ESSAYS ON ASYMMETRIC INFORMATION

By

Kevin James Wainwright

A thesis submitted in partial fulfillment
Of the requirement for the degree of

DOCTOR OF PHILOSOPHY

In the
Department of Economics

SIMON FRASER UNIVERSITY

2007

APPROVAL

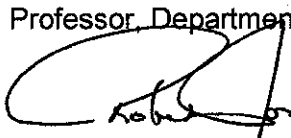
Name: **Kevin James Wainwright**

Degree: **Doctor of Philosophy**

Title of Thesis: **Essays on Asymmetric Information**

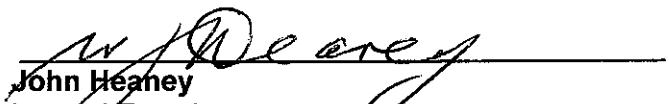
Examining Committee:

Chair: David Andolfatto
Professor, Department of Economics



Robert Jones
Senior Supervisor
Professor, Department of Economics

Douglas Allen
Supervisor
Professor, Department of Economics



John Heaney
Internal Examiner
Associate Professor, Faculty of Business Administration

Merwan Engineer
External Examiner
Professor, Department of Economics
University of Victoria

Date Defended/Approved: July 31, 2007

ABSTRACT

ESSAYS ON ASYMMETRIC INFORMATION

BY

KEVIN WAINWRIGHT, BA, MA (*SFU*)

Doctor of Philosophy

Simon Fraser University, 1998

This thesis is a collection of three papers with a common theme of asymmetric information arising from costly measurement. The first paper is an analysis of strategic investment in legal services to influence the magnitude of settlements in torts. The problem is modelled as a two-stage game. In the first stage the frequency and level of care is chosen. In the second stage a tort occurs, investments in legal services are made and a Nash equilibrium in settlement is reached. The outcome of stage two determines the levels of the choice variables in stage one. The results of the two stage game are compared to those that would arise if chosen by a social planner to maximize welfare.

The second paper presents an analysis of a phenomenon known as "The Relative Age" effect. When assessing the innate ability (or talent) of individual children who are grouped into age cohorts, systematic errors occur due to differences in biological maturity. A structural model of a multi-period progression through levels (or grades)

that employs screening and selection is developed. Through a series of simulations, impact of the relative age on the of selection process is analyzed.

The final paper extends the work of Mathewson and Winter (1985) in the field of franchising. Given the hypothesis that a franchise contract ensures quality compliance at a lower cost relative to alternative organizational structures, the existence of dual organizational structures within the same franchise chain is inadequately explained. This paper extends the basic model of Mathewson into a spatial framework, demonstrating that nonconvexities in monitoring costs will produce dual organizational structures within the same chain.

DEDICATION

James

Mary-Lou

Lou-Ann

Austin

...and Jen

ACKNOWLEDGMENTS

I would like to thank Curtis Eaton, my original supervisor, for helping me first formulate and identify the topics in this thesis, as well as for showing me his unique approach for breaking down a complex issue into the essence of the problem. A special thanks to my current committee, Robbie Jones and Doug Allen for being there to guide me through to the very (delayed) end. I would like to thank Mike Bowe, who was my first mentor and who first encouraged me to follow an academic path. My appreciation to the many faculty and fellow graduate students who's comments and discussions helped shape the way I thought about economic problems.

I would like to acknowledge all those in the department (past and present) that supplied support, encouragement and friendship along the way: Larry Boland, Nancy Olewiler, James Dean, Clyde Reed, Stephen Easton, Sherrill Ellis, Barbara Clark, and Gisela Seifert (as well as those I missed but deserve to be on this list).

A special acknowledgement is reserved for Jennifer Yoe, who had a gift for channelling my random thoughts onto paper, the significance of her support is much more than she realizes. Finally, I would like to thank Laura Nielson for putting up with me at the end....

CONTENTS

LIST OF FIGURES

PREFACE

In my first economics class, the professor told me how economics can be applied in everyday life. While I was engrossed with the subject I never did see the connection between what was taught to me in that class and what went on each day. By the time I reached graduate studies, I'd become technically proficient at economics, but I still never quite felt the connection to everyday life that the professor first told me about. It was probably my Ph.D. methodology class where I began to notice that the "light was coming on" and I began to notice how many everyday things around me had interesting economic applications. I became important to me that my thesis embody this spirit.What appeals to me about each of the three essays is that the original idea in each case was derived from my own experience; which has given me the sense of ownership of each topic necessary to complete this task.

The first paper entitled "Dogs of War" came from a summer consulting contract with a property developer while in graduate school. At the same time, the developer was in an ongoing dispute with the local municipality who kept changing the terms of the land use agreement. After yet another stakeholder meeting that failed to resolve the dispute, the vice-president of development said "Enough of this, I'm turning loose the dogs of war". I asked him to clarify what he meant, and he explained that his company had an expensive law firm on full retainer so there was no additional cost to his decision. The city would need to acquire legal representation in response to the developer; which would be costly. Further, the law firm had a practice of requesting legal opinions from other law firms with similar expertise'. Once a legal opinion is given, it disqualifies the lawyer from representing the other side, thereby reducing the

available pool of law firms that the city could retain.

The strategic use of legal services that was very different than the role lawyers had typically been modeled in law and economic papers. In questioning contacts in the legal profession, it became apparent that the strategy used by the property developer was probably more prevalent than I realized. Since most of these cases ended in settlement which, by its nature, implies a certain level of confidentiality, the challenge would be in finding supporting data. However it seemed that this was a story worth exploring; therefore it became the motivation for the first paper.

The second paper in this thesis, entitled "The Tyranny of Chronological Age" arose when my original thesis supervisor, Curtis Eaton, introduced me to Professor Jay Allen of the University of Lethbridge economics department. Jay's current research interest was relative age effects in sports and education; an area that had been extensively documented by researchers in the fields education and psychology. However, little in the way of formal modeling had occurred.

From those discussions with Curt and Jay, it seemed that the next step in this was to build a structural model of the relative age effect that could isolate the various influences that drive this phenomenon. With the right kind of dynamic model, you could isolate all the observable variables and therefore be able to extract the parameter associated with natural talent. In essence you'd be able to come up with an observation of something that is inherently unobservable.

Originally this whole thing was a fascinating modeling problem and I was quite intrigued with the idea of measuring this unobserved talent parameter. After attending the conference in Calgary, I learned more about the Relative Age Effect and all

the different ways it was manifesting itself in society, I started becoming very passionate about the issue. While the most pronounced evidence showed up in competitive sports, the part that most influenced me was the data showing the link between suicides of teenagers and their month of birth, and the data that showed that children born later in the year were more likely to be tested and tagged with a learning disability. It made no sense that more disabled children were both in December than January and even if that was an error that was detected later on by the school system, it didn't seem likely that any child that was tagged as learning disabled in his first year of school would ever fully recover from that stigma.

The final paper in this thesis is entitled "Dual Organizational Structures in Franchising". In essence it is an extension of my master's thesis which had been an opportunity for me to incorporate some of my own personal experience into my academic studies. Prior to grad school, I had spent most of my adult life working in various capacities for both the McDonald's and The Keg franchise chains.

The Keg was founded by George Tidball, who also first brought McDonald's to Canada. He modeled his new franchise after the McDonald's system, incorporating many of its operational policies, including the fundamental parts of the franchise contract. So I drew on this experience combined with my original interest in principal agent problems to write my master's thesis.

While I was attending a seminar given by Nancy Gallini, I became interested in a new aspect of the topic. Nancy gave a talk on dual structures in franchising where she was looking at the question of why franchising companies such as McDonald's would have both corporate stores and franchises simultaneously. Her hypothesis was

that franchise companies needed to own a certain number of corporate stores in order to gain credibility with potential franchisees in order to expand.

At the reception after the presentation. I posed the question to her about McDonald's specifically, saying that I found it surprising that such a large established company still needed to demonstrate credibility with potential franchisees. The evidence clearly suggested that most people that entered into franchise contracts were severely wealth-constrained. Nancy conceded that there were limitations to her hypothesis, and it was best suited for those types of franchises that were less established. It was during that conversation that I formulated the idea that the dual structure could be the result of non-convexities in monitoring costs. It seemed that the questions raised by Nancy Gallini could be addressed by re-working the franchising model explicitly within a spatial framework.

Each of these papers are linked because in its own way, they are an illustration of asymmetric information problems. On a personal level, these papers are a reflection of my eventual understanding and grasp of what my first economics instructor tried to teach me – how economics can be applied to everyday life.

Part I

The Essays

CHAPTER 1

THE DOGS OF WAR

1.1 Introduction

The phrase "*Turning loose the Dogs of War*" is sometimes used when referring to the decision by individuals or organizations to use their lawyers in a strategic manner[1]. The objective is to force an adversary to incur higher legal costs in response to the initial action. Faced with increased legal costs, respondents may choose to either enter into a favourable settlement or otherwise alter their original strategy. A recent example of this type of behavior is the *Insurance Corporation of British Columbia* (*ICBC*), the provincial government's automobile insurance company. *ICBC* instituted a policy paying a set amount for all soft tissue injury cases and litigating any claims that differed from the set amount. This policy prohibited the organization's claim adjustors from negotiating settlements with claimants, thus dramatically increasing the cost of recovering damages from low-impact car accidents[2]. Further, since ICBC typically uses staff (in-house) lawyers, most of ICBC's legal costs are pre-

[1] The title of this essay was inspired from events that arose between a property developer and a municipality in British Columbia, Canada. After a long negotiation between the two parties over perceived damages from actions of the municipality, the vice president of the development company had had enough and gave the order to turn loose the "Dogs of War".

When questioned on this about this statement, the vice president explained that he was referring to the team of lawyers his company kept on retainer. From his perspective these assets were already paid for, so he may as well use them as intended. A short time later the municipality returned with a "reasonable" settlement offer.

[2] ICBC sets a fixed amount for all soft tissue injuries ($6500 in 2005). Claimants who disputed this amount would required to seek a remedy in a court. (Source: ICBC's official website: www.icbc.com)

committed−or sunk−thus creating a credible threat for potential claimants. If such strategic uses of legal services and the associated cost asymmetries lead to a downward bias in settlements for accident claims, it brings into question the effectiveness of torts in ensuring an economically efficient level of due care by potential injurers.

A great body of literature has been devoted to analyzing the tort system to promote economic efficiency in cases of liability[3]. The fundamental issue is whether the type of damage rules employed by the courts produce a socially efficient level of care by *injurers*[4]. When the type of accidents are *unilateral,* by which is meant that only the actions of the injurers, and not the victims are assumed to influence the probability or severity of the loss, the only relevant damage rules are *strict liability* or *negligence*. When the accidents are *bilateral* in nature, the types of damage rules become much more complex[5]. This paper confines itself to types of accidents which are unilateral in nature.

A number of important articles analyze the economic effects and incentives of liability rules[6]. Shavell[7] presents one of the more noted formal treatments of negligence versus strict liability. In his model he demonstrates that, in the case of unilateral

[3]See Posner, R. **Economic Analysis of Law** Wolters Kluwer Law & Business; 7th edition (February 7, 2007) chapter 4.

[4]Strictly speaking, we are discussing *potential injurers,* since the accident may not actually happen. However, we will refer to potential injurers and potential victims as *injurers* and *victims* respectively.

[5]In fact, there are 6 possible damage rules for bilateral accident cases. For a thorough discussion of each, see John Prather Brown, Toward an Economic Theory of Liability, 323 Journal of Legal Studies (1973).

[6]See Guido Calabresi, The Costs of Accidents: A Legal and Economic Analysis (1970); Guido Calabresi, and Jon Hirschoff , Toward a Test for Strict Liability in Torts, 81 Yale Law Journal 1055 (1972); R. H. Coase The Problem of Social Cost, 3 Journal of Law and Economics 1 (1960); Harold Demsetz, When Does the Rule of Liability Matter?, 1 Journal of Legal Studies 13 (1972); Richard Posner, A Theory of Negligence, 1 Journal of Legal Studies 29 (1972); John Prather Brown, *Supra* note 5

[7]Steven Shavell, Strict Liability Versus Negligence 1 Journal of Legal Studies (1980).

accidents, both strict liability and an appropriately set negligence rule will produce socially efficient levels of care by injurers. In Shavell's model all parties are assumed risk neutral and do not engage in strategic behavior with respect to either the level of care or frequency of the activity that causes the tort.

However, the efficiency results of Shavell and others ignore the costs of using the legal system. In an extension of his own work, Shavell[8] demonstrates that once legal and court costs are considered, social and private incentives diverge. In Shavell's model social and private incentives diverge for two reasons. First, because plaintiffs are not responsible for the defendant's legal fees, plaintiffs do not consider those costs in their decision. Second, plaintiffs do not take into account the safety incentives created by the possibility of lawsuits. Shavell shows that socially inefficient suits may be brought while socially beneficial suits will not be brought.

Menell[9] challenges the second of Shavell's results. Menell purports to show that under strict liability the injurer's cost benefit equals societies cost benefit. Kaplow[10] demonstrates that Menell's result is correct, but that it fails to address the cost externality in the plaintiff's decision to bring suit. Using the Menell model, Kaplow demonstrates that under certain circumstances a prohibition on law suits may be socially desirable.

[8]Shavell, Steven, "The Social versus Private Incentive to Bring Suit in a Costly Legal System", 11 *Journal of Legal Studies* 333 (1982)

[9]Menell, Peter S., "A Note on Private versus Social Incentives to Bring Suit ina Costly Legal System", *Journal of Legal Studies* vol 12, No. 1 (Jan 1983) 41-52

[10]Kaplow, Louis, "Private versus Social Costs in Bringing Suit", 15 *Journal of Legal Studies*, (1986)

Rose-Ackerman and Geistfeld[11] demonstrates that both Menell and Kaplow fail to emphasize the essential difference between their models and Shavell. Menell believed that his result was driven by the ability of the injurer to choose any level of damages. Rose-Ackerman and Geistfeld show that this is incorrect. Instead, the distinctive feature of Menell's result is driven by the nature of the damage function. With Menell, damages are a function of the output level chosen by the injurer and occur with certainty. In Shavell's model damages are probabilistic accidents. When the problem is formulated with greater generality, Shavell and Menell become special cases. Further, with a switch to the British rule[12] , the Menell-Kaplow results hold in general.

Finally, Rose-Ackerman and Geistfeld argue that a move from strict liability to negligence will generate the optimal outcome so long as the standard of care is the same as would be optimal in the absence of lawsuits, However this conclusion requires either the British system or a situation where the plaintiff's court costs are below damages at the optimal level of care.

In the above models of tort the courts are assumed to operate at zero cost with complete knowledge of all benefits and costs. Legal fees in these models are sunk costs incurred by the relevant parties to the tort. The only uncertainty in their models is the probability of an accident (or, in some models, the size of the damage), which is a function of the care taken by one or both of the parties involved.

[11] Rose-Ackerman, Susan, and Geistfeld, Mark, "The Divergence Between Social and Private Incentives to Sue: A Comment on Shavell, Menell, and Kaplow", 16 *Journal of Legal Studies* 483 (1987)

[12] In the British system the loser pays the winner's court costs.

After an accident has occurred it is assumed that the courts will assign damages with perfect certainty, as a function of the damage rule that applies. The role of the courts in these models is the enforcement of the appropriate rule.

In fact, outcomes of court cases are not known with certainty. Instead, court cases tend to be probabilistic in nature. The probability of winning or losing tends to be a function of: (a) the particular circumstance; (b) the court's interpretation of precedence; and, (c) the efforts of the (disputing parties) legal representatives. It is the third component mentioned that allows for strategic behavior involving the amount of resources devoted to litigation and the pre-trial bargaining process.

In almost all cases, the frame of reference used by a judge is the *Learned Hand Rule*[13]. To the economist, applying the Learned Hand rule simply means asking the following question: ″Given the level of care currently taken by the injurer, would the marginal cost of an additional unit of care exceed the marginal benefit of that unit of care?". If the answer is no, then a tort has occurred. Therefore, the role of the lawyers is to convince the judge as to the location of the marginal cost and benefit curves that apply in their particular case. Beyond that, they try to negotiate the best deal for their client.

This paper presents a two stage model of negligence and legal action. In the first stage the defendant decides on the level of due care to take while carrying out an action or activity. In the second stage an accident has occurred and the plaintiff brings suit. Lawyers (agents) invest in actions that are designed to increase the probability

[13]Posner, R. *Supra Note 3*

of success on behalf of their respective clients[14]. First we present the symmetric case where all decisions on investment in legal services occur *ex-post* (after an accident has occured). We then look as the case where the potential injurer pre-commits to a level of legal services *ex-ante*. (prior to an accident) and analyze the effect that pre-commitment in legal services has on the stage one choice of care. We then present a case study from the field of labour relations where strategic pre-commitment is used to influence the settlement of grievances. The grievance process in labour relations parallel civil litigation and a grievance can be viewed as equivalent to either a tort or a breach of contract; depending on the type of grievance.

1.2 The model

This model is a two-stage game that involves an injurer and a victim. In the first stage the injurer decides on both the frequency and the level of care taken in an activity or production process. The level of care will be chosen to maximize the net private benefit of the injurer. The injurer will be referred to as the defendant and the victim will be referred to as the plaintiff. It is assumed that the victim does not influence the probability or magnitude of the loss.

In the second stage a suit is brought by the plaintiff against the defendant. Both parties invest in legal services which are assumed to increase each parties likelihood of success in court. Each party takes the level of legal services purchased by the other

[14]The exact nature of the role of lawyers in economic models of the legal system tends to be diverse. Some view legal fees as simply an additional cost associated with using the legal system. Others view the lawyer as an agent for information on the true probability of conviction in the event of a court case. In both cases, a lawyer's services enter the model only as a cost. The Lawyer plays no role in determining the outcome once a suit reaches court. The model presented here treats lawyers as an input that does influence the outcome of a court case.

party as given, and chooses their level of legal service to maximize their expected utility. It is assumed that all parties are risk neutral. The stage two equilibrium is a *Nash Equilibrium.*

If we assume all players have foresight, the outcome of the second stage game will, in turn, determine the equilibrium level of care taken in stage one. Therefore, to solve the model, we start by finding the equilibrium in the second stage game. The solution to stage two is used in the solution to the first stage of the game.

1.2.1 Initial Conditions

The defendant engages in an activity, denoted by y. The gross benefit to the defendant of activity y is

$$B(y) \qquad where \qquad B'(y) > 0, B''(y) \leq 0 \quad and \quad B(0) = 0 \tag{1.1}$$

Let x denote the level of care taken by the defendant while engaging in activity y, and $c(x)$ be the cost to the defendant of taking care of level x. Assume that $c'(x) > 0$ and $c''(x) \geq 0$.

Now suppose that the probability of loss (accident) is a function of both the frequency of the activity and the level of care. Let the probability of loss be

$$\pi = \pi(x, y) \qquad where \quad \pi_x < 0 \quad and \quad \pi_y > 0 \tag{1.2}$$

π is assumed to be a separable and additive function of all its arguments.

Let the loss incurred by the plaintiff (victim) be denoted by L. The loss incurred by the plaintiff may, or may not, be a function of the level of care taken by the defendant, i.e.

$$either \quad L = \bar{L} \qquad or \qquad L = L(x) \quad where \quad L'(x) \leq 0 \tag{1.3}$$

The expected loss faced by the victim is given by πL. Given that the signs of both π_x and $L'(x)$ are negative and any change in x will change the expected loss in the same direction,then, with no loss in generality, we shall assume that L is exogenous.

In this model all benefits and costs are accrue only to the injurer or victim; therefore, since all parties are risk neutral, social welfare can be expressed as the benefit of the activity net of the cost of care and any losses that are incurred due to an accident. The social welfare function is given by

$$W(x, y) = B(y) - c(x) - \pi(x, y)\bar{L} \tag{1.4}$$

If the injurer took into account the full cost of his actions or, equivalently, a social planner determined the frequency and level of care of the activity, the levels of x and y would be chosen to maximize equation 1.4. The first order conditions for this problem are given by

$$\begin{aligned} B'(y) - \pi_y(x, y)\bar{L} &= 0 \\ c'(x) + \pi_x(x, y)\bar{L} &= 0 \end{aligned} \tag{1.5}$$

where equation system 1.5 implicitly defines the socially optimal level of care (x^L)

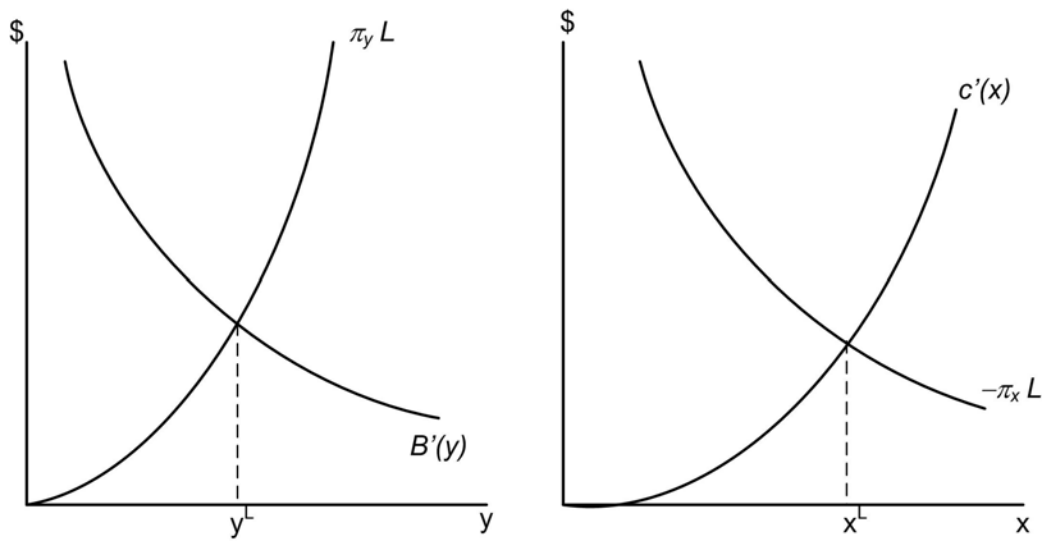


Figure 1.1: Socially Optimal Frequency and Level of Care

and frequency (y^L) of the activity. Figure 1.1 illustrates graphically the solution to (1.5). The first graph shows the optimal frequency of the injurer's activity and the second graph shows the optimal level of care.

1.2.2 Stage Two Game

In stage two an accident has happened. Each party retains legal services and the process of litigation begins. Let lawyer one represent the defendant and let lawyer two represent the plaintiff. The activity of lawyer one is denoted a_1, and the activity of lawyer two is denoted a_2. The activities of the lawyers are assumed to influence the probability of conviction if the suit goes to court. Both lawyers know the nature of probability function and the level of activity by the other lawyer.

