ELSEVIER

Journal of Corporate Finance 2 (1995) 39–74

Journal of
CORPORATE
FINANCE

The role of risk in franchising ☆

Francine Lafontaine *, Sugato Bhattacharyya

University of Michigan School of Business Administration, 701 Tappan Street, Ann Arbor, MI 48109-1234, USA

Abstract

The empirical literature on franchising suggests that the proportion of risk borne by franchisees increases as the amount of risk to be shared goes up. This has been interpreted by some as evidence that franchisors use franchising as a way to "shed" risk. This paper argues against this conclusion. First we show that the evidence is weak given the problems associated with measuring risk in franchising. Second, we show how a model emphasizing incentive issues and informational problems can give rise to the patterns found in the data. We conclude that risk shedding need not be invoked to explain franchising.

JEL classification: L22; L14; D20

Keywords: Franchising; Risk; Agency; Contracts

1. Introduction

The notion that risk plays an important role in franchising is quite pervasive in both the trade and academic literatures. First, the choice between becoming a franchisee and starting an independent business may very well hinge on considerations of risk [1]. Second, even in the context of

☆ This paper was prepared for the Conference on Franchise Contracting, Organization, and Regulation, held on May 27th, in Ann Arbor, MI. It has benefitted greatly from the comments of Dean Lueck, Scott Masten, Kathryn Shaw, Margaret Slade and Michael Weisbach. We also thank Robert Picard and Rajdeep Singh for their assistance. The usual caveat applies.

* Corresponding author

[1] See Williams (1994) and Forbes (1984) on some aspects of this issue.

SSDI 0929-1199(95)00004-6

franchising alone, risk averse franchisees need to be compensated more for bearing a greater amount of risk, especially given that their investments are typically undiversified [2]. From a modelling standpoint, too, the level of uncertainty associated with the input–outcome relationship critically determines the contract terms in the most prevalent models of franchising, namely models based on moral hazard considerations. On a somewhat related note, it has been suggested that franchise contracts may arise principally from the need to share risks. The idea is that franchising would allow for an efficient allocation of risk across the franchisee and the franchisor [3].

The main empirical implication of this last idea is that franchisors should bear more risk, either by operating more units directly or by increasing their royalty rate (and decreasing their upfront fixed fee), when there is more risk to share. This, of course, is derived under the assumption that the franchisee is the more risk averse party. This implication has been tested a number of times and, in general, the results do not establish that franchisors insure franchisees more when there is greater risk. In fact, the results support the notion of franchisors "shedding risk" in that the proportion of risk borne by the franchisee seems to go up with the amount of risk to be shared. This has been interpreted by some as evidence that the franchisor is more risk averse than the franchisee, an assumption that is contrary to what one would expect given the relative sizes of the two parties and their differing access to capital markets.

The present paper provides both a critical review of this empirical literature, and a reinterpretation of its results. In particular, we show that the conclusions one can draw from the existing literature are limited, given the problems that arise in measuring risk in franchising. In addition, we show that a model emphasizing informational problems can easily give rise to the patterns found in the data. On these grounds, we argue that the data as they now stand cannot be used to support the conclusion that franchisors use franchising to "shed" risk.

The paper is organized as follows. In the next section, we briefly review the existing empirical literature both in terms of the measures of risk used in the various studies, and their results. We then turn in Section 3 to a discussion of what one would want to be measuring to assess the role of risk in franchising, and the extent to which measures used in the literature actually achieve this goal. We argue that the conditions under which measures of the variance of sectoral sales (the level at which sales data are

---

[2] See Brickley and Dark (1987) for a more detailed argument.

[3] See Cheung (1969) for an early statement of this argument as an explanation for the existence of another form of share contract, namely sharecropping contracts, and Martin (1988) for a similar argument about franchising contracts.

typically available in franchising) would be meaningful as measures of riskiness are unlikely to hold. As a result, we suggest that discontinuation rates might better capture the degree of risk in franchising. In Section 4, we show some of the patterns in the data that have been interpreted as an indication that franchisors use franchising to reduce the amount of risk they face. We then present an alternative explanation for these data patterns in Section 5. More specifically, we propose a model that reinterprets the empirical results by emphasizing incentive and information issues over risk. Section 6 concludes.

2. An overview of the existing evidence

Several authors, notably Martin (1988), Norton (1988), Banerji and Simon (1992), and Lafontaine (1992) and Lafontaine (1993) have used different measures of risk to assess whether franchisors rely more or less on franchising (relative to company ownership) with increases in risk. In addition, Lafontaine (1992) and Sen (1993) have examined the effect of risk on franchisors' choices of royalty rates [4]. Table 1 summarizes both the way in which risk has been measured in these studies, and the results they obtained.

As can be seen from this table, the most frequently used measure of risk has been some notion of variance of sales over time, where the sales data have been obtained either on a franchising sector basis from the U.S. Dept. of Commerce's *Franchising in the Economy* (see e.g., Martin, 1988, Lafontaine, 1992), or on a SIC basis from census data (Norton, 1988). In all cases, the sales data have been aggregate: authors have used time-series data on average sales per outlet in franchising sectors defined by the Department of Commerce to generate their risk measures. Lafontaine (1992 and 1993) and Sen (1993) relied instead on some notion of failure rates of outlets as a way to capture the degree of riskiness in these same franchising sectors [5].

Table 1 also shows that in most cases where estimates are statistically significant, the effect of risk on franchisors' use of franchising, that is, on their proportion of franchised units (%F), has been positive. The evidence on

[4] Brickley and Dark (1987) rely on a measure of total investment to assess the cost of "inefficient risk-bearing". They do not make any attempt to measure the amount of risk faced by franchisees directly or assess its effect on this cost.

[5] The sectors defined by the U.S. Department of Commerce are clearly not as homogeneous as one would like for these purposes. The worse cases are probably the Construction, Home Improvement, Maintenance and Cleaning sector, which includes firms involved in remodeling kitchens and building houses all the way to cleaning services, and the Non-Food Retailing sector, which includes furniture stores, drug stores, hardware stores and vending among other things. Because of this, all these measures are necessarily very imperfect.

Table 1
A summary of the empirical literature on the effect of risk on franchising

Authors	Measures	Type of Data	Results For % F	For R
Martin (1988)	Variance of Residuals from a Regression of Average Sales per Outlet on Time 1. For Company-Owned Outlets 2. For Franchised Outlets	Sectoral Data from D.O.C.	+	not tested
Norton (1988)	Root Mean Square Error of Detrended Percentage Change in Sales	State Level Data for 3 Industries	+ (sign. in one of three industries)	not tested
Lafontaine (1992)	1. Discontinuation Rates 2. Normalized Variance of Detrended Average Sales 3. (Normalized Variance of Detrended Average Sales)· Number of Outlets	Sectorial Data from D.O.C.	+	– (not sign.)
Banerji and Simon (1992)	Standard Deviation of Monthly Stock Returns	Franchisor Level, from CRSP and Compustat	–	not tested
Lafontaine (1993) [a]	Discontinuation Rates	Sectoral Data from D.O.C.	– (not sign.)	+
Sen (1993)	Termination Rate of Franchises [b]	Sectoral Data from D.O.C.	not tested	–

Not sign. means not significant at $\alpha = 0.05$.
[a] Lafontaine (1993) focused on only new franchisors in order to assess the implications of the signaling model. The special sample that resulted might explain the different results obtained in that study.
[b] This measure is different from the discontinuation rate measure used in Lafontaine (1992) and (1993). Discontinuation refers to the outlet itself being closed down. The data on this for 1986 is given in Dept. of Commerce (1988), p. 47. Termination refers to the franchise relationship being terminated. It is unclear from reading Sen's paper whether the measure he uses includes only terminations by franchisors, or also terminations by franchisees and by mutual agreement. See Dept. of Commerce (1988), p. 53 for the 1986 data.

the effect of risk on royalty rates (R) is more sparse, but it suggests that more risk does not generally imply an increase in royalty rates as would be necessary to provide more insurance to the franchisees [6]. Hence, these results are not supportive of the risk sharing or the one-sided moral hazard

[6] These conclusions about the empirical evidence are strengthened if one excludes the one study whose sample was chosen purposefully rather than randomly, namely Lafontaine (1993). This study focused on assessing implications from a signaling model. This required analyzing the contract terms and contract mix of new franchisors only, and the results may not be general.

model. They suggest that franchisors "shed" risk instead of insuring franchisees. This would make sense, under risk sharing, only if franchisees were less risk averse than franchisors, an assumption that goes counter to the usual arguments about relative sizes and access to capital markets. Interestingly, under one-sided moral hazard, where the agent is the franchisee, the extreme assumption that the franchisor is risk averse and the franchisee risk neutral would lead the franchisor to always sell the unit to the franchisee for a fixed fee. This is because there would be no need to trade off the provision of insurance to the franchisee with the need to give him incentives if in fact the franchisee was best suited to bear risk. Hence in this case we would not have the type of franchise contracts that we actually observe or the mix of company-owned and franchised contracts that these empirical studies have focused on [7].

3. "Riskiness" in franchising: measurement issues

The idea that the efficient allocation of risk among individuals with different preferences plays an important role in firms' decisions about franchising arises from both models of pure risk-sharing and of one-sided moral hazard [8]. In the latter type of model, the franchisor is usually assumed to be risk neutral, or at least less risk averse than the franchisee. This gives rise to the need for her to insure the franchisee. Hence the risk one should be trying to measure to test the risk implication of this type of model is that faced at the level of the franchised unit, or the outlet-level risk. In the pure risk-sharing model, the idea is that franchising means that the franchisor and the franchisee each bear only a portion of total risk. In this case, one would want to assess the riskiness of the unit to the franchisor as well as the franchisee. If the franchisor is not diversified, then the risk of a unit to both parties will be the same, and we can again focus only on the outlet-level risk. If the franchisor is diversified (e.g. geographically), then the incremental amount of risk she faces if she operates an additional unit will be smaller

[7] Williams (1994) takes a different stab at the issue. He uses the Characteristics of Business Owners (CBO) database to assess what makes individuals choose to franchise rather than start their own independent business. He finds that more risk measured by the standard deviation of sales across franchised businesses in an industry leads more individuals to join a franchise. He concludes that individuals with "a greater than average aversion to risk are more likely to choose franchising over independent business ownership." As franchisors are usually firms started by individuals who choose the independent business route, Williams' conclusion is consistent with the notion that franchisees, not franchisors, are the more risk averse parties in this relationship.

