribution is developed for homogenous units, showing that heterogeneity of units is not a necessary condition for duality to emerge as an efficient device. Nevertheless, we by no means dismiss the importance of heterogeneity of units. Since previous empirical studies suggest that this factor is important, (Brickley and Dark 1987) care must be taken when testing the empirical validity of our predictions to incorporate the effect of such heterogeneity. We believe that, overall, our explanation of dual distribution is complementary to those based *ex ante* on the units' heterogeneity.

Finally, our core propositions on the efficiency of dual distribution can also be understood in a more general way: an organization facing the kind of contractual hazards we focused on is always more efficient by mixing incentive schemes provided to its members or agents; *i.e*. by combining fixed and revenue-based remuneration. In other terms, the generalization of variable remuneration inside an organization could be a sub-optimal policy since, ideally, it should not rely exclusively on fixed remuneration. This is another rationale for the prevalence of what is sometimes called "tapered integration", *i.e*. the simultaneous use of external input suppliers and in-house suppliers, in vertical relationships and supply chains.

REFERENCES

Azevedo P.F., and V.L. Dos Santos Silva., 2001, "Contractual Mix Analysis in The Brazilian Franchising", Unpublished Manuscript.

Bai, C.E, and Z. Tao, 2000, "Contract Mix in Franchising", *Journal of Economics and Management Strategy*, 9 (1), 85-113.

Baron, D.P., and D. Besanko, 1984, "Regulation, Asymmetric Information and Auditing", *RAND Journal of Economics*, 15, 447-470.

Bercovitz, J.E., 2001, "The Organizational Choice Decision in Business-Format Franchising: An Empirical Test", working paper, Fuqua School of Business, Duke University.

Besanko, D., and D. Spulber, 1989, "Antitrust Enforcement under Asymmetric Information", *Economic Journal*, 99, 408-425.

Bhattacharyya, S., and F. Lafontaine 1995, "Double-Sided Moral Hazard and the Nature of Share Contracts", *RAND Journal of Economics*, 26(4), 761-781.

Bradach, J.L., 1997, "Using the Plural Form in the Management of Restaurant Chains", *Administrative Science Quarterly*, 42, 276-303.

Bradach, J.L., and R.G. Eccles, 1989, "Price, Authority and Trust: From Ideal Form to Plural

Forms", *Annual Review of Sociology*, 15, 97-118.

Brickley, J., and F.H. Dark, 1987, "The Choice of Organizational Form: The Case of Franchising", *Journal of Financial Economics*, 18, 401-420.

Brickley, J.A., 1999, "Incentive Conflicts and Contractual Restraints: Evidence from Franchising", *Journal of Law & Economics*, XLII, 745-774.

Brickley, J.A., 2002, "Royalty Rates and Upfront Fees in Share Contracts: Evidence from Franchising", *Journal of Law, Economics & Organization*, 18(2), 511-535.

Bull, J., 2001, "Costly Evidence Production and the Limit of Verifiability", working paper.

Bull, J., and J. Watson, 2004, "Evidence Disclosure and Verifiability", *Journal of Economic Theory*, 118, 1-31.

Caves, R.E., and W.F. Murphy, 1976, "Franchising: Firms, Markets and Intangible Assets", *Southern Economic Journal*, 42, 572-586.

Gal-Or, E., 1995, "Maintaining Quality Standard in Franchised Chains", *Management Science*, 41(11), 1774-1792.

Gallini, N.T., and N. Lutz, 1992, "Dual Distribution and Royalty Fees in Franchising", *Journal of Law, Economics & Organization*, VIII, 471-501.

Holmström, B., and P. Milgrom, 1991, "Multitask Principal Agent Analysis: Incentive Contracts, Asset Ownership and Job Design", *Journal of Law, Economics & Organization*, 7, 24-52.

Khalil, F., 1997, "Auditing without Commitment"', *RAND Journal of Economics*, 28(4), 629-640.

Kaufman, P., F. Lafontaine, 1994, "Costs of Control: The Source of Economic Rents for McDonald's Franchisees", *Journal of Law & Economics*, 37(2), 417-453.

Klein, B., 1995, "The Economics of Franchise Contracts", *Journal of Corporate Finance*, 2, 9-37.

Krueger, A.B., 1991, "Ownership, Agency and Wages: An Examination of Franchising in the Fast Food Industry, *Quarterly Journal of Economics*, 106(1), 75-101.

Lafontaine, F., 1992, "Agency Theory and Franchising: Some Empirical Results", *RAND Journal of Economics*, 23, 263-283.

Lafontaine, F., 1993, "Contractual Arrangement as Signaling Devices: Evidence from Franchising", *Journal of Law, Economics & Organization*, 9, 256-289.

Lafontaine, F., K.L. Shaw, 1999, "The Dynamics of Franchise: Evidence from Panel Data", *Journal of Political Economy*, 107, 1041-1080.

Lafontaine, F., and K.L. Shaw, 2005, "Targeting Managerial Control: Evidence from Franchising", forthcoming, *RAND Journal of Economics*.