Let the probability of conviction be given by

$$P = P(a_1, a_2; x) \tag{1.6}$$

where P is a function both parties legal activities and the level of care taken by the defendant. Since x is determined by the defendant in stage one, it is treated as exogenous in stage two. P is assumed to be continuous and twice differentiable and has the following properties:

$$P_1 < 0, \quad P_2 > 0, \quad P_{11} > 0, \quad P_{22} < 0, \qquad P_x \leq 0$$

The restrictions on P imply that, for a given level of care taken prior to the court case, each lawyer can influence the outcome of the case through his own efforts; however, there are diminishing returns to either lawyer's efforts. An additional assumption that $P_{12} = P_{21} \leq 0$ simply implies that lawyers are strategic substitutes. The sign of P_x reflects the *Learned Hand* rule; that the greater the level of care taken, the more likely the courts would consider it reasonable.

As before, let L denote the loss incurred by the plaintiff and D denote the damages awarded by the courts. Note that L may, or may not, equal D. Since pre-trial bargaining is permitted in the stage two game, let S denote any out of court settlement.

Each party must make a compensation payment to their respective lawyers. The compensation functions for lawyers one and two respectively are $w_1(a_1)$ and $w_2(a_2)$.

Compensation can be a simple linear function ($w_i a_i$) such as an hourly rate, or it can be in the nonlinear, two-part pricing schedule of the form $w_i(a_i) = \alpha f + (1-\alpha)\, w_1 a_i$, where f represents pre-payment (or retainer) and α is the fraction of the lawyer's total compensation that is a sunk cost to the client. The compensation function may also be a share of the damages awarded or settlements negotiated. Such "contingency fees" tend to vary across jurisdictions and are often subject to state or provincial regulations. In this section we will assume that all compensation functions are linear.

The defendant's expected payoff function is

$$v = -[P(a_1, a_2; x)D + w_1 a_1] \tag{1.7}$$

And the plaintiff's expected payoff function is

$$u = P(a_1, a_2; x)D - L - w_2 a_2 \tag{1.8}$$

Each party will make investments in legal services (a_i) that will maximize (minimize) his expected gain (loss)[15], taking the level of legal activity of his adversary as given. Therefore the equilibrium levels of legal activity will be a *Nash equilibrium.*

Differentiating the pay-off function of the defendant gives us

$$\frac{dv}{da_1} = -\frac{\partial P(a_1,a_2)}{\partial a_1} D - w_1 = 0 \tag{1.9}$$

[15]It is assumed that there exists no agency problem on the part of lawyers, such that lawyers will not in engage in excessive legal activities to maximize their own reward.

or

$$-\frac{\partial P(a_1,a_2)}{\partial a_1}D = w_1 > 0$$

and differentiating the pay-off function of the plaintiff gives us

$$\frac{du}{da_2} = \frac{\partial P(a_1,a_2)}{\partial a_2}D - w_2 = 0 \tag{1.10}$$

or

$$\frac{\partial P(a_1,a_2)}{\partial a_2}D = w_2 > 0$$

Equations 1.9 and 1.10 implicitly define the defendant and plaintiffs' respective *best response functions*

$$a_1 = R^1(a_2) \qquad and \qquad a_2 = R^2(a_1) \tag{1.11}$$

where $a_i = R^i(a_j)$ describes the optimal amount of legal services for player i in response to a given level of legal services of player j. By taking the total differential of 1.9 and 1.10, the following can be shown to hold:

$$dR^1/da_2 > 0 \qquad and \qquad dR^2/da_1 < 0$$

The best response functions for the plaintiff and defendant are illustrated in Figure 1.2. It is of interest to note the asymmetry of the two response functions. The defendant's best response to an increase in a_2 is to increase a_1, whereas the plaintiff's

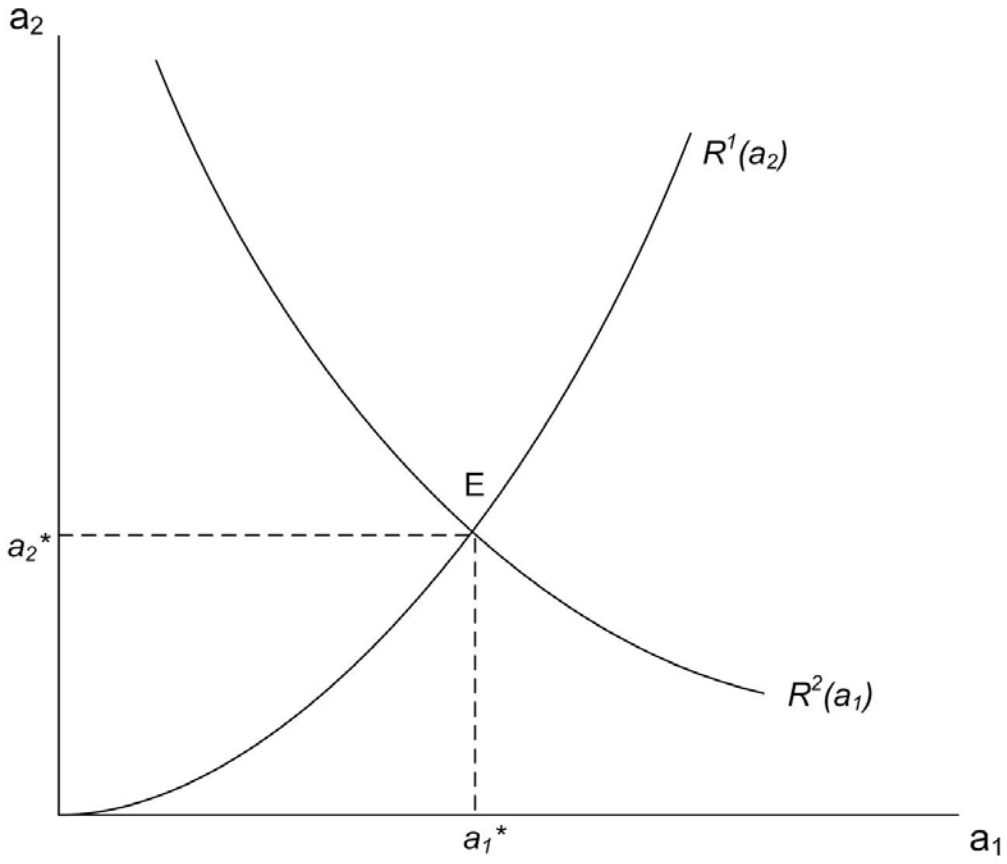


Figure 1.2: Best response functions in legal services

best response to an increase in a_1 is to reduce the level of a_2[16]. As shown in figure 1.2., the intersection of the best response functions determines the *Nash equilibrium* values, $a_1^*(x)$, $a_2^*(x)$.

Once $a_1^*$and a_2^* are determined, the probability of conviction is also known by both parties, given

$$P^* = P(a_1^*(x), a_2^*(x), x) = P^*(x) \tag{1.12}$$

In this framework it is assumed that both parties are risk neutral. Therefore the defendant would be willing to offer a settlement, S, such that

$$S \leq P^*(x) \times D + w_1 a_1 \tag{1.13}$$

and the plaintiff would accept any settlement of S that satisfied

$$S \geq P^*(x) \times D - w_2 a_2 \tag{1.14}$$

Assuming a *Nash* bargaining outcome, we get the value of S such that each side would be indifferent between going to trial or accepting an out of court settlement. In this case the equilibrium value of the settlement would be

$$S^*(x) = P^*(x)D + \frac{w_1 a_1 - w_2 a_2}{2} \tag{1.15}$$

[16]In other words, from the perspective of the defendant legal activity can be viewed as *strategic compliments*, whereas the plaintiff views legal activies as *strategic substitutes.*

In the case where the courts set $D = \bar{L}$, equation 1.15 can be re-written as

$$S^*(x) = P^*(x)\bar{L} + \frac{w_1 a_1 - w_2 a_2}{2} \tag{1.16}$$

1.2.3 The Stage One Game

Given S^* from above, the stage one objective function of the defendant can be written as

$$V(x, y) = B(y) - c(x) - \pi(x, y)S^*(x) \tag{1.17}$$

Differentiating 1.17 with respect to y and x gives us

$$\begin{aligned} B'(y) - \pi_y(x, y)S^* &= 0 \\ -\left[\pi_x(x, y)S^* + \pi\tfrac{dS^*}{dx}\right] &= c'(x) \end{aligned} \tag{1.18}$$

equation system 1.18 determines the injurer's privately optimal x and y, which are denoted x^s and y^s. Comparing the first equation in (1.18) to its counterpart in (1.5), we have the result that $y^S > y^L$ whenever $S^* < \bar{L}$. This implies that the frequency of the activity will be greater than is socially optimal. With respect to the level of care, x, the result is ambiguous. The usual assumption is that, since $S^* < \bar{L}$, the level of care would be less than the socially optimal level. However, from the second equation in (1.18) it is not to possible to determine the level of care relative to the social optimum (x^L) . Since $\frac{dS^*}{dx} \leq 0$ and $\frac{da_1}{dx} \leq 0$, then, depending on their magnitudes, the reduction in both the settlement and legal fees in the second stage

due to greater care may be strong enough that x^S would exceed the socially optimal level of care. On the other hand, if ex-ante changes in x have little impact on the stage two equilibrium, then the impact of $\frac{dS^*}{dx}$ would be small relative to the difference between $S^*$and $\bar{L}$ and the level of care would be less than the social optimum.

Figure 1.3 illustrates the results. The first graph shows the socially optimal and privately optimal frequency of the activity. When the true level of damages, or loss, are taken into account by the injurer, he will set the marginal benefit of the activity ($B'(y)$) equal to the true marginal expected damage function ($\pi_y L$) and the equilibrium will occur at point E. When the injurer equates the marginal benefit of the frequency of the activity to his personal marginal expected damages ($\pi_y S^*$) equilibrium will occur at point F. The second graph in figure 1.3 illustrates both the socially optimal and private choice of care. When the injurer equates the marginal reduction in expected damages ($\pi_x L$) to the marginal cost of care ($c'(x)$), the socially optimal level of care occurs at point K. However, when the injurer equates the marginal reduction in expected settlement costs $\left(-\pi_x S^* - \pi\frac{dS^*}{dx}\right)$ to the marginal cost of care, equilibrium occurs at point J (here we assume dS^*/dx to be small).

1.3 Strategic investment

It is often the case that large firms pre-invest in legal services. This usually involves an annual retainer which the firm views as a sunk cost and effectively gives the firm "free" legal services at the margin, up to the point that the retainer is exhausted. This creates a discontinuity in the firm's cost of legal services, whether it is the plaintiff or defendant.

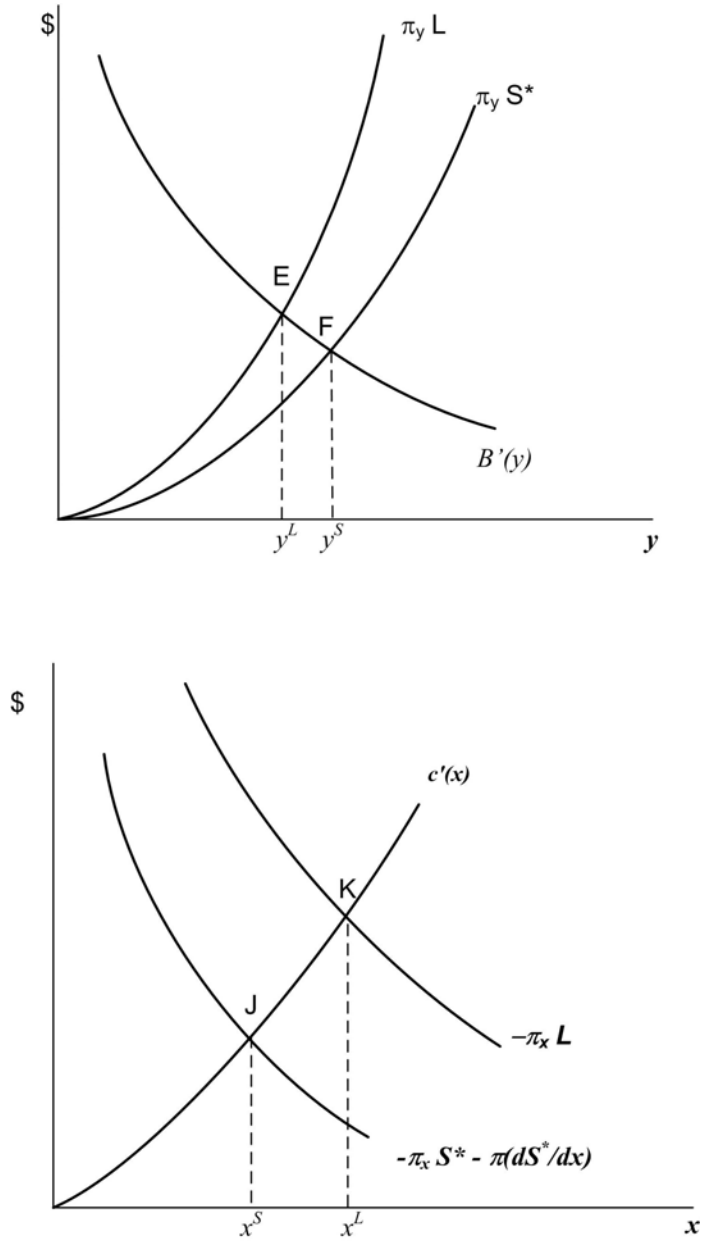


Figure 1.3: Private and socially optimal levels of the injurers activity (y) and care (x)

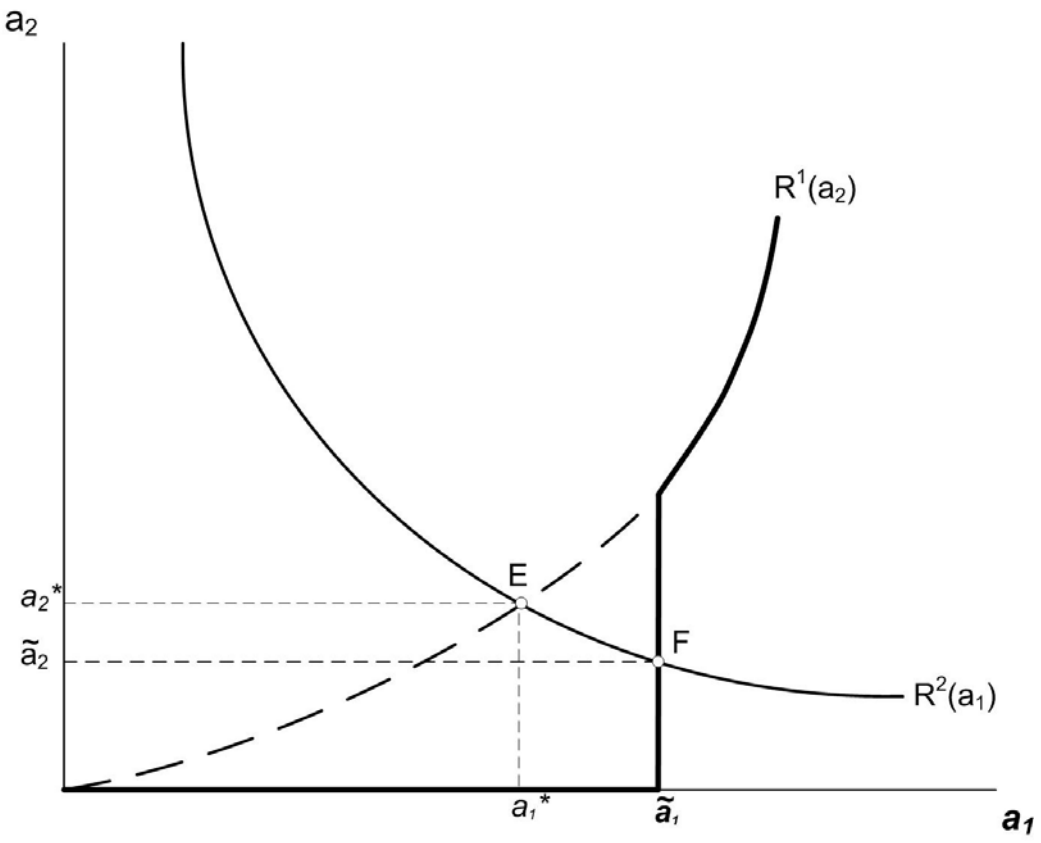


Figure 1.4: Best response with pre-commitment to legal services

1.3.1 Stage two equilibrium with sunk investment in legal services

Consider the case where the defendant (injurer) has made an ex-ante sunk investment in legal services. Let $\tilde{a}_1$ be the amount of legal services that the defendant has retained. therefore equation 1.9 becomes

$$\frac{dv}{da_1} = \begin{cases} -\frac{\partial P(a_1,a_2)}{\partial a_1} D = 0 & for \quad a_1 \leq \tilde{a}_1 \\ \\ -\frac{\partial P(a_1,a_2)}{\partial a_1} D - w_1 = 0 & for \quad a_1 > \tilde{a}_1 \end{cases} \tag{1.19}$$

This result is illustrated in figure 1.4.

In figure 1.4 the original equilibrium is point E. This represents the equilibrium illustrated in figure 1.2 above. When the defendant makes ex-ante investment in a sunk level of legal services, his response function becomes vertical at $\tilde{a}_1$ up to the point it crosses the original best response function of the defendant. At that point the response function returns to the original response function as the defendant must now incur additional legal services at the rate of w_1. Assuming the investment is large enough, the plaintiff's response function will intersect the defendant's at point such as F and the values a_1 and a_2 are denoted $\tilde{a}_1$ and $\tilde{a}_2$ respectively. In this case, the defendant has effectively changed the equilibrium point in his favour. An equilibrium at point F corresponds to a lower probability of success for the plaintiff than an equilibrium such as point E.

The probability of conviction becomes $\tilde{P} = P(\tilde{a}_1(x), \tilde{a}_2(x))$ where

$$P(\tilde{a}_1(x), \tilde{a}_2(x)) < P(a_1^*(x), a_2^*(x))$$

and the Nash equilibrium settlement offer now becomes

$$\tilde{S}(x) = \tilde{P}(x)D - \frac{w_2}{2} \tag{1.20}$$

where $\tilde{S}(x) < S^*(x)$. This occurs for two reasons: First, since $w_1\tilde{a}_1$ is sunk in the first stage, the pre-trial settlement offer by the defendant is less than equation 1.15 by the amount $w_1/2$. Second, as shown in figure 1.4, $\tilde{a}_1 > a_1^*$ and $\tilde{a}_2 < a_2^*$, thus the equilibrium probability of conviction is now lower than in the case where both parties

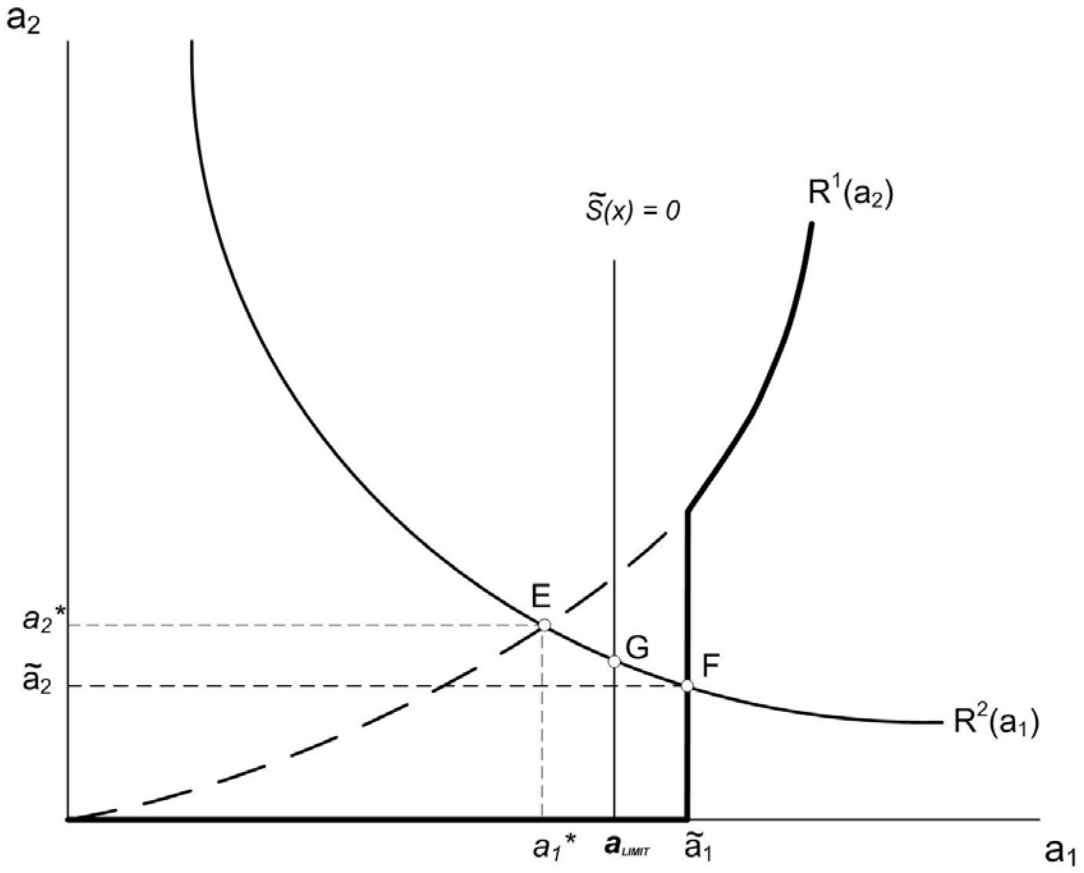


Figure 1.5: Pre-commitment to legal services and suit deterring investment

invest in legal services ex-post ($\tilde{P}(x) < P^*(x)$)

1.3.2 Stage one decision on pre-commitment to legal services

There are two possible outcomes in the stage one game depending on the size of the plaintiff's legal costs relative to damages. The first being that the defendant will choose a level of legal services to pre-commit up to the point that the marginal reduction in expected settlement just equal to the marginal cost of legal services ($w_1'(a_1)$). However, it may also be the case that there is a level of pre-commitment what will effectively deter lawsuits. This would be, in essence, a *Limit Output* or

Limit Pricing strategy.