[8] To date, most other models have assumed risk neutrality of both the franchisee and the franchisor.

than the incremental amount of risk faced by the franchisee. But the reduction in incremental risk faced by the franchisor in this case can also be modeled, under certain circumstances, via a reduction in her risk aversion parameter. The decision as to how to allocate the risk will then again be based on the outlet-level risk, making it the relevant concept for us to measure.

Under both pure risk sharing and one-sided moral hazard models, what is meant by riskiness is the degree to which some exogenous factor or shock affects outcomes. Measures of dispersion, such as the variance, provide an intuitively appealing way to capture this effect [9]. Assume, for example, that sales at the outlet level can be written as:

$$S_{it} = f(e_{it}, r_t, \theta_{it}), \tag{1}$$

that is, sales of outlet i at time t are a function of the amount of effort put in by the outlet manager or franchisee at time t, e_{it}, the effort of the franchisor at time t, r_t, and some random shock θ_{it}, with mean 0 and variance σ_θ^2. If the two inputs are fixed, then sales variability over time and across outlets arises strictly as a result of the variability in the θ_{it}. In that case, a measure of the variance of sales over time at the outlet level, or across outlets at time t, would correctly capture the degree of "riskiness" faced by an outlet.

We examine more specifically below how and when the available data on franchising can yield measures of variance that correctly identify the variance of θ_{it} or, alternatively, measures that correctly rank firms or sectors according to their degree of riskiness. First, however, we note that the variance of profits (or returns) at the franchisor or the franchisee level would not give a correct measure of risk here. This is because the division of profits between the franchisor and the franchisee is determined by the franchise contract. Suppose, for example, that the franchise contract stipulates that the franchisee will pay a fixed dollar amount of royalties per period to the franchisor [10]. Then the revenues of the franchisor would be fixed, and her profits very stable, whatever the degree of sales riskiness at the outlet level might be. More generally, if the franchise contract involves an upfront fee and royalties based on sales, as they usually do, and ignoring the time t subscripts for simplicity, we have

$$\pi_i = (1-\beta)S_i - V(e_i) \tag{2}$$

and

$$\Pi_i = \beta S_i - U(r) \tag{3}$$

[9] According to Lepper (1967), the use of variance as a measure of risk dates back at least to Makover and Marschak (1938).

[10] In some franchised chains, royalties are fixed in this way. See Lafontaine (1992).

where π_i is the per period profit remaining with the franchisee after payments of royalties to the franchisor, $V(e_i)$ is the cost of the franchisee's local effort, $U(r)$ is the cost of the franchisor's effort, and Π_i are the franchisor's profits from this unit [11]. Assuming that the effort levels e_i and r are fixed, and V is the same for all franchisees in the chain, the variance of these profits will be

$$Var(\pi_i) = (1-\beta)^2 Var(S_i) \tag{4}$$

and

$$Var(\Pi_i) = \beta^2 Var(S_i). \tag{5}$$

Unless we can correct for the sharing rule (namely β), measures of the variance of franchisor or franchisee profits (or returns as in Banerji and Simon, 1992) will not give a correct assessment of the degree of risk faced by different firms. Furthermore, because the β's vary across firms and on average across sectors, the ranking of sectors according to their "riskiness" based on such measures will also be incorrect. Fortunately, measuring the underlying riskiness of the business using sales –as opposed to profit – data does not suffer from this problem. For that reason, and because measures based on the variance of sales have been most frequently used in this literature, we focus on sales-based measures in what follows.

3.1. Measures of risk based on the variance of sales

There are two major issues that arise when one tries to measure risk or uncertainty in franchising using measures of the variance of sectoral per outlet sales. The first comes from the variance in factor input decisions, which relates to the fact that inputs are not fixed or the same for all outlets. The second is an aggregation problem that arises because of the nature of the available data. We explore below the implications of each of these in some detail. Before doing so, however, we raise two other points that have a significant impact on the results one obtains in measuring risk, or in ranking firms or franchising sectors according to their level of risk.

First, because the available data on sales are of a time series nature, one should detrend them, if they display a trend, before calculating variances. Otherwise, the observed variance in sales may include a trend component,

[11] We do not consider the effect of the upfront fixed fee in these profit equations because 1) it is paid only once at the beginning of the contract, not in each period over which the profits are observed; 2) these fixed fees are very small relative to royalty payments; and 3) as sunk costs, they do not enter the variance of profit equations. We also, for simplicity, do not consider any potential input costs beside the cost of the two types of effort.

thereby overstating the actual riskiness. For example, suppose that sales levels in outlet i are affected by time, so that

$$S_{it} = \alpha_i t + \theta_{it}, \tag{6}$$

where t is a time index and α_i is a fixed parameter. Then, assuming that there is no correlation between $\alpha_i t$ and θ_{it}, the variance of sales for outlet i will be given by

$$Var(S_{it}) = Var(\alpha_i t) + Var(\theta_{it}) \tag{7}$$

where

$$Var(\alpha_i t) = \sum_{t=1}^{T} \left(\alpha_i(t - \bar{t})\right)^2 / (T-1) > 0. \tag{8}$$

In other words, the variance of sales would overestimate the variance of θ_{it}. Of course, if all outlets in all sectors grew at the same rate, this problem would only affect the level of the measures and not the ranking of firms or sectors according to their riskiness. But this is rather unlikely. If, instead, outlets in some sectors grow faster than outlets in others, rankings will also be affected. This problem has been noticed in the empirical literature: all of the above studies used detrended data [12]. In the remainder of this paper, any mention of time-series data on sales should be interpreted to mean their detrended version [13].

Second, it is important to recognize that there are important differences in the levels of the variables whose variances are being calculated. For example, sales in corporate outlets are typically much greater than sales in franchised outlets (see Table 4 below). Similarly, average sales in the hotel industry are much larger than average sales in the dry cleaning industry. Given that variances are proportional to the square of the level of a variable, the variance of sales will usually be much larger in those sectors or in the type of outlet (company owned or franchised) that displays higher average sales [14]. Yet these differences in the standard errors do not reflect relevant differences in risk. In particular, if the franchisor and franchisee preferences can be described by mean-variance utility functions, the coefficient of variation

[12] Banerji and Simon (1992) are an exception, but this is because they measure the variance of returns rather than sales.

[13] This correction is only really valid if the sales levels of all outlets in a sector are growing at the same rate, an assumption that is unlikely to hold in reality given the degree of heterogeneity of the franchising sectors and the idiosyncratic characteristics of individual chains. However, given the nature of the data, no better alternative is available.

[14] The idea is simply that if average sales in a sector, Z, can be written as $Z = c \cdot Y$, where Y represents average sales in another franchising sector, and c is a constant of proportionality, the variance of Z, $Var(Z)$, will be given by $c^2 \cdot Var(Y)$.

will correctly assess the riskiness that these people care about, while the variance will not [15]. In general, empirically, some form of normalization will be necessary [16].

Having made these points, the main issue we are concerned about now becomes: When will the (normalized) variance of (detrended) average sales data over time provide a) an accurate assessment of outlet level risk, or b) an accurate ranking of firms, or franchising sectors, in terms of their degree of riskiness. In the remainder of this section, we note that the conditions under which the variance of average sales per outlet in a franchising sector correctly assesses the riskiness faced by franchisees are rather stringent. Given this, we suggest that discontinuation rates might better capture the degree of risk at the unit level.

The effect of factor input variability

As noted above, input levels underlying the sales per outlet data are not fixed across units or over time. Given our sales equation

$$S_{it} = f(e_{it}, r_t, \theta_{it}), \tag{9}$$

the estimated variance of per outlet sales in a franchising sector will include not only the variance of θ_{it}, but also the variance in e_{it} and in r_t (and any other input for that matter) across outlets and over time, as well as their interactions. In other words, the variance of sales will overestimate the variance of θ_{it}, σ_θ^2 [17].

There are circumstances under which some measure based on the variance of aggregate sales will still correctly capture σ_θ^2 even if input levels vary [18]. Specifically, assume that the franchisor's effort is unimportant, and that the same realization of the random shock affects all outlets of a given franchisor in a given year, that is $\theta_{it} = \theta_t$ for all $i = 1 \ldots n$, where n is the number of units

[15] A mean-variance utility function is one where $EU(\tilde{X}) = \bar{X} - \rho\sigma_{\tilde{X}}^2$, where ρ is a constant absolute risk aversion parameter, and risk neutrality can be represented by $\rho = 0$.

[16] The empirical importance of normalization can be illustrated by looking at results from Martin (1988). In his Table 1, he shows ratios of "riskiness" for company-owned versus franchised units. He also shows that sales in company-owned units are on average about 3 times higher than sales in franchised units. Had he normalized his measures of risk by dividing by the average sales of each type of outlet, all but two of his ratios of "riskiness" would have been less than one, supporting much better his argument that company-owned units are less "risky" than franchised units.

[17] This issue will be especially problematic when observed variances in sales levels achieved by franchised and by company-owned units are used to compare the degree of risk faced by each type of unit. The effort levels of both the franchisor and the franchisee, and their variance, will depend on the choice of organizational form and on the terms of the franchise contract, making them endogenous to the decisions about which outlets are franchised and which are company-owned.

[18] We thank Dean Lueck for his help on this point.

in a chain. Assume also that the sales equation is additively separable. Then we can rewrite (9) as

$$S_{it} = f(e_{it}) + \theta_t. \tag{10}$$

Average sales across all outlets in the chain in a given year, $\bar{S}_t$, can be written as

$$\bar{S}_t = \left(\sum_{i=1}^{n} \frac{f(e_{it})}{n} \right) + \theta_t. \tag{11}$$

Assuming that the e_{it}'s are bounded, by the law of large numbers, as n goes to infinity, we will have

$$\bar{S}_t = K_t + \theta_t, \tag{12}$$

where K_t is the limit of $\Sigma_{i=1}^{n}(f(e_{it})/n)$ as $n \to \infty$ [19]. In addition, assuming that the distribution from which the e_{it}'s are drawn is stationary, and that the f function remains the same each period as well, then $K_t = K$ is independent of time, and we have

$$Var(\bar{S}_t) = Var(\theta_t) = \sigma_\theta^2. \tag{13}$$

In other words, by aggregating and calculating the variance of average sales over time for all outlets in a given chain, rather than the variance across individual outlets's sales data, we can eliminate some of the bias introduced by the variability of e_{it} in the data as long *as all these outlets face exactly the same realization of the random shock at time t*. In that sense, aggregation can be used to get around the factor variability problem.