Lafontaine, F., and E. Raynaud, 2002, "Residual Claims and Ongoing Rent as Incentive Mechanisms in Franchise Contracts: Complements or Substitutes?", in E. Brousseau and J-M. Glachant, eds., *The Economics of Contracts: Theories and Applications*, Cambridge University Press, 315-336.

Lafontaine, F., and M. Slade, 2001, "Incentive Contracting and the Franchise Decision", in K. Chatterjee and W. Samuelson (eds*.), Advances in Business Application of Game Theory*, Kluwer Academic Press.

Lewin-Solomon, S.B., 1999, "Innovation and Authority in Franchise Systems: An Exploration of the Plural Form", working paper, Dept. of Applied Economics, Cambridge University.http://www.econ.cam.ac.uk/dae/people/lewin/empirical85.pdf

Mathewson, F., R. Winter, 1985, "The Economics of Franchise Contracts", *Journal of Law & Economics*, XXVIII, 503-526.

McAfee, R.P., and M. Schwartz (1994), "Opportunism in Multilateral Vertical Contracting: Nondiscrimination, Exclusivity and Conformity", *American Economic Review*, 84, 210-230.

Minkler, A., 1990, "An Empirical Analysis of a Firm Decision to Franchise", *Economic*

Letters, 34, 77-82.

Mookherjee, D., and I.P.L. Png, 1989, "Optimal Auditing, Insurance, and Redistribution", *Quarterly Journal of Economic*, 104, 339-415.

Norton, S., 1988, "An Empirical Look at Franchising as an Organizational Form", *Journal of Business*, 61, 197-217.

Oxenfeldt, A.R., and A.O. Kelly, 1969, "Will Successful Franchise Systems Ultimately Become Wholly-Owned Chains?" *Journal of Retailing*, 44, 69-83.

Pénard T., E. Raynaud and S. Saussier, 2003, "Dual Distribution and royalty rates in Franchised Chains: An Empirical Analysis Using French Data", *Journal of Marketing Channels*, 10, 5-32.

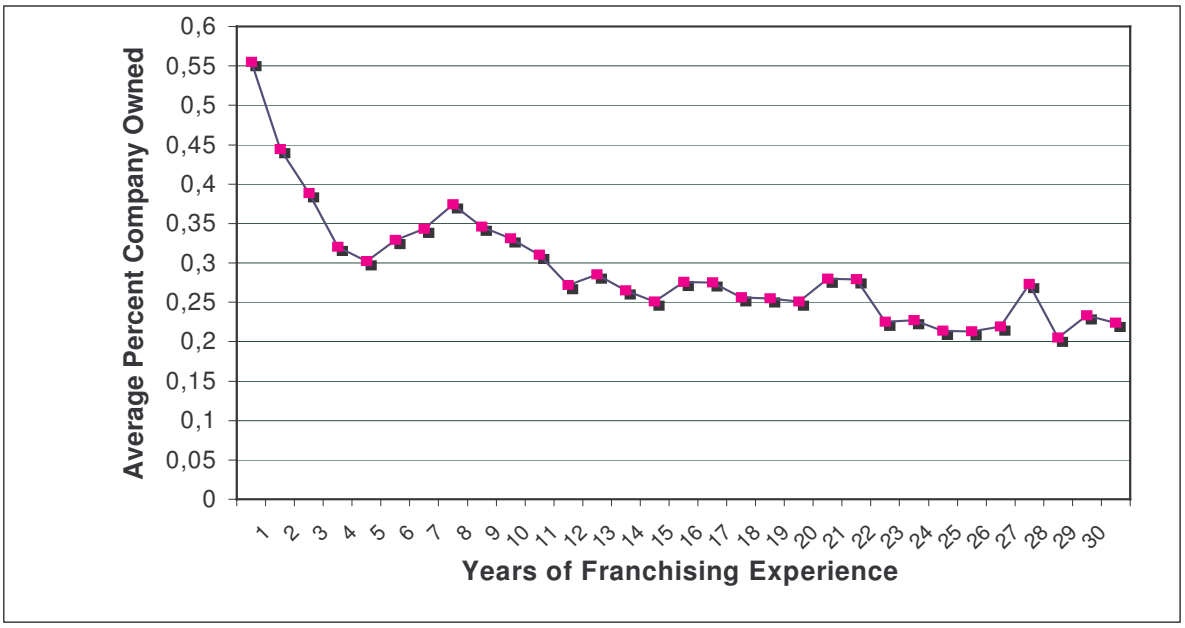
Rubin, P., 1978, "The Theory of the Firm and the Structure of Franchise Contract", *Journal of Law & Economics*, XXI, 223-233.

Scott, F.A., 1995, "Franchising vs. Company Ownership as a Decision Variable of the Firm", *Review of Industrial Organization*, 10, 69-81.

Shapiro, C., J.E. Stiglitz, 1984, "Equilibrium Unemployment as a Worker Discipline Device", *American Economic Review*, 74, 433-444.

Sorenson O., and J.B. Sorensen, 2001, "Finding the Right Mix: Organizational Learning, Plural Forms, and Franchise Performance", *Strategic Management Journal*, 22 (6-7), 713-724.

FIGURE 1. The proportion of Company Units as a Function of Franchising Experience



Source: Pénard-Raynaud-Saussier 2003