In stage one, the defendant's first-order conditions become

$$\begin{aligned} B'(y) - \pi_y \tilde{S} &= 0 \\ -\pi_x \tilde{S} - \pi \frac{d\tilde{S}}{dx} - c'(x) &= 0 \\ -\pi \left[\underset{(-)}{\frac{\partial S}{\partial a_1}} + \underset{(+)}{\frac{\partial S}{\partial a_2}} \underset{(-)}{\frac{da_2}{da_1}} \right] - w_1 &= 0 \end{aligned} \tag{1.21}$$

Let $\tilde{x}$ and $\tilde{y}$ denote the level of care and frequency of the activity, respectively, that satisfies (1.21). Because there is a pre-commitment to legal services, a_1 becomes a choice variable in the stage one game. Let the amount of legal services that satisfies (1.21) be denoted as $\tilde{a}_1$. Since the solution to (1.21) in stage one will determine the optimal settlement in stage two, there are two possible outcomes. First, if, at $a_1 = \tilde{a}_1$, $\tilde{S}(x) > 0$ then $\tilde{S}(x)$ will be the Nash equilibrium settlement in the stage two game. Otherwise, if $\tilde{S}(x) \leq 0$, then a_1 will be chosen such that $\tilde{S}(x) = 0$. Let a_{Limit} denote the level of legal services drives (1.20) to zero. In this case, the pre-commitment to legal services will effectively deter lawsuits. This result is illustrated in figure 1.5.

Consider the case when $\tilde{S}(x) > 0$. When $\tilde{S}(x) < S^*(x) < L$, this implies that $\tilde{y} > y^S > y^L$ and that $\tilde{x} < x^S$. Further, like the results in (1.18), it is ambiguous as to whether $\tilde{x}$ is less than x^L. Figure 1.6 illustrates the stage one equilibrium frequency and care associated with the activity when there is a pre-commitment to legal services. The bottom graph illustrates the case where the level of care under

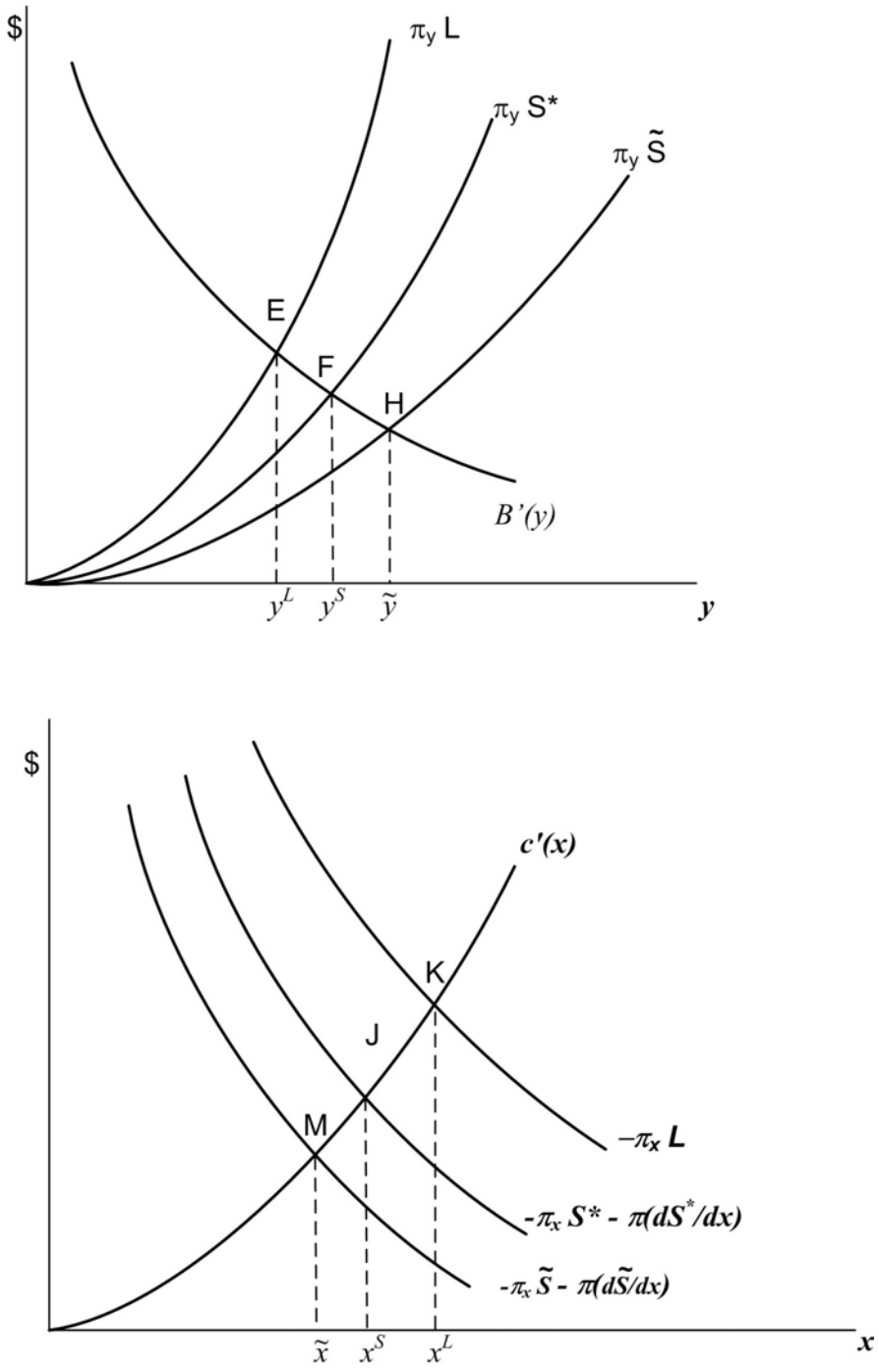


Figure 1.6: Privately optimal care and frequency when injurer pre-commits to legal services.

settlement is less than the social optimum (x^L).

1.4 Strategic legal investment in labour relations

1.4.1 A case study in the field of post-secondary education

Strategic investment in legal services also occurs in the field of labour relations. Grievances can be viewed either as a tort or a contract breach, depending on the nature of the grievance. Grievances arise in one of two ways: either via a complaint by an individual (union member), or from one of the signatories to the collective agreement (union or employer). The latter is usually a procedural or jurisdictional grievance and is often the more common form of grievance.

Grievances occur in a manner similar to the process of a civil litigation. A grievance is filed and pre-arbitration negotiations occur - ultimately leading to either a settlement or arbitration hearing[17]. Once a case is referred to arbitration, an arbitrator is agreed upon and the cost is divided between the two parties. In the province of British Columbia, an arbitration hearing can cost between $5,000 and $10,000 per day, and the typical hearing can last from three to ten days. While small organizations (unions or employers) will engage outside council to represent their interests at the arbitration, larger organizations will have in-house representation[18].

In the 1990s, there were a high number of jurisdictional grievances occurring at the British Columbia Institute of Technology (BCIT). In 1981, the provincial

[17] An arbitration hearing operates much like a court. An independent arbitrator acts as judge and legal council presents evidence and cross-examines witnesses. The ruling of the arbitrator is usually final.

[18] Individuals with legal training are highly sought after in labour relations, since they can fulfill a dual role as grievance officer and legal advocate.

government merged BCIT with the Pacific Vocational Institute (PVI), resulting in a single institute containing three different bargaining units (unions) representing the various employee groups on campus. Two of the unions, the Faculty and Staff Association (FSA) and Government Employees' union (BCGEU), represented faculty, with jurisdiction being determined by course (or program) content; a metric that was an ongoing source of debate and controversy. By the late 1990's, BCIT had more outstanding grievances than the rest of the province's post-secondary system combined[19].

The complainant responsible for most of these grievances was the FSA, the largest of the three bargaining units[20]. The FSA adopted a policy of pre-committing $100,000 per year to legal services (approximately 15% of their annual revenues) and, in 2000, the FSA decided to post their operating budget and all other financial documents on their website for public viewing. This strategy effectively signaled their commitment of resources to the grievance process. In 2002, the FSA implemented a 90-day "fast-track" policy for new grievances. 90 days after a new grievance file was opened, an arbitration hearing was booked (regardless of the state of negations). Outstanding grievances dropped from a high of 68 cases in 1999, to less than 10 by 2004.

[19] See the *British Columbia Labour Relations Board* reports found at
http://www.lrb.bc.ca/reports/

[20] Information summarized in this section was compiled from the minutes of the BCIT Faculty and Staff Association board of director's minutes published on the FSA website:
http://www.bcitfsa.ca/minutes.htm

1.5 Conclusions

In the economic analysis of the legal system, the role of lawyers has tended to fall into two categories. The first is that of information broker. In models of asymmetric information, the lawyer supplies information to the client about the probability of a particular outcome in the event of a trial. The additional information is then incorporated into decisions regarding the size of the settlement or to proceed to trial.

In the second, the lawyer is simply part of the overall cost of using the court system. In either case, the probability of success at trial is usually exogenous. Further, the cost of using the legal system is treated as an exogenous variable that determines a boundary condition within a benefit-cost analysis governing the decision to proceed with a suit.

The main premise of this paper is that the actions of lawyers do influence the outcome of a trial; that a greater investment in legal services will increase the probability of a favourable outcome to the party making the investment. Under this assumption expenditures on legal services becomes a strategic choice variable; both with respect to the amount and to whether it is an ex-ante or ex-post expenditure.

This produces some interesting results. First, with respect to the frequency of the activity that could cause harm, the model's finding match those of Shavell and others. The presence of a costly legal system causes the private and social optimum to diverge, with the amount of the activity being too high relative to the social optimum. This result holds both in the case where legal fees are paid ex-post or there is a pre-commitment in legal services.

When analyzing the level of care, however, the results do differ from that of Shavell and others. In both cases, where legal fees are incurred expost or pre-committed, it is ambiguous as to whether the level of care will be less than the social optimum, If damages (D) in the suit are equal to the true loss (L), the expected settlement cost used by the injurer in determining the level of care will be less than the actual loss incurred by the victim. Intuitively, this would suggest that care would be set lower than the optimal. However, if the choice of care leads to a reduction in both the size of the settlement and the injurer's legal costs, there would be an additional incentive for the injurer to take additional care. Therefore it is possible that the level of care may not be less than that which would be socially optimal. Depending on the productivity and cost structure of legal services, it is possible for the level of care in this case to be greater than the social optimum.

This result would be sensitive to how the courts apply the *Learned Hand* rule or to type of liability rule adopted by the courts. If, in the case of multiple defendants, the courts favoured a *relative liablity* approach[21], then we would expect the dS/dx term in equations 1.18 and 1.21 to be larger. Additionally, if the defendant viewed the courts as having a propensity to favour the plaintiffs, one would also expect additional care to be taken. More would need to be said about the properties of the $P(x)$ function than has been developed in this paper.

Finally, it is unambiguous that strategic pre-commitment will lead to lower set-

[21] In product liability cases or accidents (i.e. motor vehicle) it is not uncommon for the plaintiff to name any and all related parties in the suit (manufacturer, distributor, financial backers, etc.). In the case of multiple defendants the courts will determine how to allocate shares of the damages across defendants.

tlements and, in some cases, effectively deter lawsuits in the event of accidents. One testable implication of this strategy is the ever rising allocation of resources to legal services on the part of firms and organizations. This also implies a possible welfare gain to the use of punitive damages. There is also the possibility that the adoption of the *British system* of assigning court costs, or some other restriction on the manner by which lawyers are compensated, could lead to a welfare gain. Recent experiments in alternative dispute resolution by certain jurisdictions may imply that this is the case.

1.6 References

- Bebchuk, L. "Litigation and Settlement under Imperfect Information", RAND Journal of Economics, Vol. 15, No.2 (1984), pp. 404-415.

- Calabresi, G. "The Costs of Accidents: A Legal and Economic Analysis" (1970)

- Calabresi, G., and Hirschoff, J., "Toward a Test for Strict Liability in Torts", 81 Yale Law Journal 1055 (1972);

- Coase, R. H. "The Problem of Social Cost", 3 Journal of Law and Economics 1 (1960)

- Demsetz, H., "When Does the Rule of Liability Matter?", 1 Journal of Legal Studies 13 (1972)

- Gould, J. "The Economics of Legal Conflicts" Journal of Legal Studies, Vol. 2, (1973), pp.279-300

- Kaplow, Louis, "Private versus Social Costs in Bringing Suit", 15 *Journal of Legal Studies* Vol. 15, No. 2 (Jun., 1986), pp. 371-385

- Landes, W. "An Economic Analysis of the Courts" Journal of Law and Economics, Vol. 14 (1971), pp. 61-107.

- Menell, Peter S., Menell, Peter S., "A Note on Private versus Social Incentives to Bring Suit ina Costly Legal System", *Journal of Legal Studies* vol 12, No. 1 (Jan 1983) 41-52

- Nalebuff, B. "Credible Pretrial Negotiation", RAND Journal of Economics, Vol. 18, No.2 (1987), pp. 198-210.

- Osbonne, E. "Who Should Be Worried About Asymmetric Information in Litigation", International Review of Law and Economics, Vol. 19 iss.3 (1999), pp. 399-409.

- P'ng, I. "Strategic Behavior in Suit, Settlement, and Trial" Bell Journal of Economics, Vol. 14, No.2 (1983), pp. 539-550.

- Posner, R. **Economic Analysis of Law** Wolters Kluwer Law & Business; 7th edition (February 7, 2007)

- Posner, R., "A Theory of Negligence", 1 Journal of Legal Studies 29 (1972)

- Posner, R. "An Economic Approach to Legal Procedure and Judicial Administration" Journal of Legal Studies, Vol. 2, (1973), pp.399-458.

- Prather Brown, J. Toward an Economic Theory of Liability, 323 Journal of Legal Studies (1973).

- Priest, G. and Klein, B. "The Selection of Disputes for Litigation" Journal of Legal Studies, Vol. 13, (1984), pp.1-55.

- Reingannum, J. and Wilde, L. "Settlement, Litigation, and the allocation of Litigation Costs' RAND Journal of Economics, Vol. 17, No.4 (1986), pp. 557-566.

- Rose-Ackerman, Susan, and Geistfeld, Mark, "The Divergence Between Social and Private Incentives to Sue: A Comment on Shavell, Menell, and Kaplow", 16 *Journal of Legal Studies* 483 (1987)

- Shavell, S., "Strict Liability Versus Negligence" 1 Journal of Legal Studies (1980)

- Shavell, S. "Suit, Settlement and Trial: A theoretical Analysis under Alternative Methods for the Allocation of Legal Costs" Journal of Legal Studies, Vol. 11, (1982), pp.55-81

- Shavell, S. "The Social Versus the Private Incentive to Bring Suit in a Costly Legal System" Journal of Legal Studies, Vol. 9, (1982), pp.333-339

- Shavell, S. "Any Frequency of Plaintiff Victory is Possible", Journal of Legal Studies, Vol. 25, (1996), pp. 493-501.

CHAPTER 2

THE TYRANNY OF CHRONOLOGICAL AGE

2.1 Introduction

A review of player profiles the National Hockey League (NHL) reveals that more than four times as many players are born in January as in December. Further, 70 percent of all players are born in the first six months of the calendar year. Investigations into the highest level of amateur hockey leagues reveal similar findings. The obvious question is why do players born early in the calendar year dominate these leagues? When we look back to the years that these players were born, we find that the monthly birth rate of males does not conform to this pattern. For the years that the players in question were born, the birth rate of males in Canada is almost perfectly uniform, with no evidence of seasonality or other systematic variations. Researchers in the fields of education and psychology have labeled this observed phenomenon as the *Relative Age Effect*.

Players who end up in the NHL (or in the upper levels of Canadian Hockey) are picked or selected by a non-price allocation mechanism, which makes this issue of interest to economists. The allocation mechanism is based on relative performance – children who do well tend to stay in the system and those who demonstrate an exceptional aptitude are given special treatment. Those who excel are placed on teams that receive dedicated resources. This type of selection mechanism is expected

to filter out untalented players, to encourage talented players to stay in the system and to develop their talents. Yet observed results, summed up by what has been identified as the relative age effect, strongly suggests otherwise, raising the question as to why is this apparently sensible allocation mechanism producing such birth-date biased results?

The phenomenon described above is not unique to hockey. Similar results are found in most organized sports played in many countries and cultures[1]. Further, studies of academic performance in school systems of different countries has also shown a bias similar to that found in hockey. In national exams given to elementary and high-school students the top percentiles tend to be dominated by children and youths who are born in the first half of the calendar year[2]. In contrast, grade school retention's (failures) and even teenage suicides (see figure 2.1) tend to be dominated by children born in the last half of the calendar year[3].

Recent research in Psychology has shown that selection processes in schools and some sports are characterized by systematic errors caused by the difficulty of observing ability independent of maturity in children[4]. When children are grouped into age

[1]A study of teams in the under 23 year old world cup of soccer found that there was strong evidence of the relative age effect across all teams. This event draws on teams from 24 countries.

[2]See: Freyman, R. "Further Evidence on the Effect of Date of Birth on Subsequent School Performance." *Educational Research,* 8, 58-64, (1965); Jinks, P.C., "An Investigation into the Effect of Date of Birth on Subsequent School Performance." *Educational Research,* 6, 220-225, (1964); Russell, R.J.H. and Startup, M.J. "Month of Birth and Academic Achievement." *Personality and Individual Differences,* 7, 839-846, (1986); Sutton, P. "Correlation between Streaming and Season of Birth in Secondary Schools." *British Journal of Educational Psychology,* 37, 300-304, (1967); Thompson, D. "Season of Birth and Success in the Secondary School." *Educational Research,* 14, 56-60, (1971).

[3]See: Barnsley, R.H., Allen, J., and Thompson, A.H. "School Achievement, Grade Retention and 'The Relative Age Effect'."

[4]Barnsley, R.H., Thompson, A.H. and Barnsley, P.E.,"Hockey Success and Birthdate: The Relative Age effect." *Canadian Association for Health, Physical Education, and Recreation,* 51, 23-

categories such as calendar year of birth, there may be differences in both physical and mental maturity due to the differences in age within the group. If children are grouped by calendar year, the age difference between any two children can be as much as twelve months. During the formative years, this may produce significant differences in performance due to the difference in maturity and ability of these children. When assessing the performance of children during this time, there is a risk of mistaking differences in innate ability with differences in maturity.

Furthermore, in many institutions, such as minor hockey, children are often sorted into tiers based on observed ability. Within these tiers children will receive different levels of training. Usually those demonstrating the highest ability will be sorted into a tier which will receive a greater degree of training than those who demonstrated a lesser ability. If the relative age effect is present during this sorting process, then children of greater maturity will be mistakenly selected into the highest tiers. Once selected, the differences in training will cause the selection bias to persist beyond the point in time where relative age effects are significant. This secondary effect is referred to as the *Training Effect*. It is generally recognized that the relative age effect is transitory in nature, disappearing by early teenage years. However, if it occurs concurrent with the training effect, temporary distortions observed in the data cited will become permanent.

Barnsley and Thompson (1988) propose two hypotheses to explain the observed age distribution found in professional hockey and the highest levels of amateur hockey. The first suggests that there is a tendency for older children to continue to partic-

28,(1985);

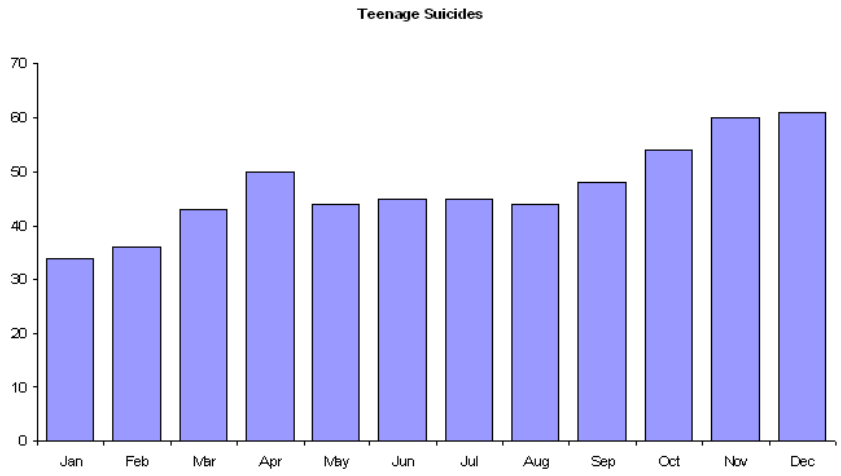


Figure 2.1: Teenage suicides by month of birth, 1979-1990 *(Source: Alberta Ministry of Health)*

ipate in hockey over time, whereas those who are younger tend to drop out. This is referred to as the *discouragement effect*. The second hypothesis suggests that the older children experience greater success due to being streamed into tiers, where the top tiers receive more intensive training. The better training causes them to continue to progress through the system in the subsequent periods and it is this top tier that supplies players to the professional leagues. The work of Anders Ericsson and others[5] strongly suggest that talent alone is never enough to explain success at the elite levels of sports. While talent may be considered a precondition, intensive training is a necessary condition for success.

Whether it is the selection of elite teams in minor hockey or allocating young students into enriched classes in primary school, an allocation of resources is taking

[5]Ericsson, K. A., Charness, N., Feltovich, P. J., Hoffman , R. R., (editors) **The Cambridge Handbook of Expertise and Expert Performance** (June 2006) Cambridge Press

place, thus making this inherently an economic problem. By allocating resources to those who have the potential to derive the most benefit, there is an implicit social gain. In the case of education, society, by streaming, hopes to produce the best doctors, researchers and teachers. If there is a limit to the number of doctors a society can produce, then the hope is to select the best candidates to fill those positions. If there exists some systematic bias the distorts the signal used in the streaming process, this creates a potential cost to society that would reduce social welfare.

The social cost of the relative age effect can be extensive. It can range from the lost opportunity of not training the best team of cancer researchers, to the associated costs of health care and crime due to those children who have been disenfranchised by society. In addition to the problem of teenage suicide, recent studies have linked problems of self-esteem and mental health to the relative age effect. The existence of the relative age effect implies that there is a potential for two types of measurement errors. First, relatively older children will be deemed to possess higher natural ability than is the case, placing pressure on these individuals to maintain a level of performance that exceeds their true ability. Second, relatively younger children will be overlooked in the process due to lack of maturity. This may cause them to become discouraged and withdraw from the process prematurely. Both of these errors lead to an inefficient use of society's resources and potential hardship for the individuals involved.

The goal of this essay is to develop a structural model that captures the various factors that influence the observed age distribution found in organized hockey.

Through the use of a structural model we hope to separate differences in innate ability from differences in maturity and training. In doing this, we can estimate the magnitude of the misallocation of players due to the relative age effect. It is hoped that this work will lead to a framework that can address the more general issues that surround all education and training processes that involve streaming.

The analysis is applied to Canadian minor hockey for two reasons. First, the CAHA is a fairly rigid system that is applied to all Canadian hockey players on a national level. The CAHA has very strict rules governing eligibility based on birthdate and geographical location. The level of hockey a Canadian youth plays is completely determined by birth date. Since the CAHA is a national body, any relocation of a hockey player will have no bearing on the player's level. Second, minor hockey tends to mirror the school system with respect to age levels, cohort groupings, entry dates and end dates. Both systems use some form of "streaming" to allocate scarce resources. Hockey has "rep-teams" for the top hockey players and schools often have enriched classes for the top students. Finally, since hockey is an elective activity, players may choose to no longer participate; which is not true in the education system; therefore hockey supplies a much cleaner set of data for analysis.