Unfortunately, while the assumptions underlying this procedure can be very reasonable in some cases, they are not really appealing here [20]. First, the sales function is not likely to be additively separable: one could reasonably expect a franchisee's effort to be affected by the realization of the random shock.

Second, the assumption that $\theta_{it} = \theta_t$ for all $i = 1...n$ is unlikely to be valid. In particular, differences in tastes and in economic activity across regions should lead to differences in the realization of θ_{it} of different units within a chain. Moreover, the data that are available to estimate the variance are sectoral average sales per outlet, not simply firm averages. The assumption that would be necessary to make this level of aggregation appropriate would

[19] This procedure also assumes that the e_{it}'s are independent for $i \neq j$, that is that the choice of effort of one franchisee is not influenced by the choice of effort of other franchisees in the same chain.

[20] See Allen and Lueck (1994) for an example, involving sharecropping, where these assumptions appear reasonable.

be that all outlets in a particular industry in the United States face the same realization of θ_{it}.

Third, we have assumed that the franchisee/manager effort enters the sale equation in a particular way, namely as $f(e_{it})$. The assumption that this can correctly represent the technology or sales function of all outlets in a chain seems reasonable. But once one considers using sectoral data to measure the variance, it has to be the case that the same f function describes the sales function in all chains in the sector, a more heroic assumption.

Finally, we have assumed away the issue of franchisor input by eliminating it from the sales equation. If the franchisor's input matters, and maintaining the additive separability assumption, we have

$$\bar{S}_t = K_t + g(r_t) + \theta_t. \tag{14}$$

To eliminate the effect of variation in this variable on the variance of average sales over time within a chain, one would have to assume that r_t is constant over time, or growing in a constant way such that detrending the sales data would correct for this effect. When calculating the variance of sectoral average sales per outlet, the assumptions that would have to hold for this variance to represent an estimate of σ_θ^2 are 1) that the g function is not chain specific; 2) that r_t is the same for all outlets in the various chains included in a sector; and 3) that r_t does not change over time, or that it follows a simple time trend.

In general, we don't expect this rather long list of assumptions to hold in franchising, especially given the rather heterogeneous definition of some of the franchising sectors. As a result, we conclude that the effect of factor input variability is not likely to be eliminated here by a switch from outlet level to aggregate data.

An alternative approach would be to recognize the existence of the upward bias of the variance measures induced by variation in input levels, and assume that the bias is of the same magnitude in all franchised chains. This requires that the extent of variation in the e_{it} and in r_t, and their effects on sales levels, be similar in all chains. Under those circumstances, assuming again that the random term enters the sales equation in an additive way, the variance of sales at the outlet level would at least rank sectors correctly in terms of their degree of riskiness. It may even provide a correct assessment of the risk differentials between sectors. While such assumptions are again rather heroic, for some empirical purposes, this might be a reasonable way to proxy the degree of risk. Alternatively, it might be possible to correct the measures ex post for such variation within each sector. In either case, this would allow us to abstract from the problem of factor input variability and focus on measuring the variance of outlets' sales, taking for granted that we can infer from it some notion of the underlying risk.

In the next section, we show that once one abstracts from the problem of factor input variability, given the sectoral nature of the available data, the main issue becomes the extent to which outlets in a sector do or do not face the same realization of θ_{it}, or put differently, the extent to which measuring the variance of average sales per outlet in a sector gives a correct estimate of the variance of individual outlets' sales within a chain or sector.

Aggregation problems

Abstracting from the issue of factor input variability, we now show how measuring risk by calculating the variance of average sales per outlet in various franchising sectors raises its own set of issues. Fundamentally, the idea is that the correspondence between the variance of sales at the outlet level and the measure obtained from aggregate data will depend on the correlations among the outlets' sales levels and, as a corollary, may depend *on the number of outlets in each sector*. Similarly, the ranking of sectors in terms of their riskiness will depend on these correlations and potentially on the number of outlets per sector.

To abstract from the factor input variability problem, assume now that outlet sales are affected only by a random term, or in its simplest form, that

$$S_{it} = \theta_{it} \tag{15}$$

where θ_{it}, and thus S_{it}, follow a distribution with mean 0 and variance $\sigma_\theta^2 = \sigma_S^2$ [21]. Under these circumstances, we want to know when the variance of average sales per outlet in a sector will give a good measure of the variance of sales across outlets within a chain or sector. From basic statistics, the variance of a mean can be written as

$$Var(\bar{S}_t) = Var\left(\frac{\Sigma_1^{N_t} S_{it}}{N_t}\right) = \frac{1}{N_t^2}\left[\sum_{i=1}^{N_t} Var\, S_{it} + 2\sum_{i=1}^{N_t-1}\sum_{j=i+1}^{N_t} Cov(S_{it}, S_{jt})\right] \tag{16}$$

where $\bar{S}_t$ represents average sales per outlet in the sector, and N_t is the number of outlets, company-owned and franchised, in the sector, both at time t [22]. Assume that the S_{it}'s are all drawn from the same population, with variance σ_S^2. Assume also, as was done in the section above, that they are

[21] Note that when using sectoral data below, this assumption says that the distribution of the random term or sales levels is the same for all chains in the sector, and that it is the same in all time periods.

[22] Under the assumption that all outlets' sales follow the same time trend, the error term in a regression of average sales on time will represent the mean of individual outlets' error terms: If $\hat{y}_{it} = \hat{\bar{y}}_t$, for $i = 1,...N$, then $\Sigma_{i=1}^N (y_{it} - \hat{y}_{it})/N = \Sigma_{i=1}^N y_{it}/N - \hat{\bar{y}}_t = \bar{y}_t - \hat{\bar{y}}_t$, i.e. the mean of the error terms equals the error term of the mean. Similarly for the error term normalized by $\bar{Y}$. Hence the formula for the variance of a mean is appropriate.

perfectly correlated across all outlets in the sector, or, in other words, that all outlets in the sector face the same realization of θ_{it} at time t. Then from (16),

$$Var(\bar{S}_t) = \frac{1}{N_t^2}\left[N_t \cdot \sigma_S^2 + (N_t - 1)N_t \cdot \sigma_S^2\right] = \sigma_S^2, \tag{17}$$

which implies that the variance of average sales will be the same as the variance of sales in the individual outlet. In this case, an estimate of Var($\bar{S}$), $\hat{V}_{\bar{S}}$, obtained by calculating the variance of the available aggregate time series data is also an estimate of σ_S^2. This, of course, is consistent with our earlier discussion.

Now assume, in contrast, that the S_{it}'s are mutually independent. Then (16) reduces to [23]

$$Var(\bar{S}) = \sigma_S^2/N. \tag{18}$$

Given an estimate of Var($\bar{S}$), $\hat{V}_{\bar{S}}$, one can get an estimate $\hat{V}_S$ for the outlet-level variance, σ_S^2, by multiplying $\hat{V}_{\bar{S}}$ by N [24].

Finally, if the sum of the covariance terms is positive, we have

$$Var(\bar{S}) < \sigma_S^2 < Var(\bar{S}) \cdot N \tag{19}$$

so that $\hat{V}_{\bar{S}}$ is a lower bound, and $\hat{V}_{\bar{S}} \cdot N$, an upper bound, estimate for σ_S^2 [25].

If the number of outlets, N, was similar across sectors, and the correlation patterns were similar as well, the difference between $\hat{V}_{\bar{S}} \cdot N$ and $\hat{V}_{\bar{S}}$ would be irrelevant from a measurement perspective. Using either one of these as a measure of risk would lead to different estimated levels of risk, but to the same ordering of sectors' risk. In fact, even relative magnitudes would be preserved. Hence either measure would lead to a correct assessment of risk. However, given that N varies across sectors, if outlets' sales levels are independent from each other, then $\hat{V}_{\bar{S}}$ will underestimate the "riskiness" of sectors with a large number of firms relative to other sectors. If, on the other hand, the correlations are positive and large, $\hat{V}_{\bar{S}} \cdot N$ will overestimate the riskiness of sectors with a large number of firms relative to the other sectors.

---

[23] Basically, if all S_{it}'s are mutually independent and drawn from the same distribution, one can view the average sales figure each year as a draw from a population of mean $\bar{S}$ and variance σ_S^2/N. The variance of these average sales figures over time, Var($\bar{S}$), then represents an estimate of σ_S^2/N.

[24] Given that N varies over time, the correction would involve multiplying the variance of the aggregate time-series data by the mean of N over time, $\bar{N}$. This is because Var($\bar{S}$) is a measure of each of the Var($\bar{S}_t$) we would estimate if we had data on sales of individual outlets each year. As a result, $\text{Var}(\bar{S}) * N_1 + \text{Var}(\bar{S}) * N_2 + \ldots \text{Var}(\bar{S}) * N_T = T * \sigma_S^2$, or $\text{Var}(\bar{S}) * (N_1 + N_2 + \ldots + N_T)/T = \sigma_S^2$.

[25] However, if the sum of the covariance terms is negative, $\hat{V}_{\bar{S}} \cdot \bar{N}$ may still underestimate σ_S^2.

This discussion begs the following question: under what circumstances should we expect to find positive, negative, or no correlation among the sales levels of different outlets in a franchising sector? Two main forces are at play here. The first is the level of economic activity or the business cycle that could affect all firms within an industry in a region in a similar way and imply positively correlated sales. In terms of the sales equation above, business cycles imply positively correlated random terms across many outlets at a given time. On the other hand, units located close to each other and involved in the same type of business compete with each other. In that case, higher sales in one outlet might occur at the expense of lower sales in another outlet.

A summary

The discussion above suggests that the (normalized) variance of (detrended) average sales per outlet in various franchising sectors will provide an accurate measure of σ_θ^2 only if:

1. The sales function is additively separable in all three of its arguments;
2. The f function is the same in all chains in a sector, and it remains constant over time;
3. The e_{it}'s for all outlets in a sector in all periods are drawn from the same distribution;
4. The same realization of the random shock affects the sales levels of all units in the sector;
5. The franchisor's input is unimportant or the same for all franchisors in a sector (and it enters the sale function in the same way in all chains as well) [26].

Under these conditions, the sales levels of different outlets will be perfectly correlated, and the variance of average sales will be exactly the same as the variance of sales at the outlet level. Furthermore, the problem of input level variability can be resolved via aggregation so that the estimated variance of average sales will in fact capture the variance of θ.

If, however, the realization of the sales levels are completely independent across outlets, then the variance of average sales *multiplied by the number of units in the sector* will correctly assess the variance of sales at the outlet level, or σ_S^2. But in this case, the realization of the random shock is different for different units, so that aggregation does not resolve the problem of factor input variability. As a result, all measures based on the variance of sales will overestimate the variance of θ. One will have to argue that the extent to

[26] One other assumption that is embedded in our formulation of the problem is that the distribution of the random shock is the same for all chains in a sector, and it does not change over time.

which input levels, e_{it} and r_t, and their effects on sales, the f and g functions, vary within sectors is similar in all sectors so that the extent to which the measures are biased would be similar for all. In that case, the variance of average sales per outlet in a sector –multiplied by the number of outlets in the sector – could be used to rank sectors according to their degree of riskiness. Alternatively, a version of this measure that controls for input level variations and their effects on sales could be used to approximate the underlying riskiness of the sales equation.

Finally, if the above set of conditions does not hold, then none of the variance-based measures will come close to capturing the extent of exogenous risk.

3.2. Failure rates as a measure of risk

The conditions above are unlikely to be really satisfied in the context of franchising arrangements. In that sense, measures of risk based on the variance of sales are generally unsatisfactory, suggesting that one look for alternative ways to capture risk in franchising. As noted in Table 1, two different notions of failure rates have been used as measures of risk in the empirical literature on franchising. Lafontaine (1992) and Lafontaine (1993) has used discontinuation rates of units within franchising sectors, as reported by the Dept. of Commerce in *Franchising in the Economy*, as her measure of risk [27]. Sen (1993) used termination rates, from the same source [28].

In franchising, discontinuation typically refers to a case where an outlet is closed down. Whether the franchisee remains affiliated with the chain in this case will depend on whether he owns additional units, or whether the franchisor decides to award him a different unit, and whether he wants to remain associated with the franchisor. Termination, on the other hand, implies that the franchise relationship is discontinued; the unit might or might not remain open, depending on the termination conditions included in the franchise contract, and on who owns the land and building at that location. Terminations by franchisors can be due to a number of factors unrelated to "exogenous" risk, and terminations by franchisees or by mutual consent are especially likely to reflect a host of other factors beside risk [29]. Most importantly, company-owned units are never terminated: if they are

[27] See Dept. of Commerce (1988) p. 47 for the 1986 data.

[28] See Dept. of Commerce (1988), p. 53 for the 1986 data. Note that non-renewals are in effect "terminations" that occur at the end of the contract. In that sense, one might want to include those in a measure of terminations.

[29] It is unclear from reading his paper whether Sen (1993) used only terminations by franchisors or also included terminations by franchisees and by mutual agreement in his estimation of the proportion of franchises terminated.

closed down, they are discontinued. If the manager is replaced by another, no "termination" has occurred either. In other words, it is not possible to use termination as a measure of risk for company-owned units. For these reasons, we focus on discontinuation rates in what follows.

One major difference between discontinuation rates as a measure of risk compared to measures based on the variance of sales is that the former do not really capture the variation of the random term in a sales equation. Instead, discontinuation rates measure "downside" risk, or the probability that a bad outcome will occur [31]. This, interestingly, may be a more appropriate way to describe the type of risk that franchisees care about [32].

A potential disadvantage of discontinuation rates is that they may underestimate actual "failure" rates. For example, company-operated outlets may not be discontinued even if they are "unprofitable" if they serve other purposes beside the generation of outlet profits, such as being used as training grounds or information gathering systems, or to increase the chain's visibility in certain critical locations [33]. This would make the proportion of company-owned outlets, and hence the proportion of all outlets, that are discontinued a biased measure of the risk faced by franchisees. However, there is a limit on the extent to which franchisors should be expected to use non-profitable units for these activities when profitable ones are at hand. In that sense, the bias from this effect should not be too large.

Advantages of discontinuation rates as a measure of risk in franchising include the fact that they are not as vulnerable to the effects of growth in the sales data or differences in levels of sales as variance-based measures are. For example, growth in sales levels, if accompanied by an equivalent growth in costs, will likely leave unchanged the proportion of discontinued outlets. This is as it should be for a good measure of risk [34]. Similarly, differences in

[31] Measures of risk based on the probability that an outcome reaches a "disaster" level have been used frequently in labor economics, for example. There, the notion of risk is often considered to be less related to income variability and more to the probability of unemployment or the probability of accidents. Note that under certain conditions, namely that discontinuation occurs whenever sales reach a given low level, and that the distribution of possible outlet sales is symmetrical, there will be a one to one correspondence between the variance of sales at the outlet level, that is the variance one would want to measure, and the probability of discontinuation.

[32] One of the major selling points of franchising to franchisees over the years has been the statistics found in the trade press on the very low failure rates of franchised businesses compared to independent operations. These statistics never had real scientific bases and recent work by Bates (1995) and Williams (1995) casts serious doubts on them. However, the extent to which they have been used by franchisors suggests that the risk of failure is indeed a major factor in individual decisions about franchising versus opening an independent business.

[33] See Smith II (1982), footnote 4, for an example of the latter in automobile retailing.

[34] If sales grow faster (slower) than costs, the proportion of discontinued units would likely decrease (increase). But this change would in fact reflect an actual reduction (increase) in the risk faced by the franchisee so that the measure would still be appropriate.

levels of sales (hotels versus dry-cleaners for example) will be accompanied by differences in the levels of costs such that the proportion of discontinued outlets should not be systematically related to the sales levels [35]. Though variance-based measures can be corrected to deal with these issues, as noted earlier, these corrections rely on a number of assumptions that may be unappealing.

Most importantly, the discontinuation rate calculated from sectoral data is just an average of the discontinuation rates at the firm level in that sector. As a result, if the discontinuation rate is assumed to be a good measure of riskiness at the firm level, it remains an equally good measure at the sectoral level, irrespective of the correlation structure of store-level sales. This is in sharp contrast to variance-based measures where the correspondence between sectoral level and outlet-level variances is vitiated by the correlation structure of outlet-level sales.

Finally, discontinuation rates are not as susceptible to factor input variability as are variance-based measures. For one thing, failure rates will not be systematically sensitive to variations in the level of franchisee or franchisor input that arise from normal optimizing behavior. Furthermore, they will not be affected by franchisee shirking to the same extent that variance-based measures are. The reason for this is that if a franchised unit is not doing well, and the franchisor feels it has potential, she will generally take it over or transfer it to another franchisee rather than discontinue it [36]. Such buybacks are especially likely to occur if the poor performance can be attributed to bad management or shirking on the part of the franchisee. In other words, when bad outcomes occur because of franchisee shirking, they do not show up as discontinuations but rather as terminations. Hence discontinuation rates will measure exogenous rather than franchisee-induced risk. However, failure rates will still be sensitive to franchisor moral hazard: if a franchisor shirks, or behaves opportunistically, this will increase the probability of failure for all units in the chain, and is likely to show up in the discontinuation statistics [37]. Assuming that most franchisors do not shirk most of the time however, measuring the discontinuation rate on a sectoral basis, and over a number of periods, should smooth out some of these effects [38].

[35] See Table 2 below for an illustration of the degree to which unnormalized measures of the variance of sales correlate with sales levels.

[36] See Thompson (1971), p.34, on this issue.

[37] The rate of franchisor failures might provide a better measure of the underlying riskiness of a product. However, this information is not publicly available on a sectoral basis. One of the authors has begun the process of gathering this type of information.

[38] For several purposes, a measure of risk that includes the effect of franchisor moral hazard might actually be appropriate. For example, the amount of risk premium requested by a franchisee should be a function not only of exogenous risk, but also of the risk of franchisor opportunism. See McAfee and Schwartz (1994) for more on the effect of franchisor opportunism on contract terms.

Given the rather important advantages of discontinuation rates over variance-based measures, we conclude that these provide a better way to capture risk in franchising.

4. Risk measures: an empirical comparison

4.1. The data

This section shows what numerical values one obtains using the different measures of risk discussed above, and how they compare to each other. It also shows some of the data patterns that have led to the conclusion that franchisors do not insure franchisees more when there is more risk, but instead use franchising to "shed" this risk.

The principal source of aggregate data on franchising is *Franchising in the Economy*, a yearly publication of the U.S. Department of Commerce. This source provides, among other things, sectoral data on total sales and total number of outlets, and data on discontinued outlets. The sectors defined in this publication include 14 Business Format franchising sectors, and a number of subsectors as well. In what follows, we focus strictly on the 14 sectors because data on discontinuation rates are not available for the subsectors.

The published data cover the period from 1969 to 1986 [39]. However, the information relative to the two rental services sectors (respectively Auto-Truck and Equipment) is not given separately prior to 1971. Similarly, data for the Food Retailing (Non-Convenience) sector is unavailable before 1971. Finally, data for the Non-Food Retailing sector in 1969 and 1970 are inconsistent with the rest of the series. For all these reasons, the calculations described below are based on data from 1971 to 1986 only.

4.2. The measures

The first two columns of Table 2 show the mean average sales per outlet, $\bar{S}$ and the mean number of outlets, $\bar{N}$, for each of the 14 Business Format

[39] In fact, data for 1987 and 1988 are also published in the February 1988 edition of Franchising in the Economy. However, these data are "estimated by respondents" rather than reports of actual sales and outlet numbers.

Note to Table:

>

Note: In the headings, *SE* stands for the standard error of the $\hat{\epsilon}$'s, which are the error terms from the regression of average sales per outlet on time, $\bar{N}$ represents the average number of outlets in the sector over the 1971–1986 period, $\bar{S}$ is the mean of average sales per outlet in the sector over the same period, and % Disc. is the proportion of discontinued units in the sector.