2.2 The Canadian Minor Hockey System

Historically, 80 percent of the players in the NHL are Canadian citizens and a product of Canadian minor hockey system. This system is governed by the Canadian Amateur Hockey Association (CAHA), a long-standing organization which has governed and regulated amateur hockey in Canada for several decades. In this role,

the CAHA has kept accurate and detailed records on all aspects of minor hockey, including team and individual statistics, player histories and demographics. Canadian NHL'ers typically entered minor hockey when they were between 6 and 8 years of age. Each subsequent year they would progress through a series of levels, or leagues, until they ultimately reached the highest amateur level, known as "Junior". It is from this level that players are selected for the NHL.

As we mentioned earlier, a large majority of the players in Major Junior hockey (17-19 year old's) are born in the first six months of the calendar year. The top box in figure 2.2 presents a summary of birth data collected from the 1983 rosters of the Western Hockey League. We see that 71.5% of the players are born between January and June. Further, when the players are grouped by quarters, players born in the first three months of the year account for 41.9 % of the players followed by a systematic decline in participation to only 9.2% by players born in the last three months of the year.

How does this result compare to the total population of males at that age? Since the average age of the Western Hockey League is 17 years, we look back at the approximate time frame that most players would have been born. This corresponds to the years 1966-1967. The lower box in figure 2.2 shows the number of live male births by month for the period of July 1, 1966 to June 30, 1967. We can see that the distribution of male births by month is fairly uniform, and not at all similar to the distribution found in the Western Hockey League.

Records from the CAHA show the participation rate by month of birth at the

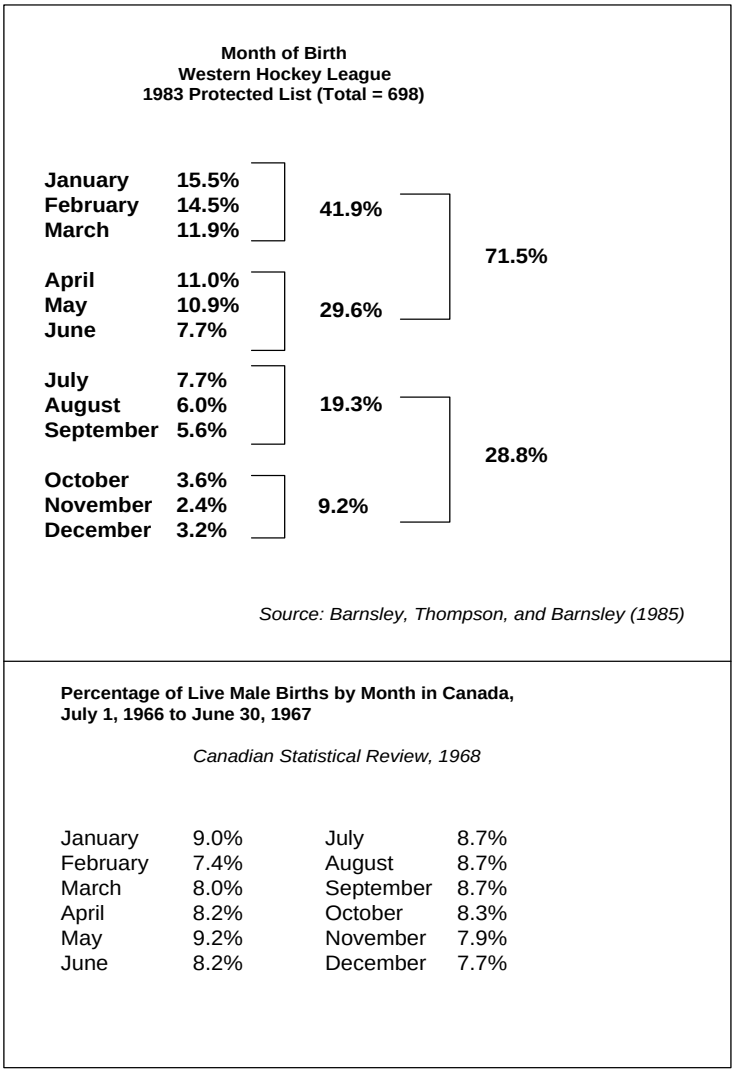


Figure 2.2: Comparison of the participation rate in the Western Hockey League (18 year olds) and the corresponding national birthrate of 18 year olds in Canada

youngest level of minor hockey very much mirrors that of the male birth rate in Canada. In other words, children grouped by month of birth tend to be uniformly represented at the entry level of minor hockey. Assuming that the distribution of natural ability is independent of the month of birth, we would expect that players at the highest level of both amateur and professional would also be uniformly distributed by month of birth. As mentioned above, this is not the case.

At this point we shall describe in more detail the process by which a player who enters minor hockey at the age of six years may progress through the system until an age of between16 and 18 years. Figure 2.3 is a summary of the number of players at each level found in a representative minor hockey system[6]. At each level players are grouped into two cohorts. Those born in the first six months of the calendar year are designated as *Old* and those born in the last six months are designated as *Young*. Also included is the percentage of the total age group each cohort represents.

At the first stage, known as "Mites", players are randomly allocated into teams in order to play recreational hockey. At this level the emphasis is on skill development and recreation rather than competition. However, there is also an evaluation process being carried out at this time which will influence the path each player will follow in subsequent years of playing minor hockey.

After playing one or two years at the Mite level, players move to the next level of organized hockey ("Pups"). At the next level there are two possible categories, or states, a player may find himself : tier one (*Rep-Team*), or tier two (*Recreational*

[6]This data set is from the Edmonton, Alberta chapter of the CAHA system.

Figure 2.3: Summary data of the minor hockey system (CAHA Edmonton, Alberta 1984). Participants born in the months January to June are grouped as "Old", participants born in the months July to December are grouped as "Young"

Level	Tier	Old	Young	total	Total all tiers	Exit
1 Mite	n/a	1132	1078	2210	2210	
8 & under		51%	49%	100%		
2 Pup	Tier 2	760	812	1572		
9-10 yr		34%	37%	71%		
	Tier 1	216	106	322	1894	316
		10%	5%	15%	86%	
3 Peewee	Tier 2	602	623	1225		
11-12 yr		32%	33%	65%		
	Tier 1	213	116	329	1554	340
		11%	6%	17%	82%	
4 Bantum	Tier 2	475	488	963		
13-14 yr		31%	31%	62%		
	Tier 1	148	73	221	1184	370
		10%	5%	14%	76%	
5 Midget	Tier 2	314	297	611		
15-16 yr		27%	25%	52%		
	Tier 1	81	44	125	736	448
		7%	4%	11%	62%	

Hockey). In addition players may also choose to leave minor hockey altogether (*Exit*). Tier one is a selective group chosen from the list of players in stage one who were evaluated to be the best hockey players in the previous level. The players who are determined to be the best are offered the opportunity to play on an elite team which offers them more ice time, better coaching, and a higher level of competition.

Players not selected for tier one, or those selected but declined the offer to play tier one, can then choose to play recreational- or tier two- hockey . This category, or state, involves less ice time and usually a lower level of competition. Tier two is open to any individual wishing to play, with the only restriction being that they must

play with others who are born in the same calendar year. Finally, some players may choose to enter state three and stop playing minor hockey.

The selection process described above repeats each year until players reach an age of 16 to 18 years. At the end of each hockey season players in both tier one and tier two are re-evaluated and a new list made of players to be offered positions in tier one. Those players not offered a position on a tier one team again have the choice of continuing in tier two or leaving the sport. The flow chart found in figure 2.4 illustrates the entire process. The transition arrows between stage two and stage three of figure 2.4 illustrate the possible paths that players may follow each year. It is the transition from stage two to stage three that describes the process by which players move through the several levels of organized minor hockey, with only difference found in the stage one to stage two transition, the initial sorting process. Note that a player in any given state during a hockey season could move to any of the three possible states the next season.

The process continues until the final period when players have reached the highest league governed by the CAHA (usually known as *"Midget"*). After this level, players cease to be governed by the CAHA. The next level available to players is the Canadian Hockey League (CHL). The CHL is made up of three leagues: the western hockey league (WHL); the Ontario hockey league (OHL); and the Quebec major junior hockey league (QMJHL). These three leagues are the highest amateur leagues in Canada. The CHL selects players from the CAHA's midget level through a combination of regional protected lists and an annual draft. The CHL is strictly an elite league

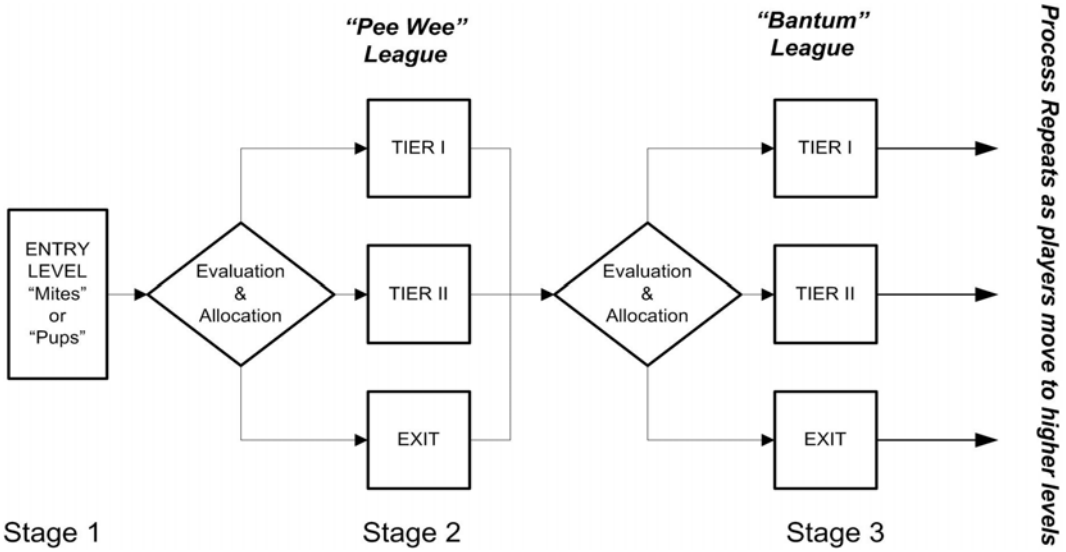


Figure 2.4: Minor hockey evaluation and progression flowchart

which produces players for the professional draft.

Players from midget hockey who are not selected by a team in the CHL may have the option of playing tier two junior hockey. Tier two junior is a provincial league that is of slightly lower calibre than the CHL. However, Tier two junior is also a competitive league that takes a limited number of players in a similar manner to the CHL. Players who fail to make either a CHL team or tier two have very few options to continue in organized, competitive hockey. Some may join a recreational men's league and some may try to play for a college or university (contingent on academic requirements). However, for most, their hockey careers end at this stage.

Comparing the participation rate by age cohort at the Mite level to the age cohorts at the Midget level in figure 2.3 demonstrates the relative age effect. Figure 2.4 shows us all the possible paths that an individual may follow to reach the final stage of the process. Obviously some paths are much more likely, while others are

highly improbable. For example, it is conceivable that an 18 year old that has never played organized hockey may show up to a junior tryout camp and make the team, such an event is almost unheard of in modern hockey. Therefore, not all the paths described in figure 2.4 can be considered relevant. This allows us to apply some simplifying assumptions to make the understand of the process more tractable. The next section presents an overview of the model where we introduce the assumptions and formulate our hypothesis.

2.3 The Model

A structural model is a framework that imposes a series of simplifying assumptions to a problem that makes the process tractable. The objective is to capture the essential features of the problem as a system of equations. The various factors influencing the model are treated as exogenous variables in the system of equations. The outcomes, or predictions, of the model are the endogenous variables whose values are the result of solving the system of equations.

The progression through the minor hockey system can be described by a *Markov* process. At any point in time a player will be in one of three states: tier one, tier two or exit. The player will then move to one of the same three states in the next period (year) based on a set of transitional probabilities. The probabilities are determined by the characteristics of hockey players. The characteristics that identify a player are natural ability, relative age, and level of training. These three characteristics completely describe a player.

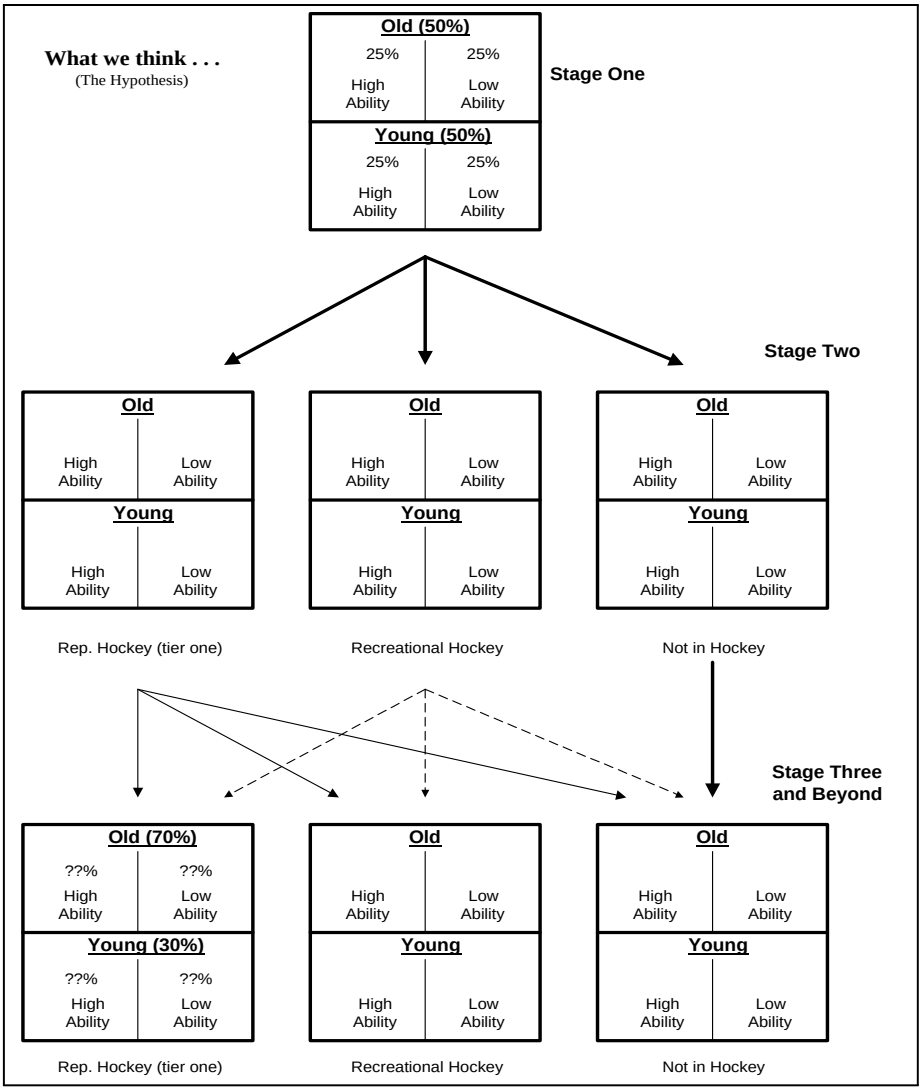


Figure 2.5: Outline of the structural model of hockey

By determining the set of probabilities at each stage, we can produce a dynamic process that replicates what is observed in minor hockey. Through the use of the computer and maximum likelihood techniques, the exogenous variables can be adjusted, or calibrated, such that the model produces a process similar to what is observed in the data. It is through such an estimation that the influence of natural ability, which cannot be directly observed, can be inferred and separated from those influences which are directly observable.

2.3.1 Assumptions

The first assumption of the model is that natural ability is independent of month or season of birth. Therefore the distribution of natural ability will be the same in each grouping of players based on birthdate. Furthermore, we assume that players are either of high ability or low ability and that there are equal numbers of both in the population. Thus any initial grouping will have 50% high ability and 50% low ability players. Second, we classify players within a calendar year as either young or old. Since both groups are born in the same calendar year, the difference between the two are what is called relative age. We assume that initially there are equal numbers of both age groups.

Based on our first two assumptions we have two characteristics that completely describe the population of players who enter the hockey system: relative age and natural ability. This implies four distinct groupings of our initial population. First they can be divided into either young and old players. then each of these two groups can be further divided into high ability or low ability. Since we assumed equal numbers

for each characteristic, then each of the four groupings will have equal numbers of players. The four groups are illustrated in the stage one box found in figure 2.5.

The third assumption of the model is that, all else held constant, players with high natural ability are more likely to move to tier one than players of low natural ability. Further, high ability players are also more likely to move to tier two and are less likely to exit than low ability players. Therefore, accounting for all other factors, we would expect tier one to be dominated by high ability players and the exit state to be dominated by low ability players. It is assumed that the influence of natural ability is unchanged throughout the entire process. We allow for some " noise", or randomness such that there will be both types of players found in all three states. The degree of noise in the model is determined from the maximum likelihood estimations.

The fourth assumption is that relative age effects the probability of a player being selected for tier one. It is assumed that old players are more likely to move to tier one than young players. Further, high ability players are also more likely to move to tier two and are less likely to exit than low ability players. Therefore, accounting for all other factors, we would expect tier one to be dominated by high ability players and the exit state to be dominated by low ability players.

The fifth assumption of the model is that players who are in tier one in the current period are the most likely to move to tier one in the next period than players who are either in tier two or out of the system. This is the *training effect*, which captures the influence of players receiving better coaching and greater ice time. The training effect is not considered to be cumulative. only the state that the player is currently in matters; not the player's overall history. In fact we make the extreme assumption

that once a player is out of hockey never re-enters either tier one or Tier Two. The assumption that players initially in tier one has the highest probability of moving to tier one in the next stage captures the *Training Effect*.

Finally, the relative age effect, is assumed to decay at each stage of the process, such that it is strongest in the stage one to stage two transition and weakest in the last transition period. The training effect is assumed to persist throughout the process.

The multi-staged process described above is known as a *Finite Markov Chain*. The transition probabilities at each step of the process make up what is referred to as the Markov *Transition Matrix*. The next section develops a formal model of the Markov process which will allow for a detailed statistical analysis of the minor hockey system.

2.3.2 Formalizing The Model

2.3.2.1 Initial Conditions

Each year a group of children born in the same calendar year begin playing hockey for the first time. The league first time players enter is named *peewee pup*. These players are randomly placed on teams where they all receive the same training and playing time. We can categorize the players in this league by two characteristics: relative age and natural ability. Any player who is born in the first six months of the calendar year as *old* and those born in the last six months of the year as *young*. Within each grouping (young or old), any given child is either a high ability player or a low ability player. We assume that half the children will be of high ability and

half of low ability. These two characteristics allows us to categorize these children into four groups:

Group I: old with high ability **(OH)**

Group II: old with low ability **(OL)**

Group III: young with high ability **(YH)**

Group IV: young with low ability **(YL)**

At the beginning of the second year, children in each group will find themselves in one of three states: either level A (tier one), level B (tier two), or they exit hockey altogether (E). Let A^i, B^i, and E^i denote the probabilities of any given child from group i $(i = I, ..., IV)$ moving to level A, level B, and Exit respectively. For example, consider group I moving from year one to year two. We can write the process as

$$OH_{t=1} \times \begin{bmatrix} A^I & B^I & E^I \end{bmatrix} = \begin{bmatrix} oh_A & oh_B & oh_E \end{bmatrix} = OH_{t=2}$$

where OH_1 is the number children in group I in year one, $\begin{bmatrix} A^I & B^I & E^I \end{bmatrix}$ is the vector of transition probabilities, and $\begin{bmatrix} oh_A & oh_B & oh_E \end{bmatrix}$ is a vector containing the number of group I children in each of the three possible states in year two; this vector is denoted by OH_2. In a similar fashion we can describe the transition process for groups II, III, and IV as follows:

$$\begin{aligned} OL_1 \times \begin{bmatrix} A^{II} & B^{II} & E^{II} \end{bmatrix} &= \begin{bmatrix} ol_A & ol_B & ol_E \end{bmatrix} &= OL_2 \\ YH_1 \times \begin{bmatrix} A^{III} & B^{III} & E^{III} \end{bmatrix} &= \begin{bmatrix} yh_A & yh_B & yh_E \end{bmatrix} &= YH_2 \\ YL_1 \times \begin{bmatrix} A^{IV} & B^{IV} & E^{IV} \end{bmatrix} &= \begin{bmatrix} yl_A & yl_B & yl_E \end{bmatrix} &= YL_2 \end{aligned}$$

In general, OH_t will denote the vector of states of group I children in year t.

Similarly, OL_t, YH_t, YL_t will denote state vectors at time t for children from groups II, III, and IV, respectively.

From the above process we see that all four groups of children are distributed across the same three states in year two. As we move from year two to year three, a child from any group found in any given state in year two can move to any given state in year three. The possible states in year three are the same as year two. For each subsequent year the possible states will be the same as the previous year's states. Starting in year two, the markov transition matrix for each of the four groups can be written as:

$$[M_t^i] = \begin{bmatrix} AA_t^i & AB_t^i & AE_t^i \\ BA_t^i & BB_t^i & BE_t^i \\ 0 & 0 & 1 \end{bmatrix} \qquad \left\{ \begin{array}{l} t = 2, ..., n \\ i = I, II, III, IV \end{array} \right\} \tag{2.1}$$

where each element in the matrix represents the probability of moving from a given state at time t to any given state at time $t+1$. For example:

- AA_t^i is the probability of a group i child in state A at time t moving to state A at time $t+1$;

- BA_t^i is the probability of a group i child in state B at time t moving to state A at time $t+1$;

- AB_t^i is the probability of a group i child in state A at time t moving to state B at time $t+1$;

- and so on...

Note that the third row comprises the vector $\begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$. This reflects the fact that once a child has exited hockey his probability of entering Level A or level B is zero, thereby making the probability of him staying out of hockey exactly one. Therefore the process is an *absorbing markov chain.*

The distribution of any group of children across states in year n can be expressed as the original vector in year two multiplied by the product of the transition matrices from year two to year n. Using our example of group I children (OH), the final distribution in year n can be expressed as follows:

$$OH_n = OH_2(M_2^I)(M_3^I)...(M_{n-1}^I)(M_n^I) \tag{2.2}$$

For each of the four groups there is an equation much like equation 2.2 that describes each group's progression through the minor hockey ranks[7]. In any given year the probabilities in each group's markov transition matrix may differ from those of the other groups'. In other words, the probability of a player in a given state (A or B) at time t moving to a particular state in year $t+1$ is a function of which group he belongs to. The reasons for these differences will be developed in the next section.

[7]If the markov transition matrix was a regular markov chain, where the probabilities were exogenous and constant over time, the previous expression would simply collapse to

$$OH_n = OH_2 \left[M_2^I\right]^n$$

2.3.2.2 The Logistic Model

So far, we have specified that the probability of success in hockey is a function of three variables: natural talent, the level of training, and the relative age of the player. Now we turn to the problem of adding structure to the model by specifying the functional form of the probabilities in the markov process. For this we have chosen the *logistic model,* commonly known in econometrics as the *Logit model*[8].