Table 2
Measures of risk for each franchising sector

Sectors	$\bar{S}$ 1971–86 \$000	$\bar{N}$ 1971–86	SE 1971–86	$SE/\bar{S}$ 1971–86	$SE\sqrt{\bar{N}}/\bar{S}$ 1971–86	% Disc. 1982–86 mean/year
Rental Services (Equipment)	170.7	1765	19.21	0.11	4.73	4.73
Business Aids and Services	134.7	34780	15.59	0.12	21.58	4.02
Educational Products and Services	117.0	3543	15.04	0.13	7.65	3.66
Hotels, Motels and Campgrounds	1205.1	6427	141.76	0.12	9.43	3.53
Const., Home Improv., Maintenance and Cleaning Services	113.3	13841	24.10	0.21	25.03	3.50
Retailing (Non-Food)	282.9	40175	33.87	0.12	24.00	3.24
Miscellaneous	153.3	3290	37.34	0.24	13.97	3.18
Laundry and Drycleaning Services	83.9	3066	11.18	0.13	7.38	3.02
Automotive Products and Services	173.7	41521	20.03	0.12	23.50	2.88
Convenience Stores	447.8	13808	52.18	0.12	13.69	2.84
Fast-Food and Restaurants	413.9	54456	13.81	0.03	7.79	2.38
Retailing (Food, Non-Convenience)	334.9	14319	61.74	0.18	22.06	2.30
Rec., Entertainment and Travel	123.9	4772	62.82	0.51	35.02	2.03
Rental Services (Auto-Truck)	339.5	8291	47.81	0.14	12.82	1.71

franchising sectors. The sales data in Column 1 are used to normalize the variance measures [40]. The outlet data in Column 2 are used to generate $\hat{V}_{\bar{S}} \cdot \bar{N}$. As can be seen from this table, both the mean average sales per outlet and the mean number of outlets per sector vary substantially across sectors. This implies that normalization will matter, and also that $\hat{V}_{\bar{S}} \cdot \bar{N}$ and $\hat{V}_{\bar{S}}$ will give rather different assessments of the riskiness of businesses in these various sectors.

The next 3 columns in Table 2 show the different variance-based measures we obtained using these data [41]. The first measure shown in the table is the non-normalized standard deviation of sales around a trend. The second is the normalized version, namely $SE/\bar{S}$, where SE stands for the standard error of the regression of average sales on time and $\bar{S}$ again represents mean average sales. As expected, this measure is substantially different from the first, showing that indeed one will assess the "riskiness" of different business sectors very differently depending on whether or not one corrects for the average size of outlets in the sectors. The third of these three columns shows the normalized standard error of the regression multiplied by the square root of the average number of units in the sector, that is $\sqrt{\bar{N}} \cdot SE/\bar{S}$.

Finally, column 6 shows the yearly average proportion of discontinued outlets between 1982 and 1986 in each sector, that is the $\Sigma_{t=82}^{86}(d_t/N_t)$, where d_t stands for the number of discontinued outlets at time t. The choice of five years of data in this case is quite arbitrary, but the measure is not very sensitive to this choice [42]. The sectors in Table 2 are ranked in decreasing order of "riskiness" according to this last measure.

From Table 2, it is clear that the measures themselves are substantially different. But are the differences meaningful? Apparently, they are. For example, it is difficult to determine from Table 2 which sectors are the most and least risky. As expected, when the size of firms in a sector is disproportionately high, as in the case of the Hotels, Motels and Campgrounds industry, or very low as in the Laundry and Dry-Cleaning sector, non-normalized measures of dispersion lead to the conclusion that the former is the

[40] The mean of the detrended sales data is by definition 0, so the mean of the original series is used for normalization.

[41] We report standard deviations in the tables because variances are too large.

[42] Interestingly, the data reported here are quite consistent with those of the "Frandata Franchise Termination Report" which suggests that about 4.4% of franchisees leave their system annually (Frandata Corporation, 1992). They obtain this figure by analyzing the discontinuation rates reported in the disclosure documents of 584 leading franchise systems in the U.S. On the other hand, the data here suggest a lower rate of discontinuation for franchised businesses than that reported by Bates (1995). He finds that 35% of the young franchised firms in his sample did not survive the 5 year period between 1987 and late 1991. The corresponding failure rate of independent firms in his data is 28%.

Table 3
Correlation coefficients across sectors for alternative measures of risk (Spearman rank correlations in parentheses)

Sectors	$SE/\bar{S}$	$SE\sqrt{\bar{N}}/\bar{S}$	% Disc.
$SE/\bar{S}$	1.00		
	(1.00)		
$SE\sqrt{\bar{N}}/\bar{S}$	0.67 [a]	1.00	
	(0.46)	(1.00)	
% Disc.	−0.33	−0.33	1.00
	(−0.34)	(−0.30)	(1.00)

Notes: In the headings, SE stands for the standard error of the regression of average sales per outlet on time, $\bar{N}$ represents the average number of outlets in the sector, $\bar{S}$, the mean of average sales per outlet in the sector, and % Disc. the proportion of discontinued units in the sector.
[a]: significant at the 0.01 level.
The significance tests are performed as follows:
(1) Under normality of the two underlying variables, and under H_0:$\rho = 0$, $(r\sqrt{n-2})/(\sqrt{1-r^2})$ follows a t_{n-2}, where r is the ordinary correlation coefficient, and n, the number of observations (14 in this case).
(2) For $n \leq 30$, Spearman rank correlation tables are found for example in Gibbons (1976). Here, n is the number of sectors, i.e. 14. Note that this test does not require any distributional assumption.

riskiest, and the latter, the least risky sector. In general, the orderings of sectors according to risk based on the non-normalized measures are not identical to the orderings in terms of sizes alone, but the two are highly correlated [43]. More importantly, rankings of sectors based on the normalized standard deviations are quite different from those obtained with the equivalent measure corrected for the number of outlets in the sector. For example, Equipment Rental firms are the least risky type of businesses according to the latter measure, while these are fairly average in terms of riskiness based on the normalized standard error. Finally, discontinuation rates disagree with both measures of standard deviation considered here.

Table 3 gives the correlation coefficients, Pearson and Spearman rank correlations, between 1) the standard error of the regression of average sales per outlet on time, normalized by the mean of average sales over time; 2) the same standard error multiplied by the square root of the average number of outlets in the sector; and 3) the average proportion of discontinued outlets. The table shows that the two variance-based measures are significantly positively correlated. This is to some extent reassuring. However, as can be seen by looking at their rank correlation of 0.46 and at Table 2, they can still

[43] The correlation coefficient between SE and average sales per outlet was 0.86.

lead to fairly different rankings of sectors in terms of "riskiness". Discontinuation rates, on the other hand, are negatively correlated with both variance-based measures, a disconcerting result suggesting that empirical work based on these different measures could give very different results.

4.3. The effect of risk on franchising

We have established that the three different measures included in Table 3 lead to fairly different rankings of sectors, and are not as highly correlated with each other as one would want. In fact, discontinuation rates are negatively, though not significantly, correlated with the other two measures, suggesting that they might be capturing something different from the other two. In general, we believe this shows how important the measurement issues discussed above really are.

Despite this, Lafontaine (1992) examined the sensitivity of her empirical results to the measure of risk employed and she obtained similar regression results with each of the three measures compared above. In all cases, she found that more risk tended to raise the extent to which a franchisor relied on franchising. Moreover, more risk led to lower or unchanged royalty rates, not higher ones. Hence she concluded that her results were not consistent with implications from either pure risk sharing or one-sided moral hazard models.

We further confirmed her results concerning these various measures by reestimating some of Martin's regressions using the three measures above. Our regression results were often not measured with enough precision to be different from zero but did not show any tendency by franchisors to franchise less when they faced more risk, as the theories would predict.

The data we report in Table 4, we believe, explain in large part the results just described. The first two columns of this table again give information on sales levels and number of units, but here in terms of the ratios of these variables for company-owned and franchised units. The data in each of the next three columns are given by 1) $\frac{SE_c/\bar{S}_c}{SE_f/\bar{S}_f}$, where the f and c subscripts refer to franchised and company-operated outlets respectively, 2) $\frac{\sqrt{\bar{N}_c} \cdot SE_c/\bar{S}_c}{\sqrt{\bar{N}_f} \cdot SE_f/\bar{S}_f}$ and 3) $\Sigma_{t=82}^{86}(d_{ct}/N_{ct})/\Sigma_{t=82}^{86}(d_{ft}/N_{ft})$. Clearly, these ratios are endogenous: The extent to which company-owned and franchised units are subject to sales variability or failures is determined in part by the decision as to which outlet should be franchised or not, and the terms of the franchise contract. But basically, this table shows that when one compares the variance

Table 4
Ratios of riskiness under alternative measures of risk: company-owned versus franchised outlets

Sectors	$\frac{\bar{S}_c}{\bar{S}_f}$ 1971–86	$\frac{\bar{N}_c}{\bar{N}_f}$ 1971–86	$\frac{SE_c/\bar{S}_c}{SE_f/\bar{S}_f}$ 1971–86	$\frac{SE_c\sqrt{\bar{N}_C}/\bar{S}_c}{SE_f\sqrt{\bar{N}_f}/\bar{S}_f}$ 1971–86	$\frac{d_c/N_c}{d_f/N_f}$ 1982–86
Automotive Products and Services	3.93	0.12	0.45	0.16	0.47
Business Aids and Services	1.29	0.19	1.46	0.64	0.19
Const., Home Improv., Maintenance and Cleaning Services	7.58	0.03	2.84	0.53	0.29
Convenience Stores	0.96	1.63	3.59	4.57	0.87
Educational Products and Services	1.44	0.15	1.93	0.75	1.99
Fast-Food and Restaurants	1.25	0.40	1.81	1.15	0.77
Hotels, Motels and Campgrounds	2.25	0.19	0.99	0.44	0.81
Laundry, Drycleaning	2.67	0.04	3.26	0.64	0.12
Rec., Entertainment and Travel	6.24	0.03	1.09	0.19	0.41
Rental Services (Auto-Truck and Eq.)	4.87	0.31	0.50	0.28	0.28
Rental Services (Equipment)	3.04	0.17	1.67	0.68	0.22
Retailing (Non-Food)	1.44	0.34	1.81	1.04	0.49
Retailing (Food, Non-Convenience)	3.26	0.12	3.43	1.21	0.55
Miscellaneous	2.04	0.09	1.63	0.50	2.92

Notes: In the headings, *SE* stands for the standard error of the regression of average sales per outlet on time, $\bar{N}$ represents the average number of outlets in the sector, $\bar{S}$, the mean of average sales per outlet in the sector, and d the proportion of discontinued units in the sector. The subscripts c and f refer respectively to company-owned and franchised units.

of sales measures and the discontinuation rates between company-owned units and franchised ones, in most cases one finds more "risk" in the franchised units than in the corporate ones. This in turn implies that a regression of the use of franchising on the amount of risk in the sector will likely yield a positive relationship. But the issue then becomes: Do we observe more sales variability and more failures among franchised units because more risky units are franchised (that is, the usual interpretation in the current literature, where franchisors "shed" risk), or because franchising, in itself, leads to decisions that make for greater sales variability (and hence possibly more failures)? We show in the next section that the latter interpretation is at least as reasonable an explanation for the existence of this positive relationship as the former. In that sense, we question the way in which conclusions have been drawn from the data to date.

5. A theoretical model

Greater sales variability in franchised units than in corporate units, and lower total sales in franchised than in corporate units, are two patterns found in the data above that have been interpreted in the literature to mean that franchisors shed less profitable and more risky units. In this section, we discuss how these can arise from a simple double-sided moral hazard model with information asymmetry. In fact, we show that greater sales variability in franchised units can be due to efficiency considerations: it occurs in our model because of the incentives embedded in the franchise contract. In showing that these data patterns can arise in circumstances that have nothing to do with "risk-shedding", we cast doubt on how they have been interpreted in the past.

There have been various attempts to model the franchisor–franchisee relationship. These attempts have used a variety of approaches from signalling and screening to single-sided and double-sided moral hazard [44]. What these attempts have in common is that they all model the franchisor as a person who has monopolistic access to some productive technology but has to utilize the services of another party to productively utilize this technology. The franchisor is assumed to make a take-it-or-leave-it offer of a contract to the franchisee. In doing so, the franchisor takes into account the incentives and participation constraints that the franchisee may face. We also follow this general approach.

The model we have in mind is one of double-sided moral hazard with information asymmetry. (See the appendix for a more detailed description of the model.) We begin by assuming that both parties are risk neutral, thereby eliminating any need to allocate risk efficiently. Hence franchising arises here completely independently from the motive of risk allocation. Later, we also argue that our results hold under franchisee risk aversion. Hence despite the fact that efficient risk allocation would require that more risk be borne by the franchisor, we still find that sales variability may be greater, and sales levels lower, in franchised than in company-owned units.

We have argued elsewhere that there are compelling reasons to model the franchise relationship as one that involves double-sided moral hazard [45]. Thus, we assume that both the franchisor and the franchisee/employee provide efforts that are unverifiable to the other unless some monitoring technology is employed upfront. We denote the franchisor's effort by r and the franchisee/employee's effort by e. We therefore assume that the produc-

[44] See namely Rubin (1978), Mathewson and Winter (1985), Lal (1990), Bhattacharyya and Lafontaine (1995), Gallini and Lutz (1992), and Minkler (1992).

[45] See Bhattacharyya and Lafontaine (1995).

tion function of a unit can be written as (ignoring the i subscript for simplicity) [46]:

$$X = K \cdot f(e,r) + \epsilon \tag{20}$$

where X is the total monetary return produced and K is a random term with mean $\bar{K}$ and variance σ_k^2. K denotes the level of a productivity shock that only the franchisee gets to observe. This informational advantage of the franchisee over the franchisor is the main factor that will lead the franchisor, under certain circumstances, to franchise the unit. The franchisee can use his information about K to make better decisions in terms of his own effort. We assume that the franchisor never gets to observe K and make input decisions based on its realization. As a result, there is always a source of inefficiency in company-operated units relative to franchised units, an inefficiency that will be the more costly to the firm the more important information about K is. On the other hand, under a franchising arrangement, the franchisor will have to provide incentives for the franchisee to utilize his information to mutual benefit. The provision of these incentives will also be costly to her.

The ϵ term in the production function denotes environmental uncertainty and has mean 0 and variance σ_ϵ^2. The uncertainty inherent in utilizing the productive technology is then composed of two terms –the uncertainty about productivity and the uncertainty about the general environment. It is possible to think of the first term as denoting supply-side uncertainty and the second term as representing demand-side uncertainty. We take $f(\cdot,\cdot)$ to be a standard neoclassical production function. We further assume that $f(0,r) = 0$ and $f(e,0) = 0$. This assumption simply emphasizes the team production aspect of the production technology by stating that both inputs are required for any production to occur. We also assume that the franchisor's and the franchisee/employee's disutility of effort functions are separable and given by $U(r)$ and $V(e)$ respectively, both of which we assume to be increasing and strictly convex in effort.

As a base case, we assume that all parties are risk-neutral. In this setting, it can be shown that, if e is not monitored, the optimal outcome can be implemented via contracts that are linear in output [47]. As a result, we confine our attention to such contracts for the franchise relationship.

We envision a company-operated operation to involve the hiring of an employee whose efforts are monitored. Of course, such monitoring is only possible if the franchisor sets up a monitoring system, which we assume is

[46] In the model we abstract away from the question as to whether sales revenue or profit should be shared in this type of arrangement. As a result, we use X to denote any verifiable output that is shared in the relationship.

[47] See Bhattacharyya and Lafontaine (1995) for a proof.

costly [48]. In fact, consistent with much of the existing literature, we assume that the cost of monitoring goes up with the physical distance of the operations from the franchisor's home base. The franchisor's choices are then straightforward. She can either employ a manager for a unit whose effort she monitors perfectly, in which case she offers an effort contingent contract. However, because the franchisor never observes K, she must choose the level of effort she requires from the manager without relying on this information. In addition, she must incur the cost of monitoring the effort provision of the employee. It should be evident, then, that such an organization of affairs should be optimal when the information on productivity is not very important and/or when the monitoring technology is not very costly. Alternatively, she can franchise a unit, which implies that she will not monitor but will instead offer contract terms that provide incentives for the franchisee to contribute appropriate information-contingent effort levels. Where the value of the information is high enough, franchising will be preferred over employment and vice versa. In addition, the model shows that franchisee-run operations generate, on the whole, more variable output than franchisor-run operations. Finally, franchisee-run units are also smaller in scale than franchisor-run units. Both these observations are consistent with the empirical evidence presented earlier.

The choice between company operation (or effort-contingent compensation) and franchising (or output-contingent compensation) in this model is, in fact, a more general phenomenon than what may be apparent up to now. Effort-contingent compensation has the advantage of letting one party have first-best incentives to maximize the level of the pie, but suffers from the defect that it does not provide any incentives for the informationally advantaged party to divulge his information for the common good. Output-contingent compensation, on the other hand, provides only second-best effort incentives to both parties, but gives incentives to the informationally advantaged party to productively utilize his information. Put another way, an effort-contingent contracting scheme does not distort *ex-ante* incentives but causes departures from productive efficiency at the *interim* stage. An output-contingent scheme, however, cannot escape from the *ex-ante* inefficiency caused by the inability of fractional sharing rules to provide first-best incentives to all parties. However, it does promote a greater degree of productive efficiency at the *interim* stage by promoting the adjustment of effort provision by the party observing the information. Whenever the value of productivity information is high enough, it makes sense, then, to adopt an output-contingent contracting scheme. Not surprisingly, this would happen

[48] We use the term "franchisor" to denote the person who has ownership rights over the productive technology, irrespective of whether she chooses to franchise or to hire an employee.

Table 5
Simulation Results: Cobb–Douglas case with exponential utility functions

	Case 1	Case 2	Case 3	Case 4
Parameters:				
α	0.5	0.1	0.1	0.1
γ	0.5	0.9	0.9	0.9
m	2.0	2.0	2.0	2.0
n	2.0	2.0	2.0	3.0
δ_r	0.5	0.5	0.5	0.5
δ_e	0.5	0.5	1.0	0.5
Max K	10.0	10.0	10.0	10.0
Min K	1.0	1.0	1.0	1.0
Results:				
β	0.5	0.80	0.80	0.83
Co-Own – e	5.0	4.43	2.30	4.32
Franch. – e(max)	3.29	5.73	2.97	5.93
Franch. – e(min)	1.58	0.71	0.37	0.73
Co-Own – r	5.0	1.48	1.08	1.11
Franch. – r	2.54	0.60	0.44	0.60
Co-Own – X	50.0	21.85	11.71	20.76
Franch. – E_X	25.71	18.02	9.66	18.56
Co-Own – X(max)	75.0	39.74	21.29	37.75
Franch. – X(max)	43.33	45.77	24.53	47.14
Co-Own – X(min)	25.0	3.97	2.13	3.77
Franch. – X(min)	10.01	0.70	0.37	0.72
Co-Own – $Var(X)$	208.33	106.58	30.61	96.17
Franch. – $Var(X)$	93.43	177.38	50.94	188.18
Co-Own – $CV(X)$	0.29	0.47	0.47	0.47
Franch. – $CV(X)$	0.38	0.74	0.74	0.74
Co-Own – Surplus	25.0	10.93	5.86	10.73
Franch. – Mean Surplus	19.28	11.37	6.09	11.52

Assumes $X = K(e^{\gamma} r^{\alpha}) + \epsilon$, $U(r) = \delta^{r} \cdot r^{n}$, and $V(e) = \delta_e \cdot e^{m}$. In the results section, β is the optimal royalty rate for franchised units, e stands for employee/franchisee effort, r is the franchisor effort, X is output, $Var(X)$ and $CV(X)$ are the variance and coefficient of variation of output respectively, and *Surplus* stands for value to the franchisor which is equal to the total surplus generated by the relationship.

here when the efforts of the party accepting the contract are of primary importance to the production process, and/or when the variance of K is high. Note that this trade-off would exist even with a costless monitoring technology.

We show how this trade-off operates in Table 5, which presents simulation results obtained under specific parameterizations of the production and cost of effort functions. More specifically, these results assume the function $f(e,r)$ to be of the Cobb–Douglas form, so that

$$X = K(e^{\gamma} r^{\alpha}) + \epsilon \tag{21}$$

and the cost of effort functions to be $U(r) = \delta_r \cdot r^n$ and $V(e) = \delta_e \cdot e^m$. As in Bhattacharyya and Lafontaine (1995), the optimal β parameter in this parameterized version of the model turns out to be independent of the parameters δ_r and δ_e. In addition, it is also independent of any of the distributional characteristics of the parameter K. That is, the optimal sharing parameter is not influenced by either the mean or any higher moments of the distribution of K [49].