The first period probabilities are characterized as

$$\begin{aligned} A^i &= \frac{e^{Z_A^i}}{1+e^{Z_A^i}+e^{Z_B^i}} \\ B^i &= \frac{e^{Z_B^i}}{1+e^{Z_A^i}+e^{Z_B^i}} \\ E^i &= \frac{1}{1+e^{Z_A^i}+e^{Z_B^i}} \end{aligned}$$

where

$$Z_A^i = \alpha_0^1 + \alpha_1^1 AGE + \alpha_2^1 TALENT$$

$$Z_B^i = \beta_0^1 + \beta_1^1 AGE + \beta_2^1 TALENT$$

In the above specification, the $\alpha's$ and $\beta's$ are coefficients to be determined. AGE is a binary variable which takes on a value of 1 if the player is old and 0 if the player is young. $TALENT$ is also a binary variable which takes on a value of of 1 if the player has talent (high natural ability) and 0 if the player has no talent (low natural ability).

[8] *The logistic model has long been popular in qualitative analysis for a couple of reasons. First, it's simplicity allows researchers to economize on the number of variables to be estimated– frequently a serious issue in empirical work– as well, it's form makes it manageable to manipulate, allowing for closed form solutions. Second, the logistic model closely approximates a normal cumulative distribution function.*

For periods *2* through n, the probabilities in the markov transition matrix ($[M_t^i]$) specification

$$\begin{bmatrix} AA_t^i & AB_t^i & AE_t^i \\ BA_t^i & BB_t^i & BE_t^i \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{e^{Z_{AA}^{it}}}{1+e^{Z_{AA}^{it}}+e^{Z_{AB}^{it}}} & \frac{e^{Z_{AB}^{it}}}{1+e^{Z_{AA}^{it}}+e^{Z_{AB}^{it}}} & \frac{1}{1+e^{Z_{AA}^{it}}+e^{Z_{AB}^{it}}} \\ \frac{e^{Z_{BA}^{it}}}{1+e^{Z_{BA}^{it}}+e^{Z_{BB}^{it}}} & \frac{e^{Z_{BB}^{it}}}{1+e^{Z_{BA}^{it}}+e^{Z_{BB}^{it}}} & \frac{1}{1+e^{Z_{BA}^{it}}+e^{Z_{BB}^{it}}} \\ 0 & 0 & 1 \end{bmatrix}$$

where

$$Z_{AA}^{it} = \alpha_0^t + \alpha_1^t AGE + \alpha_2^t TALENT + \alpha_3^t TRAINING$$

$$Z_{AB}^{it} = \beta_0^t + \beta_1^t AGE + \beta_2^t TALENT + \beta_3^t TRAINING$$

$$Z_{BA}^{it} = \alpha_0^t + \alpha_1^t AGE + \alpha_2^t TALENT$$

$$Z_{BB}^{it} = \beta_0^t + \beta_1^t AGE + \beta_2^t TALENT$$

Since the relative age effect is assumed to diminish as children reach maturity, a decay variable was included in the model. The decay variable entered the model through α_1, β_1, the coefficients on the binary variable, AGE. The specific form of the decay process was

$$\alpha_1^t = \alpha_1^0 e^{-rt} \quad and \quad \beta_1^t = \beta_1^0 e^{-rt} \tag{2.3}$$

where r is the rate of decay, t is time measured in periods of the process and α_1^0,

β_1^0 is the initial magnitudes of the relative age effect at time $t = 0$.

2.4 Simulations of The Model

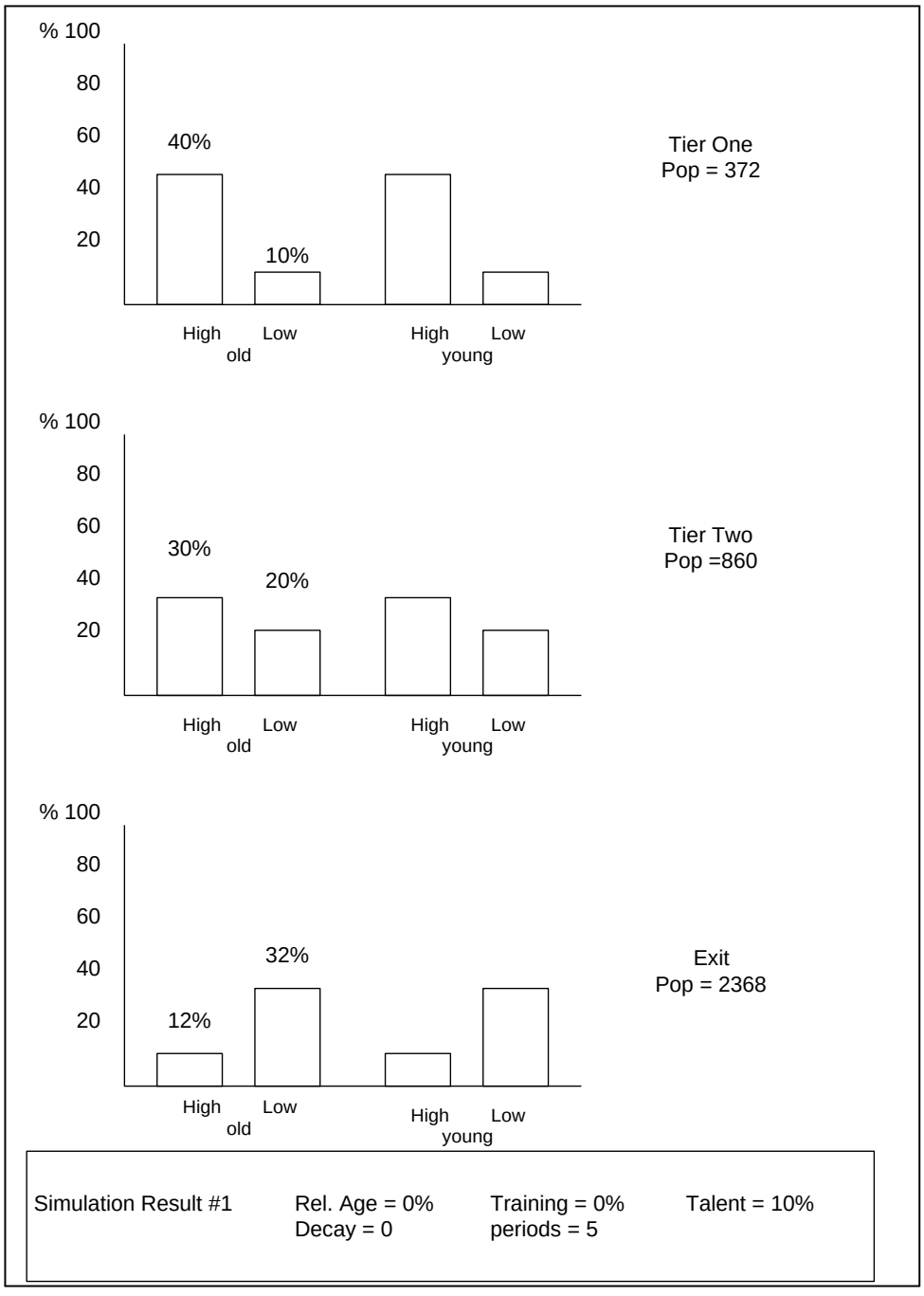
2.4.1 Initial Conditions

This section presents the results of simulations run on the model. An algorithm was written in *Maple* which carried out a multi-period markov process. Inputted into the program was the relative age variable, the training effect variable, the natural ability parameter, and the population in each category. In the simulations the total population was set at 3600. The population was further divided into four equal groups of 900, based on whether they were young or old, and were of high or low ability. The number of periods was set at five, consistent with the number of levels found in most Canadian minor hockey systems.

The model was then calibrated to produce a final period aggregate distributions across the three states (tier one, tier two, exit) that was consistent with observed populations found in the empirical data[9]. The calibration was carried out with the relative age and training effects set equal to zero and applying restrictions to the size of each state at every period. The value of the natural ability coefficient was estimated from the data for the first stage transition using maximum likelihood. A more detailed discussion of the calibration is found in Appendix I.

[9] Barnsley and Thompson (1988)

Figure 2.6: Simulation #1



2.4.2 Simulation Results for a variety of Relative Age and Training effects

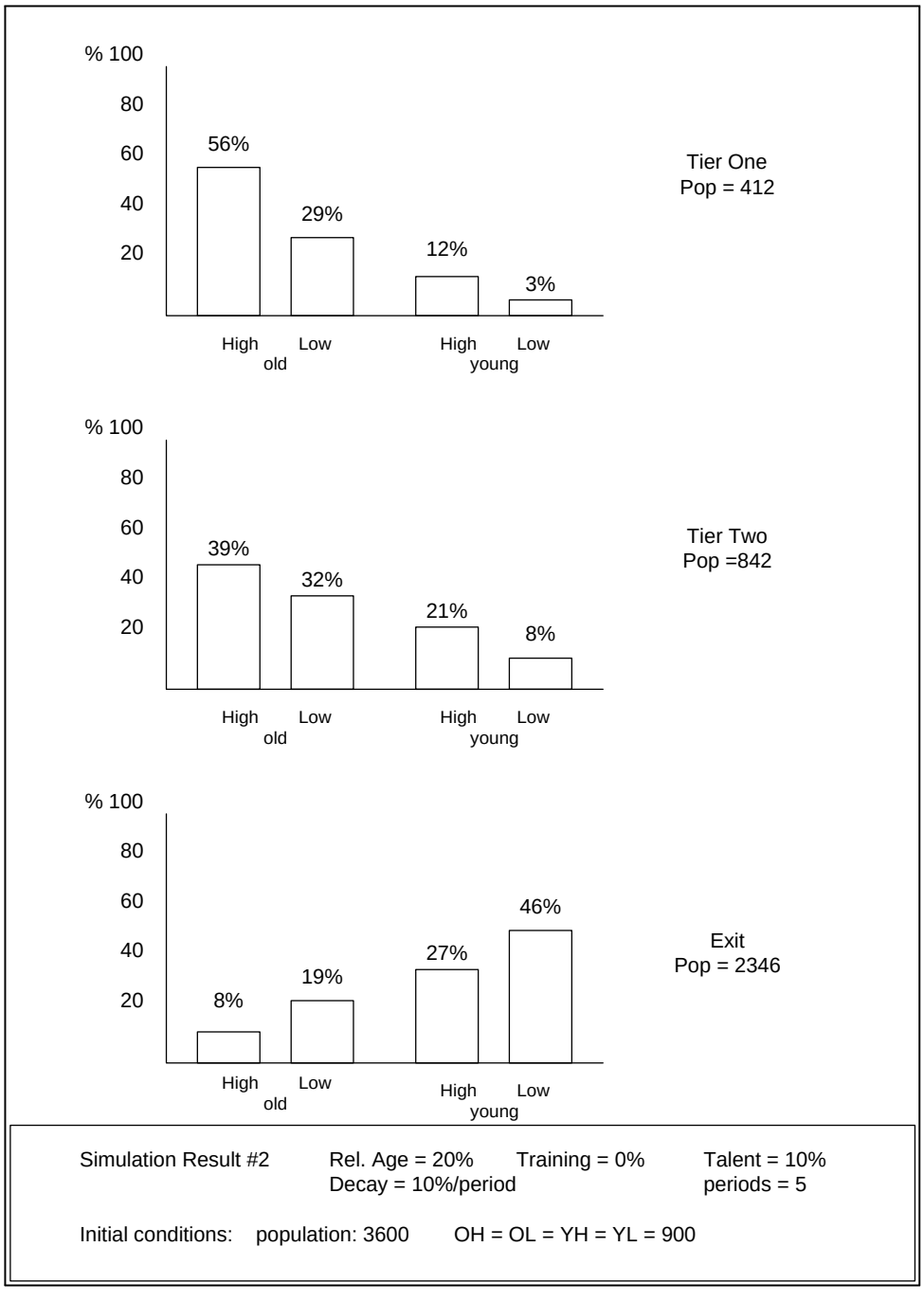
2.4.2.1 Simulation 1: calibration

The top graph in figure 2.6 shows the distribution of individuals found in tier one after 5 periods. Of the 3600 individuals in the simulation, 372 reached tier one by the fifth period. The 372 were equally divided into young and old players, since the relative age effect in this case was zero. Furthermore, 80% of the players were of high ability and 10% were of low ability.

The middle and bottom graphs illustrate the results for tier two and exit, respectively. By the fifth period there were 860 individuals in tier two and 2368 individuals had exited the system. Again the distribution between young and old in each state is symmetric due to the zero relative age effect.

Once the model was calibrated, a series of simulations were carried out varying the both the magnitude and the rate of the decay of the relative age effect. Initially the simulations were carried out with the training effect set equal to zero, then they were repeated while varying the level of the training effect. A sample of the results are presented in figures three through six.

Figure 2.7: Simulation #2



2.4.2.2 Simulation 2: Relative Age = 20%, Decay = 10%, Training = 0%

In figure 2.7 the magnitude of the relative age effect was set at 20%[10] while the rate of decay was set at 10%. After five periods, the results are quite pronounced. In tier one, 56% of the individuals are of the older, high ability category, while only 12% are from the younger, high ability category. Furthermore, 29% of tier one is made up of older, low ability players and only 3% younger, low ability players. In comparing the results to figure 2.6, we see that older, high ability player participation increased by 16% in tier one, where younger, high ability player participation dropped by 28%. The participation by older, low ability players increased by 19%. In total, older players make up 85% of tier one participation after five periods.

Similar results are found in tier two after five periods. Older individuals represented 71% of tier two, whereas only 29% were from the younger categories. The older, high ability individuals increased their participation in tier two by 9% and older, low ability individuals increased their participation by 12%. younger, high ability individuals participation dropped by 9% and younger, low ability individual dropped by 12%.

Examining the exit state in figure 2.7, we see that 73% of the individuals who exit the system by the fifth period are from the younger categories; 27% of the high ability type and 46% of the low ability type. Only 8% of the older, high ability individuals have exited the system and 19% of the older, low ability individuals have exited. In

[10] A 20% relative age effect in this context means that, *ceteris paribus*, an individual born in the first half of the year has a 20% greater chance of being selected for tier one as compared to an identical individual born in the last half of the year in the first period.

comparison to figure 2.6, older, high ability individuals went from 12% to 8% of the number who exited whereas the younger, high ability individuals who exited went from 12% to 27% of the total.

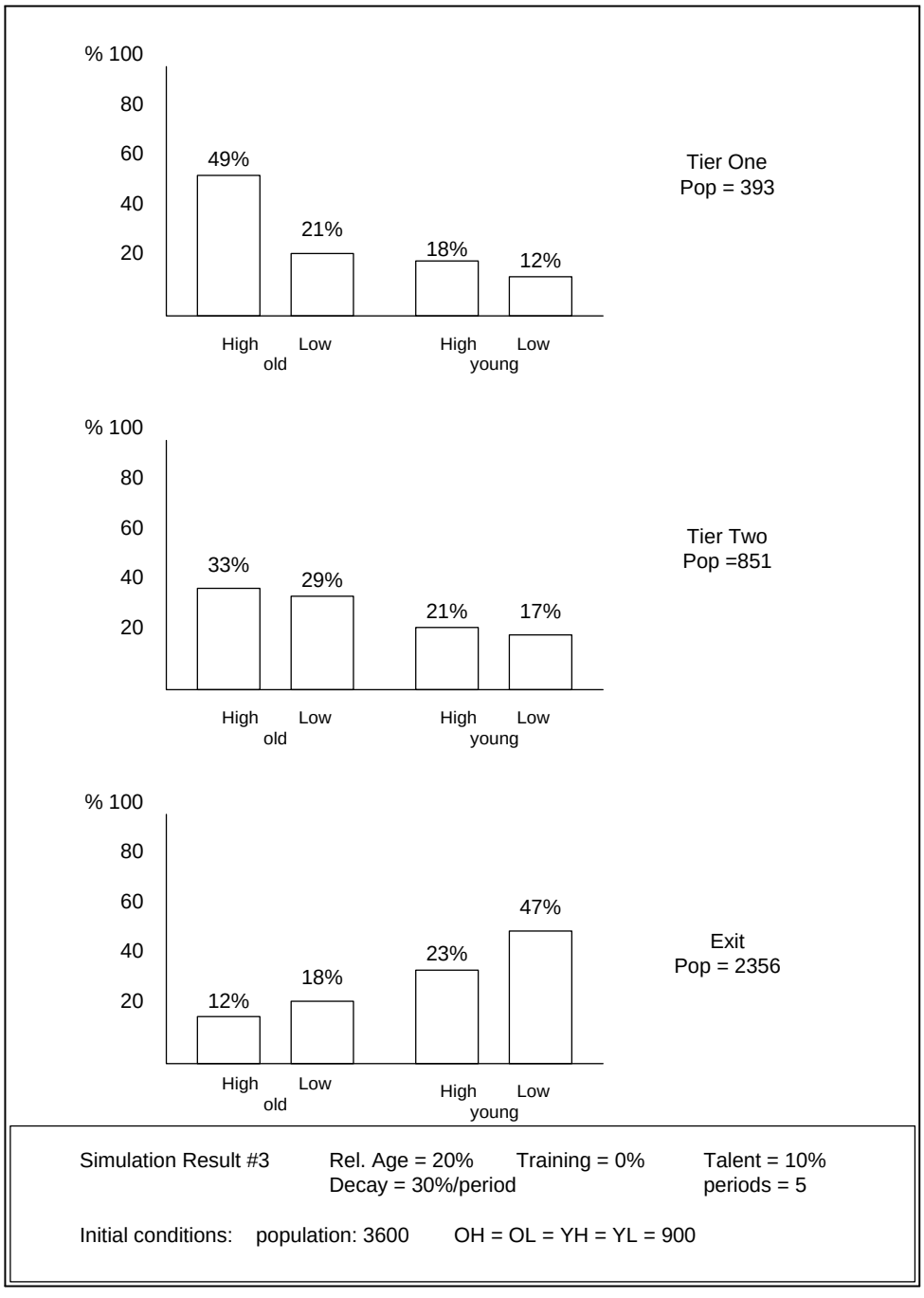
2.4.2.3 Simulation 3: Relative Age = 20%, Decay = 30%, Training = 0%

In the next simulation the rate of decay of the relative age effect (r) was increased to 30% while all other variables remained the same as before. The results are illustrated in figure 2.8. In tier one, 49% of the individuals were of the older, high ability type; 21% were of the older, low ability type; 18% were of the younger, high ability type; and 12% were of the younger, low ability type. In total, 70% of tier one were older individuals and 30% were from the younger categories. In comparison to figure 2.7, participation by older players in tier one fell from 85% to 70% of the total number of tier one individuals. Participation by individuals from the younger categories increased from 15% to 30% of the tier one population.

In tier two, 62% of the population were from the older categories; 33% from the high ability type and 29% from the low ability type. 38% of tier two were made up of younger players. 21% of tier two were of the younger, high ability type and 17% were of the younger, low ability type. In comparison to figure 2.7, participation by older players dropped by 9% of the tier two population whereas participation by younger players increased by 9%.

Of the individuals who had exited by the fifth period, 12% were older, high ability types; 18% were older, low ability types; 23% were younger, high ability types; and 47% were younger low ability types. In comparison to figure two, exit by older

Figure 2.8: Simulation #3



players increased by 3% of the total, while exit by younger players decreased by 3%. Interestingly, exit by high ability individuals from both age cohorts increased each by 4% of the total, as compared to figure 2.6. The number of low ability individuals from both age cohorts actually decreased as a percentage of total exits; each low ability age cohort fell by 1%.

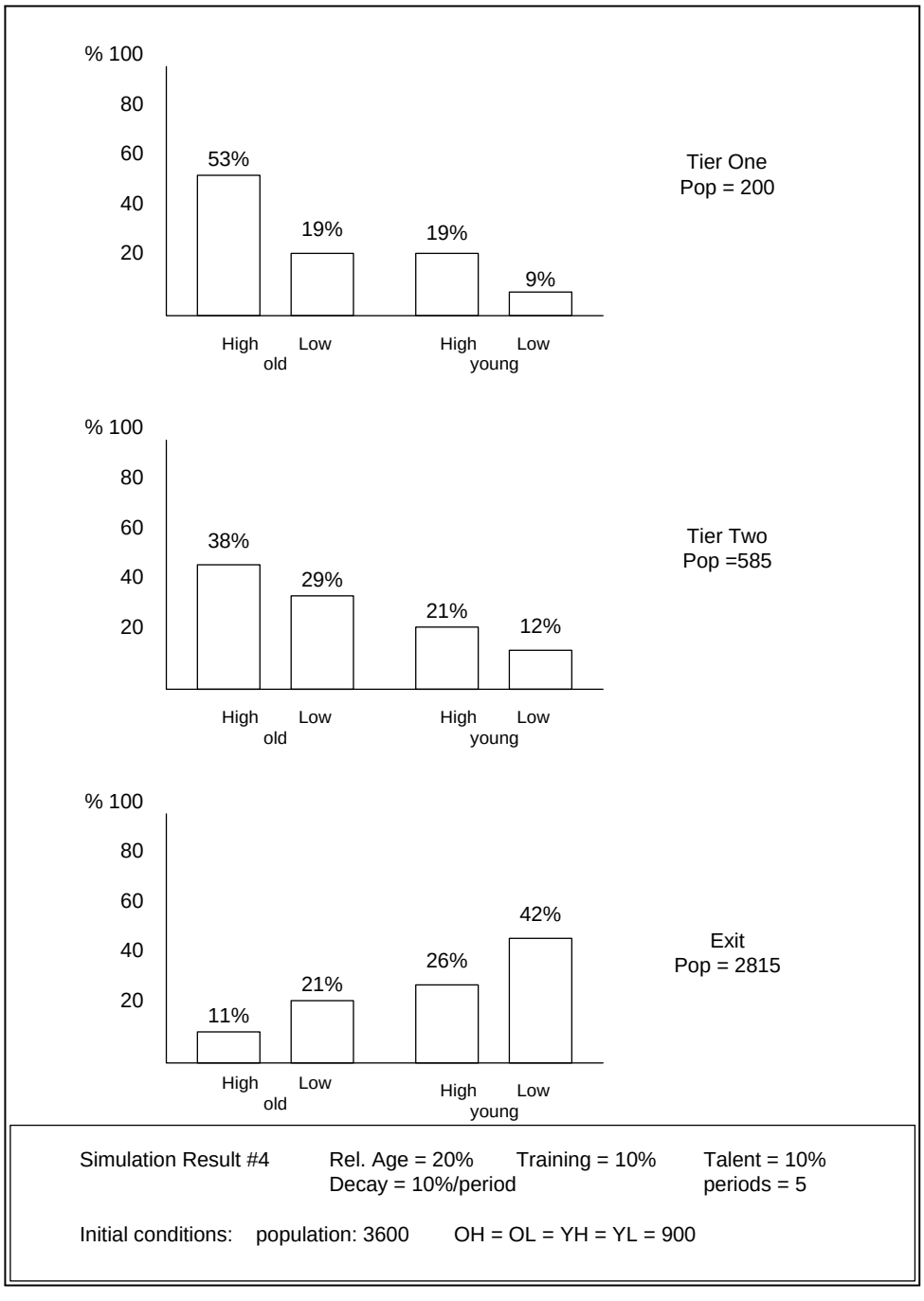
With a 20% initial relative age effect, increasing the rate of decay from 10% to 30% produced the largest change in the mix of tier one participants. The smallest effect occurred on the mix of individuals who exited the system. The number of low ability individuals from both age cohorts who exited the system was practically unchanged, while the number of high ability individuals from both age cohorts that exited increased slightly.