Table 5 shows the effort levels, surplus generated and the variability of output using various values of the exogenous parameters [50]. In these simulations, the parameter K is always assumed to be uniformly distributed over the interval [1,10]. Similarly, the cost of the monitoring technology is always assumed equal to zero. The first column in this table contains results obtained under symmetry, namely with $\alpha = \gamma = 0.5$, and m and n both set equal to 2. The next 3 columns in the table give results when $\alpha = 0.1$ and $\gamma = 0.9$, for 3 different combinations of δ_r, δ_e, m, and n. We report the value to the principal of utilizing effort-contingent contracting as *Co-Own Surplus* and the value under output-contingent contracting as *Franch. Surplus*.

The results are consistent with the intuition presented above regarding the trade-off between effort-contingent contracts and output-contingent contracts. They show that when both parties' efforts are equally important to the production process, the possessor of the productive technology finds it optimal to utilize effort-contingent contracting. Where the informationally advantaged party's effort is especially important, the franchisor prefers using output-contingent contracting.

Furthermore, in all cases presented in Table 5 the coefficient of variation of the production under a franchise contract is higher than that under an employer–employee relationship. This is due to two effects. First, the franchise contract suffers from the provision of second-best incentives and, hence, produces a lower level of mean output. Second, it allows for flexibility in effort provision by the informed party and, thus, leads to variability in effort provision which, in turn, leads to variability in production. In addition, the franchisee's disutility of effort is influenced by the information via the effect it has on effort. As a result, the coefficient of variation is uniformly higher under franchising. However, although mean production is always lower under franchising, that does not imply that the franchisor is worse off under this type of contract. The productive efficiency that is achieved by utilizing the productivity information goes to reduce the compensation paid

[49] Both in the employer–employee formulation and the franchisor–franchisee formulation, the Cobb–Douglas form introduces another regularity. With $\sigma_\epsilon^2 = 0$, the coefficient of variation of sales depends only on the statistical properties of K. This feature shows up below in Table 5.

[50] Our simulations were done with $\sigma_\epsilon^2 = 0$. The qualitative results do not depend on such an assumption.

to the franchisee since he incurs much less effort costs in the less productive states of nature. The convex nature of these costs implies substantial savings that, in some cases, can obliterate the benefits associated with the first-best incentives that obtain in the employer/employee relationship.

Incorporating the cost of monitoring the manager employee into the arguments above alters the point at which regime switching in contractual choice occurs. We have two effects in mind here. First, we expect monitoring costs to go up with distance from the home base, as mentioned earlier. Second, we think the information advantage of the employee/franchisee is likely to be higher the further is the venue of operation of the technology from the home base of the franchisor. This can be modeled as an increase in the variability of K, leaving the mean level $\bar{K}$ unchanged. The first effect diminishes the profitability of company operations with distance, and the second effect enhances the profitability of franchise operations with the same. As a result, at a certain distance from her home base, the franchisor will find it optimal to switch from an employment to a franchise relationship. This switch will be accompanied by a drop in the mean level of production, and a jump in the coefficient of variation of production, a result consistent with the empirical evidence presented earlier.

Finally, we have presented here a model in which both parties were assumed to be risk neutral. However, the same intuition would apply to the case where the franchisee is risk-averse. More specifically, assume that the informed party has a mean-variance utility function as defined earlier [51]. In addition, assume that the variability in output is principally due to variation in K rather than variation in ϵ [52]. Consider first the case of the employment contract. Since the compensation in this regime is not variable, our analysis requires no change. The franchise contracting regime undergoes significant changes in analysis since the agent will have to be compensated for bearing risk. However, the production decision, *at the moment that the franchisee undertakes his effort*, does not involve a substantial deviation from our earlier analysis since the franchisee by then knows the level of K realized! Hence, even with risk aversion on the part of the franchisee, the productivity enhancement caused by the utilization of the asymmetric information remains substantially intact. What changes is the compensation that has to be paid to the franchisee in order to make his participation worthwhile. With a low enough degree of risk aversion on the part of the franchisee, the loss to the franchisor on this account is going to be correspondingly low. As a result,

[51] The use of mean-variance utility functions is widespread in Finance. It can be shown that this is consistent with a CARA utility function and normally distributed uncertainty.

[52] That is, assume σ_ϵ^2 is almost equal to zero.

regime switching should occur even with risk aversion on the part of the franchisee.

What is important to note in this model is that the uncertainty engendered by K does more than just impose risk on the franchisee. Utilization of this information by means of a suitably designed incentive contract also allows for a more efficient alignment of incentives, leading to more efficient production arrangements. This productivity enhancement can, under suitable parameters, overwhelm the cost of offering insurance to the franchisee and make an output-sharing franchise contract optimal. Again, the switch from company ownership to franchising is accompanied by an increase in the variability of output and a drop in the mean level of output.

We conclude that the identification of more variable production in a franchise situation need not imply that the franchisee is less risk averse than the franchisor, and that the franchisor "sheds" risky units. The key insight that we have utilized is that variability in production may be due to two separate effects. The first effect, which is due to variations in exogenous demand or supply side parameters, does indeed only serve to increase the risk borne by the franchisee. The second effect, which arises due to the franchisee *efficiently adjusting his effort provision in line with the exogenous shocks about which he has private information*, or put differently due to the endogeneity of the franchisee's effort decision, does more than just increase the risk he bears. The more efficient utilization of payoff-relevant information may very well overwhelm the risk effect that much of the past literature has concentrated on.

6. Conclusion

The goal of this paper was first to provide an assessment of the existing empirical literature on the role of risk in franchising. In doing so, we discussed a number of issues related to the measurement of risk which we feel are important and make it difficult to interpret much of the existing empirical literature on this subject. We concluded that discontinuation rates might be a better way to measure risk in franchising than variance based measures.

Using discontinuation rates as well as some variance-based measures, we then showed that there are patterns in the data that seem to imply that franchisees bear more risk than franchisors. Under models based on efficient risk allocation this leads to the conclusion that franchisors are more risk averse than franchisees, a conclusion we find unappealing given their respective sizes and differential access to capital markets. As a result, we followed Goldberg's (1990) suggestion and looked for an explanation that is not fundamentally based on preferences towards risk. Some models of franchising, namely those based on double-sided moral hazard, have already typically deemphasized the role of risk (see Rubin, 1978, Lal, 1990 and Bhattacharyya and Lafontaine, 1995). Starting from a similar framework, we showed that a model emphazising franchisee information, and his efficient utilization of this information, can give rise to the kind of patterns found in the data, namely, more output variability under franchising than under company management, and lower sales in franchised than in corporate units. However, in our model, these results have nothing to do with franchisors alloting less profitable or riskier units to franchisees. Rather they arise as a result of the incentives embedded in the franchise contract for both the franchisee and the franchisor. By showing that these data patterns can arise in circumstances that have nothing to do with "risk-shedding", we cast doubt on how they have been interpreted in the existing literature.

In general, our model emphasizes among other things that "variability" and risk are two separate things. It also illustrates what we believe to be a promising approach to modelling some aspects of the franchising relationship and others similar to it. This is because it focuses on the franchisee's comparative advantage in obtaining local market information, an advantage that corresponds quite well to what franchisors point to when pressed to explain why franchisees would buy "risky" units that franchisors choose not to operate themselves. We hope that the present paper will foster more work in similar directions.

Appendix A

Appendix: The Model

We examine the choice of franchising versus direct operation in the context of a double-sided moral hazard model with information asymmetry. Thus, we assume that both the franchisor and the franchisee/employee provide efforts that are unverifiable to the other. We denote the franchisor's effort by r and the franchisee/employee's effort by e. We assume that the production function can be written as

$$X = K \cdot f(e,r) + \epsilon \tag{22}$$

where X is the total monetary return produced and K is a random term with mean $\overline{K}$ and variance σ_k^2. K denotes the level of a productivity shock that only the franchisee gets to observe. The ϵ term denotes environmental uncertainty and has mean 0 and variance σ_ϵ^2. We take $f(\cdot,\cdot)$ to be a standard neoclassical production function. Letting subscripts denote partial derivatives, this implies that f_e and f_r are positive, that f_{ee} and f_{rr} are negative, and that f_{er} is positive. We further assume that $f(0,r) = 0$ and $f(e,0) = 0$. This assumption simply emphasizes the team production aspect of the production technology. We also assume that the franchisor's and the franchisee/employee's disutility of effort functions are separable and given by $U(r)$ and $V(e)$ respectively, both of which we assume to be increasing and strictly convex in effort.

As a base case, we assume that all parties are risk-neutral. In this setting, it can be shown that, if e is not monitored, the optimal outcome can be implemented via contracts that are linear in output. As a result, we confine our attention to such contracts for the franchise relationship.

The Franchisor–Employee Relationship

We first set up the problem for the franchisor when she elects to monitor the effort of the other party and enter into an employer–employee relationship. We assume that such a contract calls for an effort level of e^* for which a fixed wage of w is paid. The monitoring technology can verify whether or not e^* was contributed and, hence, large fines can be imposed to sustain this level of effort provision once the employee agrees to the contract [53]. In this setting, the franchisor does not have to worry about satisfying any incentive constraint on the part of her employee and the optimal program can be written as

$$\max_{e,r,w}\left[\bar{K}\cdot f(e,r) - w - U(r) - Z(d)\right] \tag{23}$$

subject to

$$w - V(e) \geq 0 \tag{24}$$

where $Z(d)$ is the cost of monitoring the employee, which we assume is increasing in the distance of the unit from monitoring headquarters, and where the employee's reservation utility level is normalized to zero. Note that due to the global risk neutrality assumption, the multiplier in the constraint will turn out to be 1. Hence, the optimal program can be decomposed into two stages. In the first stage, the program to be solved is one that maximizes the joint surplus, that is

$$\max_{e,r}\left[\bar{K}\cdot f(e,r) - U(r) - V(e) - Z(d)\right]. \tag{25}$$

[53] We do not consider incentive contracts that can condition compensation based on output for effort levels above the minimum one specified. Doing so would complicate the analysis without changing any of the main results.

The second stage sets the wage level w to just satisfy the agent's reservation utility level given the optimal effort level e^* dictated by the employer.

In short, then, we are assuming that the employer–employee relationship can be modeled as a constrained first-best problem in which the constraint is that effort levels are chosen before the actual level of productivity is known. Hence, the employer–employee relationship does not utilize the information that may be available to the employee before he is required to contribute his effort. The first-order conditions for the maximization of joint surplus can be written as

$$\overline{K}\cdot f_e(e,r) = V'(e) \tag{26}$$

and

$$\overline{K}\cdot f_r(e,r) = U'(r). \tag{27}$$

Denoting the arguments to the solution of the problem by e^* and r^*, we have the following expressions:

$$E(X) = \overline{X} = \overline{K}\cdot f(e^*,r^*) \tag{28}$$

$$Var(X) = \left(\frac{\overline{X}}{\overline{K}}\right)^2 \cdot \sigma_k^2 + \sigma_\epsilon^2 \tag{29}$$

$$Coef.Var(X) = \frac{\overline{X}}{\overline{K}^2}\cdot \sigma_k^2 + \frac{\sigma_\epsilon^2}{\overline{X}}. \tag{30}$$

Note that e^* and r^* are both independent of σ_k^2 and σ_ϵ^2 due to our risk-neutrality assumptions. Note also that a mean preserving spread in either the supply- or the demand-side uncertainty increases both the variance in production and the coefficient of variation without affecting the mean production level. This feature is caused by the fact that the employee does not get to choose his effort level optimally contingent on the productivity information that he may have.

The Franchisor–Franchisee Relationship

Having modeled the employer–employee relationship, we now turn to modeling the franchisor–franchisee relationship. Note that now there is no monitoring of effort on the part of the franchisor. Instead, due to the double-sided moral-hazard problem, the optimal contract has to be based on revenue sharing, as in Bhattacharyya and Lafontaine (1995), and the choice of the sharing parameter has to be determined in accordance with the incentive constraints of both parties. Where we depart from Bhattacharyya and Lafontaine (1995) is that we now have the agent choose his optimal action after he observes the realization of the productivity shock K. As a result, the optimal program looks as follows:

$$\max_{\beta,F,e(K),r} F + \beta \cdot E_{K,\epsilon}[K \cdot f(e(K),r) + \epsilon] - U(r) \tag{31}$$

subject to

$$\begin{array}{ll} \text{i)} & (1-\beta) \cdot K \cdot f_e(e(K),r) = V'(e(K)) \\ \text{ii)} & \beta \cdot E_K[K \cdot f_r(e(K),r)] = U'(r) \\ \text{iii)} & (1-\beta) \cdot E_{K,\epsilon}[K \cdot f(e(K),r) + \epsilon] - F - E_K[V(e(K)))] \geq 0 \end{array}$$

where $E_s(\cdot)$ denotes the expectation operator with respect to s.

The introduction of the asymmetric information on the realization of the productivity parameter K causes several departures from a standard double-sided moral hazard framework. First, the franchisee's incentive compatibility constraint has to take into account the information that he has in his possession. As a result, the franchisee's effort choice is now a function of K. We will denote the functional dependence of the franchisee effort choice under the optimal contract by $\bar{e}(K)$. The franchisor, however, not being aware of the realization of K, does not get to choose an effort level that is contingent on it. Hence, the optimal effort level she chooses –denoted by $\bar{r}$ –now has to respect both the fact that the franchisor has to be provided incentives to contribute effort, and that she has to take into account the dependence of the franchisee's effort level on the realized value of K. Lastly, since the contract is signed before the realization of K, the participation constraint of the franchisee has to account for the expected cost of the effort that the franchisee has to incur. By invoking this form of an *ex-ante* participation constraint, we are, of course, assuming that the franchisee cannot quit after observing a "bad enough" realization of K [54]. As a result, in some states of the world, the franchisee would be strictly worse off than if he had not entered into the franchise relationship. On the other hand, in other states of the world, he would be strictly better off than not participating in the relationship. Note, however, that *given the lack of an exit option*, he has incentives to provide the effort called for instead of just providing zero effort and saving on his disutility of effort.

This formulation of the problem is responsible for three differences between the employer–employee relationship and the franchisor–franchisee

[54] One could allow the franchisee's participation constraint to be based on *interim* information rather than *ex-ante* information. This would significantly complicate the analysis and the major qualitative change would be that *ex-ante* rents would be left with the franchisee. Kaufmann and Lafontaine (1994) provide some evidence that there are *ex-ante* rents left downstream at McDonald's. At the present stage, however, we stick to a simpler formulation. We plan to extend our modeling to the case of *interim* participation constraints in the future.

relationship. First, the monitoring technology is no longer being used. Second, the lack of monitoring generates the need to use an outcome-contingent contract that respects incentive constraints for its implementability. The need to respect these additional constraints obviously imposes a cost from the point of view of productive efficiency compared to the earlier case. However, this effect can be offset by the third factor which is the positive effect on productive efficiency that is brought about by allowing the franchisee to adjust his effort provision levels contingent on the level of information that he has about his productivity.

With the specific functional forms used in the simulations, and with our current formalization of the participation constraint with global risk neutrality, the multiplier of the participation constraint is again 1, so that the program under the franchise relationship can be rewritten as:

$$\max_{\beta, e(K), r} E_K\left[K \cdot e(K)^{\gamma} r^{\alpha}\right] - \delta_r \cdot r^n - E_K\left[\delta_e \cdot e(K)^m\right] \tag{32}$$

subject to

i) $(1-\beta) \cdot \gamma \cdot K r^{\alpha} e(K)^{\gamma-1} = m \cdot \delta_e \cdot e(K)^{m-1}$

ii) $\beta \cdot \alpha \cdot r^{\alpha-1} E_K\left[K \cdot e(K)^{\gamma}\right] = n \cdot \delta_r \cdot r^{n-1}$.

This parameterized problem can be solved algebraically for the optimal β. As in Bhattacharyya and Lafontaine (1995), the optimal β parameter turns out to be independent of the parameters δ_r and δ_e. In addition, it is also independent of any of the distributional characteristics of the parameter K. That is, the optimal sharing parameter is not influenced by either the mean or any higher moments of the distribution of K. This program can then be solved numerically for the optimal effort levels and surplus to the franchisor, as was done to generate the results in Table 5.

References

Allen, D.W. and D. Lueck, 1994, Risk, Uncertainty and Contracts, mimeo, Louisiana State University.

Banerji, S. and C. Simon, 1992, Franchising Vs. Ownership: A Contracting Explanation, mimeo, University of Chicago.

Bates, T., 1995, Analysis of Survival Rates Among Franchise and Independent Small Business Startups, Journal of Small Business Management, 33.

Bhattacharyya, S. and F. Lafontaine, 1995, Double-Sided Moral Hazard and the Nature of Share Contracts, RAND Journal of Economics, forthcoming.

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

Cheung, S., 1969, The Theory of Share Tenancy (University of Chicago Press, Chicago).

Forbes, 1984, No Entrepreneurs Need Apply, Dec. 3, 124.

Gallini, N.T. and N.A. Lutz. 1992, Dual Distribution and Royalty Fees in Franchising, Journal of Law, Economics, and Organization 8, 471.

Gibbons, J.D., 1976, Nonparametric Methods for Quantitative Analysis (Rinehart and Winston, Holt).

Goldberg, V.P., 1990, Aversion to Risk Aversion in the New Institutional Economics, Journal of Institutional and Theoretical Economics 146, 216.

Kaufmann, P.J. and F. Lafontaine, 1994, Costs of Control: The Source of Economic Rents for McDonald's Franchisees, Journal of Law and Economics, 37, 417.

Knight, F., 1921, Risk, Uncertainty, and Profit (Houghton Mifflin, New York).

Lafontaine, F., 1992, Contract Theory and Franchising: Some Empirical Results, Rand Journal of Economics 23, 263.

Lafontaine, F. 1993, Contractual Arrangements as Signaling Devices: Evidence from Franchising, Journal of Law, Economics, and Organization 9, 256.

Lal, R., 1990, Improving Channel Coordination through Franchising, Marketing Science 9, 299.

Lepper, S., 1967, Effects of Alternative Tax Structures on Individuals' Holdings of Financial Assets, in: D.D. Hester and J. Tobin, eds, Risk Aversion and Portfolio Choice, Cowles Foundation Monograph 19 (John Wiley and Sons).

Makover, H. and J. Marschak, 1938, Assets, Prices and Monetary Theory, Economica 5, 261.

Martin, R.E., 1988, Franchising and Risk Management, The American Economic Review 78, 954.

Mathewson, G.F. and R.A. Winter, 1985, The Economics of Franchise Contracts, Journal of Law and Economics 28, 503.

McAfee, R.P. and M. Schwartz, 1994, Opportunism in Multilateral Vertical Contracting: Nondiscrimination, Exclusivity, and Uniformity, American Economic Review 84, 21.

Minkler, A.P., 1992, Why Firms Franchise: A Search Cost Theory, Journal of Institutional and Theoretical Economics 148, 240.

Norton, S.W., 1988, An Empirical Look at Franchising as an Organizational Form, Journal of Business 61, 197.

Rubin, H., 1978, The Theory of the Firm and the Structure of the Franchise Contract, Journal of Law and Economics 21, 223.

Sen, K.C., 1993, The Use of Initial Fees and Royalties in Business Format Franchising, Managerial and Decision Economics 14, 175.

Smith II, R.L., 1982, Franchise Regulation: An Economic Analysis of State Restrictions on Automobile Distribution, Journal of Law and Economics 25, 125.

Thompson, D.N., 1971, Franchise Operations and Antitrust, (Heath Lexington Books).

U.S. Department of Commerce, 1988, Franchising in the Economy, prepared by Andrew Kostecka. Also other editions (Washington, Government Printing Office).

Williams, D., 1994, Why Do Entrepreneurs Become Franchisees?: An Empirical Analysis of Organizational Choice, mimeo, Department of Economics, UCLA.

Williams, D., 1995, A Comparison of Business Discontinuations in States With and Without Franchise Termination Laws, Proceedings of the 9th Conference of the Society of Franchising, Institute for Franchise Management, University of St. Thomas.