2.4.2.4 Simulation 4: Relative Age = 20%, Decay = 10%, Training = 10%

The next simulation involved incorporating the training effect. The results are illustrated in figure 2.9. In this simulation the relative age effect was set at 10%, the rate of decay was set at 10%, and the training effect (s_t) was set at 10%. A 10% training effect means that any individual who was in tier one in any given period will have a 10% better chance of moving to tier one in the next period (ceteris paribus). In this case training effects are not cumulative, therefore, only the previous period matters.

After five periods, the tier one population contained 53% older, high ability types, 19% older, low ability types. 19% younger, high ability types, and 9% younger, low

Figure 2.9: Simulation #4



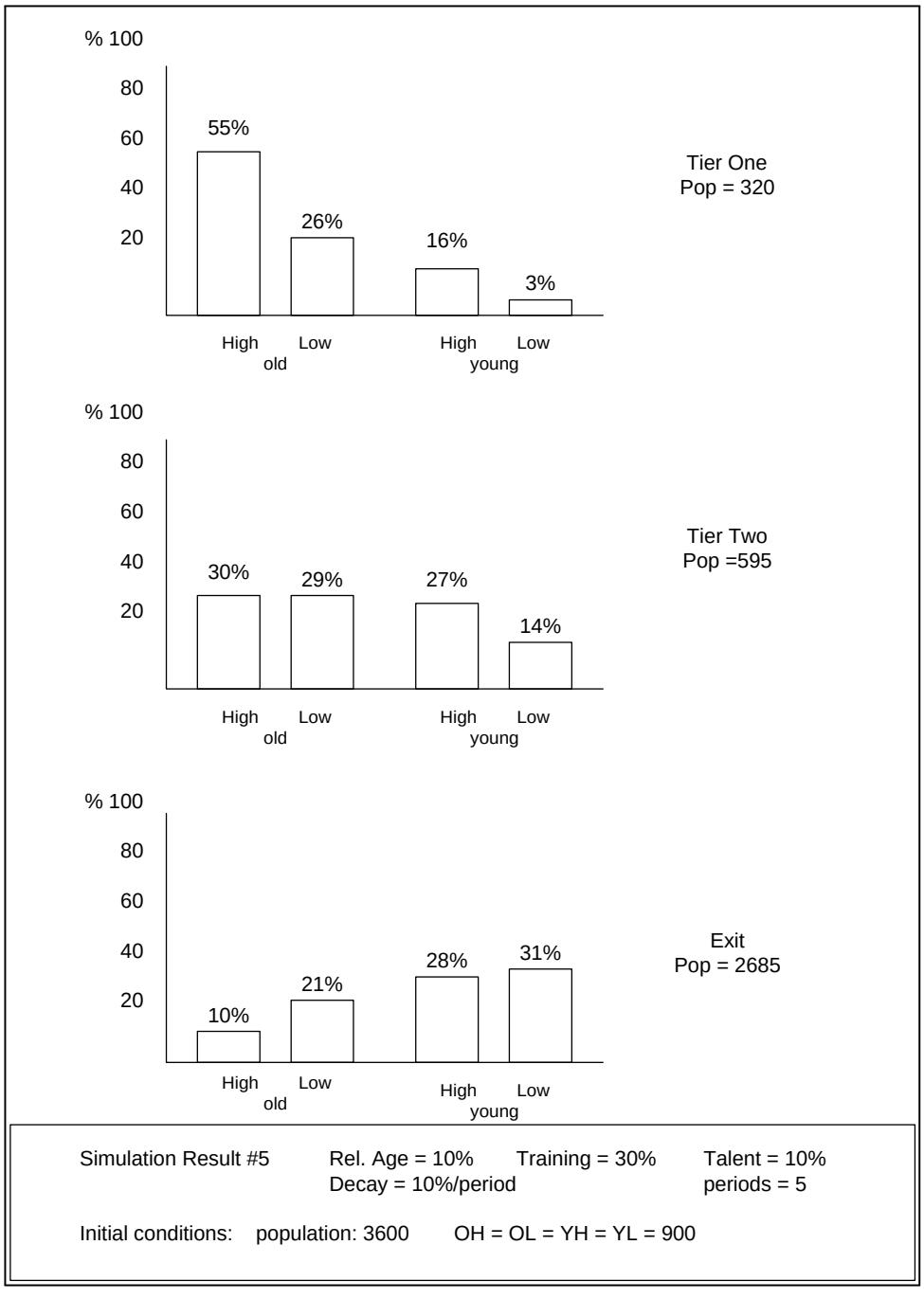
ability types. In total,72% of tier one was made up by individuals from the older age cohort and only 28 % from the younger age cohort. Tier two contained 38% individuals who were of the older, high ability type, 29% older, low ability type, 21% younger, high ability type, and 12% younger, low ability type. In total, tier two contained 67% older individuals and 33% younger individuals.

2.4.2.5 Simulation 5: Relative Age = 10%, Decay = 10%, Training = 30%

The training effect was then increased to 30% while all the variables were set to the same values as in figure 2.9. The results are illustrated in figure 2.10. In tier one, 55% of the population were older, high ability individuals; 26% were older, low ability individuals; 16% were younger, high ability individuals; and 3% were younger, low ability individuals. In all, 81% of tier one were older players and 19% were younger players. Increasing the training effect from 10% to 30% had a small effect on the percentage of older, high ability individuals (53% to 55%) but the percentage of older, low ability individuals did increase from 19% to 26% of tier one. The increased training effect lowered the younger, high ability individuals from 19% to 16% and lowered the younger, low ability individuals from 9% to 3% of tier one.

In tier two, 30% of the population were older, high ability individuals; 29% were older, low ability individuals; 27% were younger, high ability individuals; and 14% were younger, low ability individuals. In comparison to figure 2.9, the percentage of older, high ability individuals decreased while the other three categories all had percentage increases in participation. In total, 59% of tier two were older individuals and 41% were younger individuals. Interestingly, the distribution of both types of

Figure 2.10: Simulation #5



older players and younger, high ability players was approximately uniform (around 30%), and only the younger, low ability players were significantly different, making up only 14% of the tier two population.

In the exit category, 10% were older, high ability types, 21% were older, low ability types, 28% were younger, high ability types, and 31% were younger, low ability types. Younger players made up 59% of the total number who exited after five periods and 31% were older players.

2.4.3 General Findings from the Simulations

An extensive number of simulations were run, a sample of which has been presented above. Different functional forms for the probability functions were experimented with, including quadratic and linear models. In general the results were invariant to choice of functional form. Some results were quite obvious and fully expected. First, the smaller the difference in high and low talent the more pronounced was the relative age effects in the final period distributions across tiers. Second, the training effect magnified the relative age effect in tier one. However, when the training effect was relatively large (see figure five) distortions due to relative age in tier two and the exit category were less pronounced.

One interesting result from the simulations was the effect of the decay parameter, r. Even when the rate of decay per period of the relative age was quite large ($r \geq 30\%$) the final period distributions in all three categories showed a strong relative

age effect. This suggests that it is the magnitude of the relative age effect found in the first period that influences the final distribution rather than the persistence of relative age effects over time.

There were two significant findings that resulted from the model. The first was that even if the relative age effect was small but persisted for several periods (small r value) there would still be the distributions in the final period that matched those found in the data cited. The second result was that the relative age effect could be small and decay rapidly yet the existence of training would produce the same distribution as found when relative age effect was strong. While better data is required for further refining of the relative age and training parameters, one immediate conclusion at this stage is that child development is very sensitive to differentials in training in the early years.

Up to this point the focus of this paper has been on the relative age effect as found in sports; specifically minor hockey. Similar results have also been found in the education system. Research has found a strong correlation between biological maturity and cognitive development. This suggests that the results found in the analysis of sports may also be applied to the education system. The next section presents a discussion of the findings of relative age research in educational performance and learning disabilities.

2.5 Relative Age Effects in Education

In 1934 Elizabeth Bigelow, a grade one teacher in Summit, New Jersey, published a study which compared the performance of her students to their month of birth[11]. Her basic findings are consistent with the relative age hypothesis. Besides the quantitative results in her study, Ms. Bigelow assessed psychological and sociological problems observed in the classroom that was attributed to the relative age effect. Her conclusions suggested a high social cost to allowing chronologically younger children into a school system too early. Ms. Bigelow's work is probably the first documented study that identified the relative age effect in education. The more recent studies of the relative age effect in education cited in the introduction of this paper show results consistent with the findings of Ms. Bigelow.

The previous section introduced stylized facts about the relative age effect found in organized minor hockey. Here we turn our attention to the education system. In most countries the education system closely parallels the Canadian minor hockey system in structure. The school system is not as rigid as minor hockey; some children can be held back and others can be moved ahead by a year. But, by in large, most children found in a given grade are born within the same calendar year. Therefore we would expect to observe the relative age effect in performance among the early grades of school.

[11] Bigelow, Elizabeth B., "School Progress of Under- Age Children," Elementary *School Journal* (1934) 25, 186-192

Birthdates	Number of Boys	Number of Girls	Percentage Totals
January			
February			4%
March	0	2	4%
April	3	1	8%
May	1	0	2% 24%
June	5	2	14%
July	3	2	10%
August	3	2	10% 31%
September	4	2	12%
October	3	2	10%
November	6	4	20% 41%
December	4	2	12%
	32	19	100%

28%

72%

Figure 2.11: Summary of Grade one retentions (Alberta School District, 1985)

2.5.1 Relative Age in Grade One

Much of the focus of relative age research in education has been on the impact on students when they first enter the education system, typically grade one. Figure 2.11 presents a summary of retentions in grade one by month of birth. We see that 72% of children who are held back in grade one are born in the second half of the calendar year, and 41% born in the last three months. Similar studies have shown that the results found in figure 2.11 are consistent across school districts[12]. Further, findings from an ongoing study by Barnsley, Allen, and Thompson[13] show that retentions at the grade three, six and nine levels display similar results as those found in figure 2.11. They found that 10% of grade nines retained were born in the first quarter of the calendar year, while 40% were born in the last quarter of the calendar year. Their findings suggest that the relative age has a long run, or permanent, effect in academic performance.

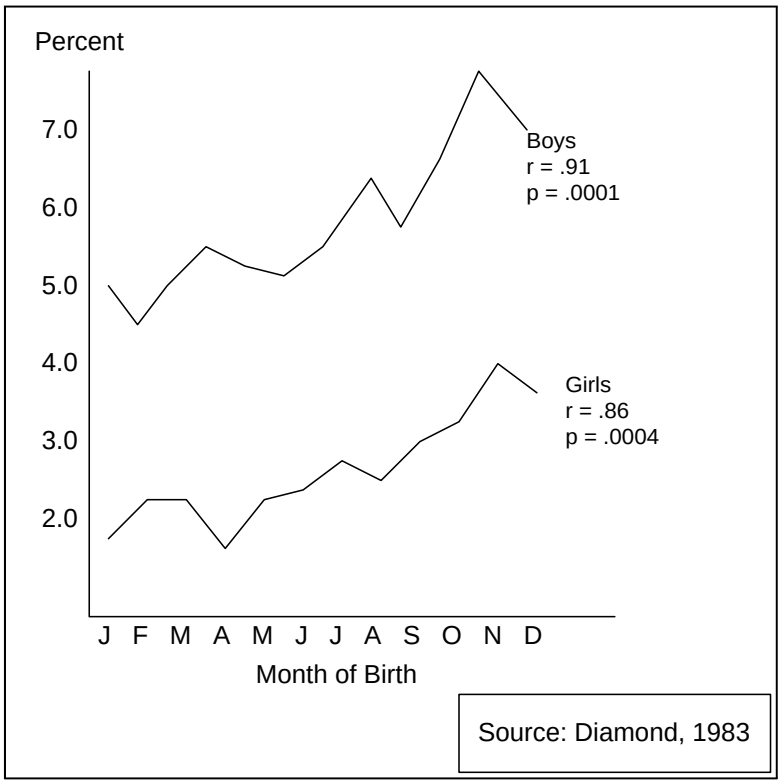
2.5.2 Relative Age and Special Education Placements

Research has shown that the age of entry to school is related to the incidence within some classifications of exceptional children. Diamond (1983) and Maddux (1980) both found that children born in later in the calendar year were over represented in programs for children with learning disabilities. Figure 2.12 shows the percentage of children born in each month classified as learning disabled. Further,

[12]See Barnsley, R.H. "Children Starting School: Readiness vs. Relative Age" *Educational Leadership*, (1986) 43, 91-92

[13]Barnsley, R.H., Allen, J., and Thompson, A.H. "School Achievement, Grade Retention and 'The Relative Age Effect'," *ongoing research.(from personal communication with authors)*

Figure 2.12: Percentage of children born each month classified as learning disabled



Maddux has shown that children who possess a relative age advantage are over represented as a group in programs for gifted children.

2.5.3 Relative Age and Academic Achievement

A common feature of most education systems is the periodic administering of achievement tests. In Canada such tests are administered at the grade 3, grade 6, and grade 9 levels. In Great Britain, there are two national tests, known as the *O-levels (age 13)* and *A-levels (age 17)*. Historically, these types of tests have lead to explicit and/or implicit streaming of the young. Figure 2.13 shows a summary of results of the nationally administered *Canadian Achievement Test (CAT)* for grades three, six and nine from the Lethbridge school district No 51. What is interesting about the data is that the persistence of the relative age effect through the first six years of education. Further, when comparing the 80% and 90% percentile in grade three, the relative age effect appears strongest when moving to the "tails" of the distribution; where scores are used to identify both "learning gifted" and "learning disabled" students.

2.6 Conclusion

A large number of researchers in the fields of education, psychology, and child development have extensively documented the relative age effect in sports and education. This effect has been linked to issues of grade retention and systematic biases in identifying children who have been labeled as *learning disabled*. Two hypotheses

RELATIVE AGE AND ACADEMIC ACHIEVEMENT
ALBERTA SCHOOL DISTRICT, 1985

	Winter	Summer	Autumn	N	χ^2
GRADE 3					
Total	34.5	34.7	30.6	480	
80th %tile +					
Math CAT	56.1	31.6	12.3	57	12.7 (p<.005)
ReadingCAT	52.5	30.5	16.9	59	9.9 (p<.010)
90th %tile +					
Math CAT	71.4	21.4	7.1	14	8.9 (p<.025)
ReadingCAT	75.0	18.8	6.3	16	12.2 (p<.005)
GRADE 6					
Total	32.7	34.3	32.9	434	
80th %tile +					
Math CAT	41.2	34.1	24.7	86	3.1 (p<.250)
ReadingCAT	39.6	33.1	27.3	139	3.2 (p<..250)
GRADE 9					
Total	31.9	33.9	34.1	504	
80th %tile +					
Math CAT	36.3	32.9	30.8	88	0.8 (n.s.)
Reading CAT	32.6	31.1	36.4	132	0.5 (n.s.)
Math& Read	34.2	32.2	33.6	77	0.2 (n.s.)

Figure 2.13: Relative age effect in Canadian Achievement Tests (CATs)

have been proposed regarding the relative age effect. The first hypothesis has been the training effect, where children are streamed into different classifications for the purposes of advanced training and development. With the existence of the relative age effect there will be a bias towards older children. Though the relative age effect is considered transitory, the influence of differential training will produce long run effects on the distribution of participants in any form of enriched or elite program.

The second hypothesis driving relative age effect is what is termed as the discouragement effect, which refers to children who are born late in the year being identified as learning disabled or poor performers. Therefore, whenever children later born in the year score poorly in tests or perform poorly in program and sports, they wind up with a stigma that they are in some way untalented, or that they lack the ability or skills. Both of these hypotheses are what drives the permanent nature of the relative age effect, so that long past the point where -biologically- relative age becomes irrelevant. Adults will wind up permanently categorized as either overachievers and high talented, while others will be low ability and underachievers. This gives rise to a classic type-1/type-2 measurement problem. Some talented but relatively young children will be overlooked in the streaming process while relatively older children may be mistakenly labeled as gifted when they are simply average. Both of these errors imply a social cost. There is the cost to the individuals from either the loss due to missed opportunities or the stress and frustration of being expected to perform at a level greater than their abilities will allow.

Of the studies cited, none have been able to analyze the relative age effect in

any systematic way. What this extensive body of research has accomplished is to document a widespread nature of this phenomenon. However without any kind of structural model, no true analysis can be carried out, and the ability to derive any meaningful policy recommendations are limited due to the inability to identify the magnitude or the duration of the Relative Age Effect. As an example, on several occasions provincial governments have explored the possibility of going into a dual entry model; where students would begin school in six month intervals rather than in an annual model. The most recent version would be the British Columbia provincial government's *Project 2000* initiative, which was eventually put on hold due to its potential. Given that the education budget is second only to health at the provincial level, the implications of the costs of this initiative would be immense. While there was a recognition of the relative age effect, analysts where not able to offer sufficient evidence to guide the scale or the duration of the project.

The question to be asked is whether or not this program is needed for the entire duration for a student's academic career, or if it could be dealt with within the first 3 or 6 years of education. So the purpose of the structural model developed in this paper would allow us to address those kind of questions directly. By using the structural model to extract different influences of relative age, natural ability and training, we're capable of comparative static exercises that may address some of the policy issues this problem raises.

The model allows us to identify when the relative age effect will cease to distort the screening mechanisms. We can also determine at what point differential training

effects compound the problem. Through the use of this kind of model, we would be able to comment whether or not there should be prohibition on enriched classes in first few years of schooling. Given the immense cost of the dual entry system, such as *Project 2000*, it would be prudent to be able to identify for how many years such a structure should be put in place. One of the results of the simulations was that even if the Relative Age effect decayed rapidly, the existence of any significant training during that period would still produce the kinds of long run distributions currently observed in the kind of studies now done in both sports and education.

One of the features of the structural model is that with the appropriate data it we have the ability to extract the natural ability parameters from the other influences. The data gives us a set of observable variables: we know who has received differential training and the relative age of members of cohorts. What can not be observed directly is - for any given group-, is the number who possess high ability or talent. By using maximum likelihood techniques in structural model, we're able to estimate and isolate natural ability from the influences of relative age and training. In essence, we get to observe and unobservable.

There were two significant findings that resulted from the model. The first was that even if the relative age effect was small but persisted for several years, there would still be the distributions in the final period that matched those we have observed. The second result was that the relative age effect could be small and decay rapidly yet the existence of training would produce the same distribution as found when relative age effect was strong. While more research and better data is required for further

refining of the relative age model built here, one immediate conclusion at this stage is that child development is very sensitive to differentials in training in the early years. Therefore deterring the use of explicit or implicit enriched programs in school systems during the formative years would go a long way to reducing the distortions that have been associated with relative age effect.

2.7 References

- Barnsley, R.H. "Children Starting School: Readiness vs. Relative Age" *Educational Leadership*, (1986) 43, 91-92

- Barnsley, R.H., Allen, J., and Thompson, A.H. "School Achievement, Grade Retention and 'The Relative Age Effect'," *ongoing research.(from personal communication with authors)*

- Barnsley, R.H., Thompson, A.H. and Barnsley, P.E.,"Hockey Success and Birthdate: The Relative Age effect." *Canadian Association for Health, Physical Education, and Recreation,* 51, 23-28,(1985);

- Bigelow, Elizabeth B., "School Progress of Under- Age Children," *Elementary School Journal* (1934) 25, 186-192

- Davis,B.D., Trimble, C.S., and Vincent, D.R., "Does Age of Entrance Affect School Achievement?" *The Elementary School Journal,* 80 133-143,(1980)

- Diamond, G.H. "The Birthdate effect- a Maturational Effect?" *Journal of Learning Disabilities*, (1983) 16, 161-164

- Ericsson, K. A., Charness, N., Feltovich, P. J., Hoffman , R. R., (editors) **The Cambridge Handbook of Expertise and Expert Performance** (June 2006) Cambridge Press

- Freyman, R. "Further Evidence on the Effect of Date of Birth on Subsequent School Performance." *Educational Research,* 8, 58-64, (1965)

- Jinks, P.C., "An Investigation into the Effect of Date of Birth on Subsequent School Performance." *Educational Research,* 6, 220-225, (1964)

- Maddux,C.D., "First Grade Entry Age in a Sample of Children Labeled Learning Disabled." *Learning Disability Quarterly,* 3, 79-83 (1980)

- Maddux,C.D., Stacey,D., and Scott, M. "School Entry Age in a Group of Gifted Children." *Gifted Child Quarterly,* 25, 180-184, (1981)

- Rosenthal R., and L. Jacobson. *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development.* New York: Holt, Rinehart & Winston, 1968

- Russell, R.J.H. and Startup, M.J. "Month of Birth and Academic Achievement." *Personality and Individual Differences,* 7, 839-846, (1986)

- Sutton, P. "Correlation between Streaming and Season of Birth in Secondary Schools." *British Journal of Educational Psychology,* 37, 300-304, (1967);

- Thompson, D. "Season of Birth and Success in the Secondary School." *Educational Research,* 14, 56-60, (1971).

2.8 APPENDIX I: Calibration of the Model

This section presents a brief description of the calibration of the model. The simulation process described earlier was designed to replicate a typical minor hockey system. Therefore the model was calibrated to fit a representative sample. The data was taken from the Edmonton, Alberta minor hockey association for the years 1983-1984[14]. The data covered five levels of organized minor hockey: (i) 9 and 10 year olds (Mite); (ii) 11 and 12 year olds (Pee Wee); (iii) 13 and 14 year olds (Bantum); (iv) 15 and 16 year olds (Midget); (v) 17 and 18 year olds (Juvenile). The data is summarized in 2.14, which shows the number of players in each level based on quarter of birth. For the levels Mite through Midget, the number of tier one players by quarter of birth is also given. The final level (Juvenile) is a tier one league. The relative size of the population at each level implicitly determined the number of players that exited the system.

The model was calibrated to approximate the aggregate date at each level as determined by a five period Markov process. For purposes of the simulation the first two quarters (Q-1 and Q-2) were aggregated into Old and the last two quarters were aggregated into Young (Q-3 and Q4). It was further assumed that each group could be further divided into two equal groups: high and low natural ability. In the model there are the three binary variables whose coefficients need to be determined (Talent, Age, and Training). Furthermore, there is a fourth variable; the rate of decay of the relative age effect.

[14]Source Barnsley and Thompson (1988)

Level	Tier	Old	Young	total	Total all tiers	Exit
1 Mite	n/a	1132	1078	2210	2210	
8 & under		51%	49%	100%		
2 Pup	Tier 2	760	812	1572		
9-10 yr		34%	37%	71%		
	Tier 1	216	106	322	1894	316
		10%	5%	15%	86%	
3 Peewee	Tier 2	602	623	1225		
11-12 yr		32%	33%	65%		
	Tier 1	213	116	329	1554	340
		11%	6%	17%	82%	
4 Bantum	Tier 2	475	488	963		
13-14 yr		31%	31%	62%		
	Tier 1	148	73	221	1184	370
		10%	5%	14%	76%	
5 Midget	Tier 2	314	297	611		
15-16 yr		27%	25%	52%		
	Tier 1	81	44	125	736	448
		7%	4%	11%	62%	

Figure 2.14: Minor Hockey Data (CAHA, Edmonton Alberta 1983-84)

The calibration process was carried out in two steps. The first step was a maximum likelihood estimate of the relative age and talent parameters. Of the four variables in the, only relative age is observable. Since the data is aggregate, the training effect can not be directly identified. This would require historical data on individual players- which was not available. However, since there is no training effect in the first period, only the age and talent parameters would be relevant. The results of a maximum likelihood estimate would allow for prediction of the number of high and low ability players in each cohort that moved to tier one, tier two, or exited in the first period transition. The derivation of the first period likelihood function is given below.

Using the results of the maximum likelihood estimates, we then carried out stage two of the calibration process. First we assumed that the coefficient on talent would remain constant across all periods of the Markov process. Then the age and training

effects were initially set equal to zero for periods 2 though 5. Then, by imposing restrictions on the aggregate populations found in tier one and tier two at each level, estimates for α_0^t and β_0^t $(t = 2..5)$ using a linear programing algorithm. The associated probabilities in each period's transition matrix produced the results found in simulation one of the chapter.

Deriving The First Period Likelihood Function

Initially, all individuals of each of the j cohorts are at the Mite level ($j = old, young$). Let N_j denote the number of individuals in cohort j. Of the N_j individuals, it is assumed that, initially, half are of high ability and half are of low ability. From the N_j individuals, a certain number are selected for tier one, (state A). Let m_j denote the number of players from cohort j at the Peewee Pup level who are selected to state A Peewee in the next period. The progression tree from Peewee Pup to Peewee is given in figure one.

Therefore, for any one player selected at random, the probability that he was of either type was 0.5. If the player was a high ability type, then he would be selected for A with probability $P_{Hj}(A)$. If he was a low ability type then he would be selected to A with probability $P_{Lj}(A)$. Given an individual is selected at random from cohort j, the probability that that individual would move to state A hockey in the next period would be

$$P(A|O) = 0.5P_A^{OH} + 0.5P_A^{OL} \tag{2.4}$$

$$P(A|Y) = 0.5P_A^{YH} + 0.5P_A^{YL} \tag{2.5}$$

Therefore, the likelihood function associated with exactly m_j individuals form a population of N_j being selected for state A hockey is

$$\begin{aligned} L_{Old} &= \sum_{i=0}^{m_o} \begin{pmatrix} m_o \\ i \end{pmatrix} (P(A|O))^i \begin{pmatrix} N_o \\ m_o - i \end{pmatrix} (1 - P(A|O))^{(N_o+i-m_o)} \\ &= \sum_{i=0}^{m_o} \begin{pmatrix} m_o \\ i \end{pmatrix} (P(A|O))^i \begin{pmatrix} N_o \\ m_o - i \end{pmatrix} (P(B|O) \\ &\qquad +P(E|O))^{(N_o+i-m_O)} \end{aligned} \tag{2.6}$$

$$\begin{aligned} L_{young} = &\sum_{i=0}^{m_y} \begin{pmatrix} m_y \\ i \end{pmatrix} (P(A|y))^i \begin{pmatrix} N_y \\ m_y - i \end{pmatrix} (1 - P(A|y))^{(N_y+i-m_y)} \\ &= \sum_{i=0}^{m_{yg}} \begin{pmatrix} m_y \\ i \end{pmatrix} (P(A|y))^i \begin{pmatrix} N_y \\ m_y - i \end{pmatrix} (P(B|y) \\ &\qquad +P(E|y))^{(N_y+i-m_y)} \end{aligned} \tag{2.7}$$

To estimate the relative age and talent parameters across all four cohorts simultaneously, the likelihood function to be estimated, denoted LL is simply the product of the likelihood functions for each of the j cohorts, or

$$LL = (L_{old})(L_{young}) \tag{2.8}$$

In the first stage the parameters to be estimated are α_0, α_1, α_2, $\beta_{0,}$ β_1, and β_2.

CHAPTER 3

DUAL ORGANIZATIONAL STRUCTURES IN FRANCHISING

3.1 Introduction

Franchising is often viewed as a contractual arrangement between two independent firms. The extent to which this is a useful description of a franchise agreement depends on the type of good and the nature of the activity being governed by the franchise contract. The presence of vertical restraints blur the distinction between two firms and one vertically integrated organization. In the case of franchising, the amount of vertical control arising from the contract is observed to vary across industries.

According to Sen[1], franchise operations can be divided into two types: "trade name franchising" and "business format franchising". The former includes sectors such as automobile dealerships, gasoline service stations, and soft-drink bottlers, while the latter format encompasses sectors resembling restaurants, hotels, real estate agents, business aids, and educational services. Typically, trade name franchising is a license to sell a wholesaler or parent company's product in a local market. Business

[1]Sen, Kabir C. "The Use of Initial fees and Royalties in Business-Format Franchising" Managerial and Decision Economics, Vol. 14, No. 2, Special Issue: Transactions Costs Economics. (Mar. – Apr., 1993), pp. 175-190.

format franchising involves licensing the use of a brand name for a locally produced product. With business format franchises, the local producer receives a business plan, specialized training and some form of ongoing support.

Each type of franchise operation poses several interesting issues that have been explored in the literature. While both types of franchises have similar terms and conditions within the standard contract, they differ by their focus. Trade name franchise issues tend to be found in the vertical restraint literature. They typically focus on issues of exclusive dealing, inventory controls and the problem of *double marginalization*[2]. Business format franchises, on the other hand, focus on the issue of moral hazard arising from informational asymmetries. These types of problems are found in the principle-agent literature, where the agent may shirk on a margin that is costly to measure (quality or effort). This paper is concerned only with the business format category of franchising.

There are two observations found in the empirical data on franchising that have not been adequately addressed in the literature. The first is the existence of both corporate-owned (and managed) outlets and franchised outlets within the same organization. Business Analysts have attributed this to the different types of activities carried out by a firm's corporate stores and franchise outlets. However, many of the chains that have both types of structures have homogeneous operations[3]. Models

[2] "Double marginalization" refers to the problem of both a wholesaler and retailer using a price markup formula. The wholesaler sell his good to a retailer at the wholesaler's profit maximizing price. The retailer, in turn, treats the wholesale price as marginal cost and marks it up a second time. To address this type of problem the wholesaler can choose from a variety of remedies such as: retail price maintenance (RPM), minimum quantity order, or two-part tariff.

[3] Lafontaine, Francine. "Agency Theory and Franchising: Some Empirical Results." Rand J. Econ. 23 (Summer 1992): 263-83. (a)

that address the choice between corporate and franchise outlets have predicted that one form or the other would come to dominate the organization[4]. This is not surprising since, with a couple of exceptions, such models tend to focus on the contractual arrangement between the franchisor and a single franchisee; with the optimal franchise contract defined in terms of the optimal franchise fee structure and monitoring levels.

The second unaddressed aspect is the apparent rigidity in various organizations' franchise fee structure over both time and between individual franchisees. According to the models, since franchise fees and royalties are chosen to extract economic rents subject to some type of incentive compatibility constraint, it is expected that franchise fees will vary across markets with different demand and (therefore) profit opportunities. This is not born out by the empirical evidence[5].

Two of the more prominent explanations put forward for the coexistence of corporate and franchise outlets in the same organization are signaling and brand reputation[6]. In the former, the franchisor uses corporate stores to signal to potential franchisees his commitment to the venture. In the latter, differences in brand reputation across markets determine the choice of arrangement (corporate or franchise). However, in both cases as the firm matures, the franchise fees would rise and a sin-

[4]Ozanne, U.B. and Hunt, S.D. The Economic Effects of Franchising (Washington, D.C.; U.S. Government printing office, 1971). Rubin, P. "The Theory of the Firm and the Structure of the Franchise Contract," Journal of Law and Economics , 21 (1978) 223-233.

[5]Lafontaine, Francine; and Shaw Kathryn L. "The Dynamics of Franchise Contracting: Evidence from Panel Data" The Journal of Political Economy, Vol. 107, No. 5. (Oct., 1999), pp. 1041-1080.

[6]*For signaling see* Gallini, N. T. and Lutz, N.A. "Dual Distribution in Franchising" J. Law, Econ., and Organization 8 (October 1992): 471-501; *and for brand reputation see:* Mathewson, F. and Winter, R. "The Economics of Franchise Contracts," The Journal of Law and Economics, Oct. (1985) 503-526.

gle structure would dominate. This paper offers an alternative explanation based on monitoring costs. It is the nature of the monitoring costs that determine both the decision to expand and to alter the choice of arrangement.

3.2 Structure of the Franchise Contract

In a franchise contract, a parent company contracts out the right to produce or market its product to an agent. Contractual stipulations specify rules governing the behavior of the agent, including pricing, mode of production, and territorial or market restrictions. A frequently observed feature of a franchised industry is that certain aspects of the parent company's product have limited scale economies that require production at the local market level.

A principle characteristic of franchise contracts is the agent's right to use a national brand name in exchange for a share of the profits. The brand name is a signal to consumers in a local market that the agent supplies a product of a certain quality. The effectiveness of the brand name as a quality signal will decide its value to consumers. Given the nature of brand names and the characteristics of certain industries that rely on them, franchise contracts, as a form of governance structure, may be the most efficient means of enhancing and protecting the value of the brand name.

Franchise contracts have certain common characteristics[7]. The franchisor sells or leases the right to produce or sell some product to a franchisee, and written into the contract are various obligations and commitments required by both parties. First,

[7]See, for example, Rubin, P. "The Theory of the Firm and the Structure of the Franchise Contract," *Journal of Law and Economics*, 21 (1978) 223-233; or Caves, R.E. and Murphy, W.F. "Franchising: Firms, Markets and Intangible Assets," *Southern Economic Journal*, 42 (1976)

with the right to use the franchisor's brand name, the franchisor also agrees to supply various types of assistance. This includes orientation with the production process, managerial and accounting assistance, site selection and development, and any on-going assistance or advice as required. The franchisor also takes responsibility for national marketing and advertising in addition to any research and development of the product.

Second, the franchisee agrees to operate the business in the manner stipulated by the franchisor, which includes hours of operation, pricing scheme, inventory levels, and adherence to the operating manual – if one is supplied. Third, the franchisee agrees to pay royalties to the franchisor. The royalties are usually in the form of a non-linear outlay schedule comprised of a fixed fee plus a share of the revenues.

Fourth, there will be a monitoring and auditing clause in the contract. This may be delineated explicitly, but will usually give the franchisor arbitrary and discretionary power. Fifth, the contract will have a termination clause that tends to favour the franchisor, who also has the ability to terminate the contract at will. However, the termination clause is less forgiving of the franchisee, who still has the ability to terminate the contract, but runs the risk of doing so at unfavourable terms and incurring heavy penalties. Finally, the contract will contain miscellaneous clauses dealing with the sale of the franchise, rights of heirs, territorial restrictions and any other conditions that may be specific to the particular product.

3.3 Explanations Of Franchising

Factors that have been put forth to explain the existence of franchise contracts include: risk pooling and capital-market imperfections, moral hazard on the part of the agent (franchisee), moral hazard on the part of the principal (franchisor), and information asymmetries on either the agent's or the principal's side.

3.3.1 Franchising As a method of capital accumulation and risk pooling

It was believed that franchising first arose as a form of capital accumulation and rapid expansion[8]. This line of reasoning can be discredited on two accounts. First, if an individual is to buy a franchise, he bears all the risk of that one outlet, whereas the franchisor has his risk spread across all outlets. To bear this higher risk, a risk-averse franchisee will demand a higher risk premium. The franchisor could therefore design a package of shares from all the outlets and sell them to the individual store managers, effectively lowering the risk premium he must pay and still maintain full control of the outlets. This form of organization will dominate since it is less costly.[9].Martin (1988) used risk sharing to explain franchise contracts. According to Patricia Lafontaine, the main empirically testable result from this model is that if the franchisor is less risk averse than the franchisee, the optimal royalty rate will increase as the amount

[8]See, for example: Hunt, S.D. "The Trend Toward Company-owned Units in Franchise Chains," *Journal of Retailing*, vol. 49, 2 Summer (1973), "Firms often choose the route of franchised units because they simply do not have access to the capital required . . ."; Caves and Murphy, *Supra note 7* , "For financing outlets the capital supplied by franchisees has no ready substitute...".

[9]Rubin *Supra note 7*

of risk increases. If the franchisor is risk neutral, this model implies that the chain should be wholly company-owned[10].

Second, evidence suggests that most franchisees have limited wealth[11], and therefore the funds they invest in a franchise must be acquired. With imperfect capital markets, it is unlikely that an individual would be more successful than a well-established firm at raising the needed capital[12]. Therefore, capital accumulation is not an adequate explanation of franchising[13].

3.3.2 Franchising as a solution to moral-hazard (agency) problems

This type of model assumes that franchisors cannot observe the behavior of franchisees in terms of their provision of local input, service for quality level. They also cannot infer it from the observed level of sales if there is a stochastic element in local demand conditions. Franchise contracts are a solution to a monitoring problem when reputation is an important factor in the exchange of a good[14]. Franchise contracts allow an agent to earn a quasi-rent stream from producing and/or selling a parent company's product in a local market. The purpose of the quasi-rent is to ensure compliance on the part of the agent to the terms of the franchise contract.

Klein and Murphy[15] argue that franchise contracts allow an agent to earn a quasi-

[10] Lafontaine and Shaw *Supra Note 5*

[11] Mathewson and Winter *Supra note 6*

[12] In an interview with George Tidball, founder of *the Keg* restaurant chain, it was reported that the Keg corporation usually financed the franchisee's purchase of the franchise rights. This was a loan agreement where the terms of repayment were independent of the annual royalties that the franchisee would pay to *the Keg*.

[13] Rubin, P. *Supra note 7.*

[14] Mathewson, F. and Winter R. *Supra note 6.*

[15] Klein, B. and Murphy, K. " Vertical Restraints as Contract Enforcement Mechanisms," *The Journal of Law and Economics,* Oct. (1988) 265-297

rent stream from producing and/or selling a parent company's product in a local market. The purpose of the quasi-rent is to ensure compliance on the part of the agent to the terms of the franchise contract. They further demonstrate that, in equilibrium, there must be a positive level of monitoring; the rent stream by itself is not sufficient to ensure compliance.

3.3.3 Franchising and reverse moral-hazard problems

A third explanation for franchise contracts relies on moral hazard problems for both parties. Here the franchise contract arises due to the mutual need for incentives. As in the previous section, the franchisee may have an incentive to shirk on the supply of an input whereas may also have an incentive to renege on his part of the agreement. To solve the moral hazard problem on the part of the agent, the franchisor could require that the franchisee put up a forfeitable bond that would be lost with non-compliance[16]. However, this creates a reverse moral hazard problem. If the bond is sufficiently large, the franchisor may renege on his promise to maintain the brand name and therefore abscond with the bond.

In addition, if the franchisee had sufficient wealth to afford an adequately sized bond, he would invest in a more diversified and less risky asset that has fewer constraints on his managerial sovereignty than a franchise. This implies a wealth constraint on the franchisee, which is a necessary condition for a franchise contract[17].

[16]For further dicussion on this form of constraint see: Klein, B. "Borderlines in Law and Economics: Transaction Cost Determinants of 'Unfair' Contractual Arrangements," *American Economic Review*, 70, 2 May (1980) 356-362.

[17]It is a lack of collateral that makes a franchise contract superior to any privately negotiated loan agreement a bank could offer the individual. A limited wealth condition is equivalent to a default

3.4 Geographic Issues of Franchise Contracts

The explanations for franchising described above typically focus on a single franchisor-franchisee contract. As a consequence, their results center on the terms of the optimal contract - the royalty rate and level of monitoring. None of these models can directly explain the coexistence of both franchises and corporate stores as it occurs in franchising. These models typically imply a different optimal royalty rate for each franchisor-franchisee pair, even within a single chain. This would suggest that, in cases where market conditions differ across locations, there would be a high degree of heterogeneity in terms of franchise fees and royalty rates.

As previously mentioned, there are two observed facts in industries that use franchising to produce and distribute their product that have not been adequately explained[18]. The first is the breakdown between corporate owned and franchised outlets found within a given organization. It is repeatedly observed that an organization that engages in franchising will frequently buy back certain franchised outlets and operate them as corporate stores, but simultaneously issue franchises in new areas. Furthermore, there appears to be little correlation between the size of the economic rent that individual outlets are earning and the decision to buy them back.

The second unexplained observation is the fact that franchise fees remain relatively fixed, both across outlets and over time, while there is a wide variability in rent

option on loans to franchisees so that banks incapable of writing performance contracts superior to franchisors will rationally limit their loans to franchisees that ease the purchase of the local right to the brand name, knowing incentives in a franchise contract. The limited wealth constraint as a necessary condition for franchising is a well established result in the literature. See, for example Mathewson, F. and Winter, R. *Supra note 6*; or Rubin, P. *Supra note 7.*

[18]See: Lafontaine, Francine.*Supra note 3; and* Simon, Carol J., "Franchising vs. Ownership: a contracting explanation", *University of Chicago working paper* (1991). This paper presents the results of an extensive survey of franchise contracts across the midwest United States.

being earned across outlets[19]. This fact appears to be inconsistent with the proposition that franchise fees allow the parent company to capture some of the economic rent being earned by the agent[20].

Two alternative frameworks that allow for the coexistence of franchising and corporate outlets are Gallini and Lutz's signaling model[21] and Mathewson and Winter's brand reputation model[22]; however both imply that franchisors will want to reduce their royalty rates and increase their franchise fees over time. This occurs in the former because information about franchisor quality is revealed over time, and occurs in the latter because of the franchisor's increased brand equity.

Incentive compatibility constraints determine the extent to which a parent company can capture the economic rent being earned by the individual outlets. If one assumes that individual franchisees have similar opportunity costs, then one would expect the economic profit required to ensure compliance to be the same across franchises. Therefore, if economic rents vary across outlets, the residual would be captured by an adjustable franchise fee. One would expect the parent company to set each outlet's franchise fee based on local market conditions.

A common characteristic to franchise industries is that aspects production and distribution are carried out by many small, geographically displaced outlets. Therefore, when the parent company wishes to monitor its outlets, the monitor must incur considerable transportation costs[23]. In a large chain, this will require the monitor

[19] Lafontaine and Shaw *Supra Note 5*
[20] See Tirole, J.**The Theory of Industrial Organization,** chapter 4 (1988).
[21] Gallini and Lutz *Supra note 6*
[22] Mathewson and Winter *Supra Note 10*
[23] The costs of procedure - or quality control - audits are nontrivial. For example, McDonald's

to cover a large area in the execution of his duties. One would therefore expect the remoteness of an outlet to have a bearing on the choice of contractual arrangement between the parent company and the local operator.

If the location of outlets and the distance between outlets were a function of market density, one would expect to see a clustering of outlets in more densely populated areas. This gives rise to an asymmetric distribution of stores, which will have a significant effect on the costs of monitoring. If the monitor has to travel a significant distance to inspect a particular outlet, then frequent monitoring will be quite costly. However, if there is a second outlet in close proximity to the first outlet, then the marginal transportation cost of monitoring the second store will be quite low.

This implies a *non-convexity* in the monitor's cost function and it is this non-convexity that will affect the choice of contract between the parent company and the individual outlets. In the case of one outlet being geographically displaced from the monitor, it may be more profitable to give the local agent an economic rent rather than frequent monitoring to ensure compliance. However, if a second store is established in close proximity to the first, it may be more profitable for the parent company to switch to extensive monitoring and reclaim the rents.

While this point may seem straightforward with respect to the parent company's decision to franchise a new outlet, it also implies something more. The decision to expand the number of outlets and the decision to change the form of the contract between the parent company and the local operator are two aspects of one decision.

---

will send a team of 2-3 auditors to a given outlet for up to a week each time they engage in a scheduled audit.For this type of audit every aspect of the operation is scrutinized. In addition, remote monitoring is carried out almost continuously and any anomalies can trigger a site audit.

This may explain why one form of contract has not come to dominate the other over time, which is something that has been predicted by analysts of these industries[24].

With respect to the observed rigidity of franchise fees, this too may be best explained in a geographic context. When a local market grows, so does the rent earned by the local franchisee. So why doesn't the parent company increase the franchise fee accordingly? One would expect that this would be a fairly straightforward clause to include at the outset of the franchise agreement.

Viewed as a principle-agent problem, it is assumed that, if the franchisee has better knowledge of local market conditions than the parent company, the franchisee would be in a better position to judge whether the local market could support expansion. In most franchise agreements the franchisee has the right of first refusal when a second outlet is being considered within his territory. However, a second outlet would be subject to the same structure of franchise fees and royalties regardless if it was operated by the incumbent or a new franchisee.

Furthermore, given diminishing returns to the ability of a single outlet to service a growing market, the parent company could better increase total royalty revenue from a given market by establishing a second outlet. The profitability of expansion will be further enhanced because of the nature of the monitoring costs. The existence of the second store will lower the economic rent that was going to the first store before expansion. The lower profit will give the agent in the first store a greater incentive to shirk and therefore greater monitoring will be required. However, with the existence of the second store, a decrease in the *relative cost* of monitoring the first store may

[24]The list includes: Caves and Murphy *Supra note 7;* Hunt, S.D. *Supra note 8*

now make an increase in the frequency of monitoring worthwhile compared to the pre-expansion period.

3.5 The Model

3.5.1 Initial conditions

A parent firm, or franchisor, sells his product in a set of geographically dispersed markets, or nodes. In each market there is an outlet where final production and sales are carried out by an agent, or franchisee. Demand conditions are assumed to vary across markets and each agent is assumed to have better information about local market conditions than the franchisor. In each market it is assumed that the agent faces a downward sloping demand function for the final product or service.

At any given outlet the agent may be a franchisee or simply an employee of the franchisor. If the latter is the case, then the outlet is referred to as a corporate store. Denote the location, or address, of a local market by x_0 ($x_0 > 0$). The location of the franchisor will be normalized to be zero. Therefore x_0 represents the distance between the franchisor and the local market.

The franchisor produces a good at a constant cost of v per unit. The good is distributed by the agent to the local market. The agent also contributes additional input into the final good in the form of services or some other quality enhancing attributes. Let s denote the level of service provided by the agent and let $c(s)$ be the agent's cost of s where $c'(s) > 0$ and $c''(s) > 0$. Finally, each outlet incurs fixed cost of K.

Demand in the local market is a function of both price, p, and the level of services, s.[25] Let q denote quantity demanded at location x_0 and the demand function is given by

$$q = D(s, p) \tag{3.1}$$

where

$$\partial D/\partial s > 0 \quad and \quad \partial D/\partial p < 0$$

The franchise contract specifies a payment schedule plus a level of s. The schedule for which the agent pays the franchisor royalties takes the form of a two-part tariff with a fixed and variable component:

$$f + \alpha(p - v)D(s, p) \qquad (where\ 0 \leq \alpha \leq 1) \tag{3.2}$$

f is the lump-sum franchise fee and α is the share of sales revenue that accrue to the franchisor.

3.5.2 The decision to shirk

Given the franchise contract, the agent may have an incentive to shirk on the level of services he is to supply. The decision to shirk will be a function of (i) the profits from shirking; (ii) the probability of detection by the franchisor; and, (iii) the

[25] Local demand is also a function of the strength of the national brand name. For our purposes, this is assumed exogeonous and is therefore supressed in the model.

penalty, or sanction from shirking. The probability of detection will, in turn, depend on the level of monitoring activity that the franchisor engages in and the degree by which the agent lowers the level of services below the contractually specified level.

Define ϕ as the frequency of monitoring carried out by the franchisor, which is normalized to be between 0 and 1. Furthermore, define Δs as the difference between the contractual level of s (denoted s^*) and the actual level of s supplied by the agent (i.e. $\Delta s = s^* - s$). Therefore the lower the actual level of services relative to the level specified in the contract, the greater will be Δs ($\Delta s = 0$ implies no shirking). The probability of the agent being detected shirking will be a function of both the frequency of monitoring and the degree of shirking by the agent[26]. Let δ denote the probability of detection, which can be expressed as follows:

$$\delta = \delta(\Delta s; \phi) \tag{3.3}$$

where

$$\partial\delta/\partial\phi > 0 \quad and \quad \partial\delta/\partial\Delta s > 0$$

In most franchise contracts the penalty for shirking is termination of the franchise agreement[27]. Therefore the expected profit from supplying a low level of services can be expressed as

[26]It is also possible that $s > s^*$, in which case the franchisee is supplying a level of service greater than the level specified in the contract. This may lead to *intra franchise* competition which lowers the franchisor's rents. Most franchise contracts will also attempt to minimize this form of behavior. For a more formal treatment, see Winter, R. A. "Vertical Control and Price versus Non-price Competition", The Quarterly Journal of Economics Vol. 108, No. 1 (Feb., 1993), pp. 61-76.

[27]See Klein, B. *Supra note 16*

$$\pi_L = (1 - \delta(\Delta s; \phi))\pi(p, \Delta s) \tag{3.4}$$

where $\pi(p, \Delta s)$ is the agent's profits as he deviates from the contracted level of s.

Differentiating (3.4) with respect to Δs solves for Δs (and therefore s) that maximizes the agent's expected profits from shirking, or

$$(1 - \delta(\Delta s; \phi))\partial\pi/\partial\Delta s - \pi(p, \Delta s)\partial\delta/\partial\Delta s = 0 \tag{3.5}$$

For any given level of ϕ the profit function of the agent (3.4) is at first increasing, then decreasing in Δs. Intuitively this results from the fact that as the level of service falls, the expected profits for the agent rises from the cost savings. However, as the level of services falls, the probability of detection rises, thus lowering expected profits. ϕ is a shift parameter in the expected profit function. Expected profits as a function of shirking on services (Δs) are illustrated in figure 3.1.

Therefore, if $\pi_H = \pi(p, 0)$ is the profits of the agent when no shirking occurs, then the agent will choose to shirk if, at the s that maximizes (3.4),

$$\pi(p, 0) < (1 - \delta(\Delta s; \phi))\pi(p, \Delta s) \tag{3.6}$$

Equation 3.10 represents the incentive compatibility constraint faced by the franchisor.

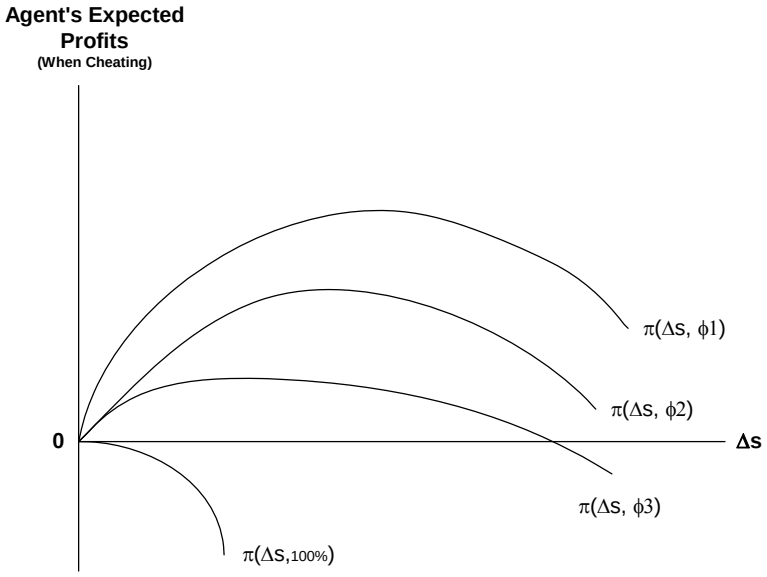


Figure 1: The agent's expected profit as a function of the level of shirking, given some known frequency of monitoring (ϕ). Changes in the frequency of monitoring will shift the agent's expected profit function.

Figure 3.1: Expected profits from shirking

If the franchisor decides to monitor the agent he will incur monitoring costs which are denoted as M. The costs of monitoring will be an increasing function of both the frequency of monitoring and the remoteness of the agent. Therefore the costs of verifying performance are

$$M = M(\phi, x_0) \tag{3.7}$$

where

$$\partial M/\partial \phi > 0 \quad and \quad \partial M/\partial x_0 > 0.$$

The Franchisor's objective is to maximize

$$\Pi(s, p, \phi) = f + \alpha(p - v)D(s, p) - M(\phi, x) \tag{3.8}$$

by choice of p, s, α, ϕ, and f subject to

$$\pi(p, s) = (1 - \alpha)(p - v)D(s, p) - c(s) - f - K \geq 0 \tag{3.9}$$

and

$$\pi(p, 0) \geq (1 - \delta(\Delta s; \phi))\pi(p, \Delta s) \tag{3.10}$$

Equation 3.9 is a non-negativity constraint on the agents profits[28] and equation 3.10 is the incentive compatibility constraint. Let λ_1 and λ_2 denote the lagrange multipliers for equations (3.9) and (3.10) respectively. Applying Kuhn Tucker conditions and noting that (3.9) is non-binding in the presence of (3.10), we get the following results:

$$p(1 - 1/\epsilon_H) = v + k(1 - \delta)(p(1 - 1/\epsilon_L) - v) \tag{3.11}$$

$$\alpha(p - v)\partial D/\partial s = kc'(s) \tag{3.12}$$

$$\partial M(\phi, x_0)/\partial\phi = \lambda_2((\delta - 1)\partial\pi_L/\partial\phi + \pi_L\partial\delta/\partial\phi) > 0 \tag{3.13}$$

$$where \quad k = \frac{1}{\alpha/\lambda_2 + (1 - \alpha)} = \frac{\lambda_2}{\alpha + \lambda_2(1 - \alpha)} > 0$$

In equation (3.11), ϵ_H is the price elasticity of demand in the local market for a given s^* and ϵ_L is the price elasticity of demand when the agent chooses to shirk. Equation (3.11) implies that the price of the final product will be higher when the incentive to shirk is absent. Intuitively, the franchisor is forced by the incentive compatibility constraint to engage in a quality/quantity trade-off in order to reduce

[28] For simplicity, it is assumed that the agent's opportunity cost is zero.

the marginal returns to shirking.

Equation (3.12) determines s^*. If $\alpha < k$ then the level of services will be set below the first best level. Equation (3.13) sets the level of monitoring and implicitly determines the rent stream accruing to the agent. Since λ_2 is the shadow price of compliance, it can be interpreted as the marginal benefit of increased local market demand (and profits) when the service level is maintained. Thus a growth in the local market would lead to a higher value of λ_2 and, from equation 3.13, imply an increase in monitoring. Also, from equation 3.12, an increase in λ_2 would increase the incentive for the franchisee to reduce s, therefore creating an incentive for the franchisor to implement an off-setting reduction in the royalty α (As cited earlier, this result is consistent with the findings of other models but is not supported by the data).

Since $\partial M/\partial x_0 > 0$, we can see from equation (3.13) that as the distance between the franchisor and the outlet increases, the level of monitoring will decrease and the rent stream to the agent will rise. This result is illustrated in figure 3.2. Figure 3.2 illustrates the marginal benefit to monitoring and the marginal costs with-and without- the effect of transportation costs. In this case, the graph shows an interior solution (point A) due to the transportation costs ($x = x_0$). When there are no transportation costs ($x = 0$), a corner solution arises with 100% monitoring. With perfect monitoring there is no need to an economic rent stream to create an incentive to maintain the service level specified by the franchisor. At this point there would be a change in the contract mix as the outlet becomes corporate. In this framework

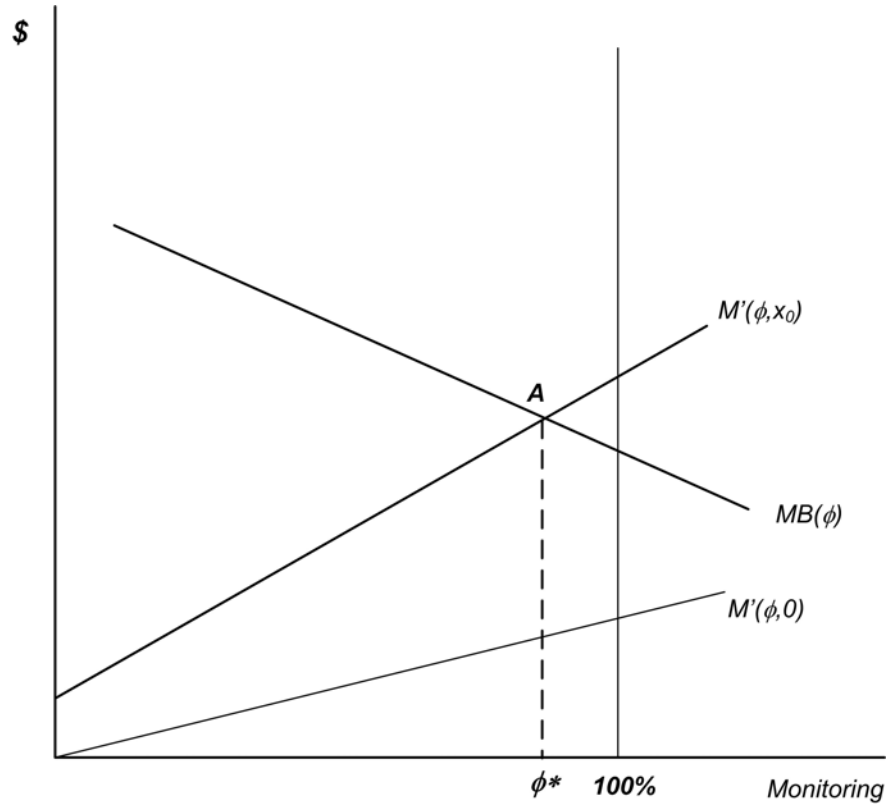


Figure 3.2: Monitoring Equilibrium

$\phi = 100\%$ implies complete vertical integration[29].

3.5.3 Expansion of the Market

Now at a certain point in the future the population allowed to grow. The increase in demand increases sales for the outlet. However, due to diminishing returns at the local level, the franchise is not able to fully supply the extra output at the given level of s. This will lead to an increase in the incentive to shirk. Therefore the franchisor will have to either increase the level of monitoring or allow the agent's rent stream to increase. The rent stream is implicitly increased whenever the fixed component of

[29] Mathewson and Winter (*Supra note 6*) use the same definition of vertical integration.

the franchise fee is held constant in the presence of growing demand.

The increase in demand may create an incentive for the franchisor to install a second outlet in close proximity to the first franchise. At this point the franchisor must decide whether it is more profitable to convert the franchises back to corporate owned and operated outlets or let them remain as franchise outlets. There are two factors that the franchisor must consider in making the decision to convert a franchise back to a corporate store. The first is the costs of monitoring two outlets which exist in close proximity to each other. The second is how the two outlets will interact while operating under a franchise arrangement.

3.5.4 The monitoring problem with two outlets

As before, the franchisor must incur transportation costs in order to engage in monitoring; therefore the cost of monitoring the first outlet is $M(\phi, x_0)$. Now suppose a second outlet is located in the same market. Since the transportation costs must be incurred to monitor a single outlet at location x_0, they become sunk costs, thus the cost of monitoring the second store is $M(\phi, 0)$ (where $M(\phi, 0) < M(\phi, x_0)$). The marginal cost of monitoring function for each of the stores is illustrated in figure 3.3.

In figure 3.3 $M'(\phi, x_0)$ intersects the original marginal benefit of monitoring schedule, MB_{old}, at point A. The increase in demand shifts the marginal benefit schedule up to MB_{new}. The marginal cost of 100% monitoring of the first store is given by point E. If a second store is also located at distance x_0, then the marginal cost of monitoring the second store is given by point H. The marginal benefit of 100% monitoring is given by point F. If the distance F to E is greater than the height to point

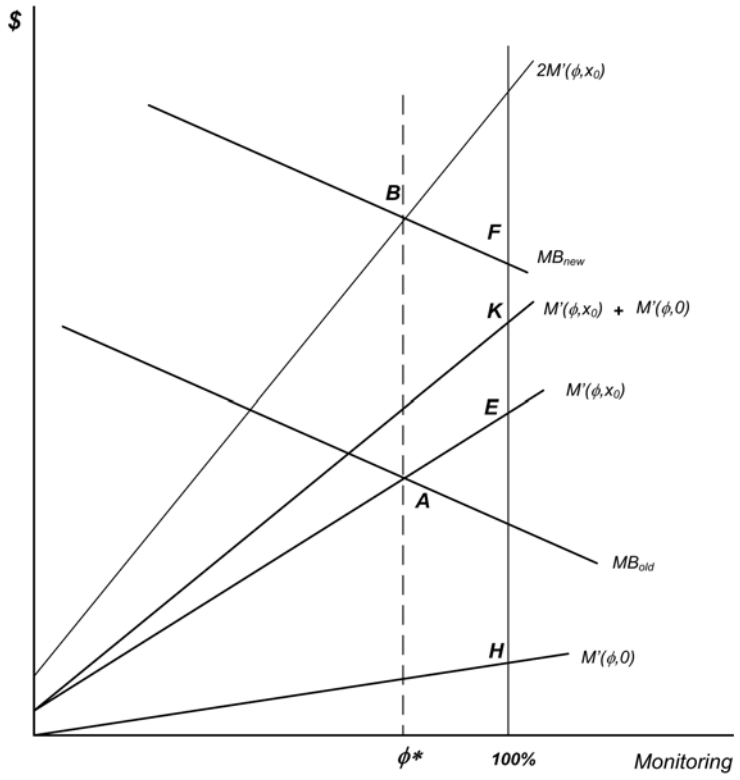


Figure 3.3: Expansion of the Market

H, then it will be worthwhile for the franchisor to convert the two stores to corporate outlets and engage in full monitoring. Regardless of the choice of contractual arrangement that the franchisor finally settles on, it is clear from figure 3.3 that the marginal cost of monitoring a second store is less than that of a single store in the same geographical area.

In addition to the *economies of scope* in monitoring costs described above, there exists a further potential reduction in monitoring when there is a second outlet. The franchisor can use information from one outlet to infer local demand conditions of the second outlet. For example if one store reports high sales in the same market that the other outlet reports low sales, the franchisor may be able to form a better prior about the likelihood that the second store is shirking on quality rather than suffering from a random drop in demand.

3.6 Conclusion

According to Francine Lafontaine (1992), "*Franchising offers a rare opportunity to assess theories concerning firms' contractual decisions*". Theoretic models that consider factors such as risk, moral hazard and capital accumulation offer explanations for the decision to enter into a franchise agreement, but say little regarding the specific terms of the contract. Most of these models focus on a single franchisor-franchisee pairings and magnitude of the franchise fee and royalty. The implication of these models is that differences in contracts should be a result of heterogeneous local markets. This would imply non-standardized franchisee fees and royalties - a result not supported by the empirical evidence.

Models that considered reverse moral hazard on the part of the franchisor address the coexistence of franchise and corporate outlets. Signaling, market saturation, and brand equity models offer explanations for contract-mixing; however, each of these three approaches have the franchisor adjusting the franchise fee with changes in the environment.

This paper has presented a simple model of a franchise contract. While capturing the essential elements of an incentive compatibility contract, the model is able to address some of the geographic issues inherent in franchise contracts. Specifically, the model focuses on the issue of the costs of monitoring to explain the contractual choices observed in franchise industry.

Two implications arise from the analysis. First, that in the presence of positive monitoring costs and incentives to shirk, increases in rent due to market growth may accrue to the agent rather than the franchisor. Second, the firm's choice of contract-mix and the decision to expand are mutually exclusive. When geographical considerations are taken into account, non-convexities in monitoring costs may arise that affect both the decision to expand and the decision to convert a franchise outlet to a corporate owned-store. While other models suggest that a change in market conditions resulting in greater economic rent would lead to higher franchise fees, this model demonstrates the reasons as to why the franchisor may potentially choose to convert the franchise to a corporate outlet. The decision to convert to a corporate outlet would be concurrent with the decision to increase the number of outlets, thereby capturing a greater portion of the additional rent due to the reduction in the relative cost of monitoring.

The model in this paper is limited to the set of franchise contracts where some input on the part of the franchisee is a major component of the final product. The model does not apply to all forms of franchising observed in the economy, but particularly franchise arrangements that are classified as business concept relationships. Such industries that experience large economies of scale in centralized production of the final product may find franchising simply an efficient method of delegating the responsibility of distribution.

3.7 References

- Barzel, Y. "Measurement Costs and the Organization of Markets," The Journal of Law and Economics , 25, April (1982) 27-48.

- Bhattacharyya, Sugato, and Lafontaine, Francine. "Double-Sided Moral Hazard and the Nature of Share Contracts." Rand J. Econ. 26 (Winter 1995): 761-81.

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

- Cheung, S. N. S. The Theory of Share Tenancy, with Special Application to Asian Agriculture and the First Phase of Taiwan Land Reform. Chicago: Univ. Chicago Press, 1969.

- Cheung, S.N.S. "The Contractual Nature of the Firm," The Journal of Law and Economics , 26 April (1983) 1-21.

- Demsetz, H. "The Exchange and Enforcement of Property Rights," Journal of Law and Economics, 7 Oct. (1964) 11-26.

- Gallini, N. T. and Lutz, N.A. "Dual Distribution in Franchising" J. Law, Econ., and Organization 8 (October 1992): 471-501.

- Goldberg, V.P. "Toward an Expanded Economic Theory of Contract," Journal of Economic Issues, 10, 1 March (1976) 45-61.

- Helmers, H.O., Davisson, C.N., and Taggart, H.F. Two Studies in Automobile Franchising , The University of Michigan Ann Arbour, Michigan.

- Hunt, S.D. "The Trend Toward Company-owned Units in Franchise Chains," Journal of Retailing , vol. 49, 2 Summer (1973) 3-13.

- Klein, B. "Borderlines in Law and Economics: Transaction Cost Determinants of 'Unfair' Contractual Arrangements," American Economic Review, 70, 2 May (1980) 356-362.

- Klein, B. and Leffler, K. "The Role of Market Forces in Assuring Contractual Performance," Journal of Political Economy, 89, 4 (1981) 615- 641.

- Klein, B. and Saft, L.F. "The Law and Economics of Franchise Tying Contracts," Journal of Law and Economics , 28 May (1985) 345-361

- Klein, B. and Murphy, K. "Vertical Restraints as Contract Enforcement Mechanisms," Journal of Law and Economics, Oct. (1988) 265-297

- Kroc, R. Grinding It Out: The Making of McDonald's Henry Regnery Co., Chicago, Illinois (1977).

- Lafontaine, Francine. "Agency Theory and Franchising: Some Empirical Results." Rand J. Econ. 23 (Summer 1992): 263-83. (a)

- Lafontaine, Francine; and Shaw Kathryn L. "The Dynamics of Franchise Contracting: Evidence from Panel Data" The Journal of Political Economy, Vol. 107, No. 5. (Oct., 1999), pp. 1041-1080.

- Martin, Robert E. "Franchising and Risk Management." A.E.R. 78 (December 1988): 954-68.

- Mathewson, F. and Winter, R. "The Economics of Franchise Contracts," The Journal of Law and Economics, Oct. (1985) 503-526.

- McAfee, R. Preston, and Schwartz, Marius. "Opportunism in Multilateral Vertical Contracting: Nondiscrimination, Exclusivity, and Uniformity." A.E.R. 84 (March 1994): 210-30.

- Minkler, A.P. "Why Firms Franchise: A Search Cost Theory" Journal of Theoretical and Institutional Economics (1991)

- Norton, Seth W. "Franchising, Brand Name Capital, and the Entrepreneurial Capacity Problem" Strategic Management Journal, Vol. 9, Special Issue: Strategy Content Research. (Summer, 1988), pp. 105-114.

- Ozanne, U.B. and Hunt, S.D. The Economic Effects of Franchising (Washington, D.C.; U.S. Government printing office, 1971).

- Rubin, P. "The Theory of the Firm and the Structure of the Franchise Contract," Journal of Law and Economics , 21 (1978) 223-233.

- Seltz, D.D. The Complete Handbook of Franchising Addison- Wesley Publishing Company Inc. (1982).

- Sen, Kabir C. "The Use of Initial fees and Royalties in Business-Format Franchising" Managerial and Decision Economics, Vol. 14, No. 2, Special Issue: Transactions Costs Economics. (Mar. – Apr., 1993), pp. 175-190.

- Simon, Carol J. "Franchising versus Ownership: a contracting explanation," *University of Chicago working paper* (1991)

- Stiglitz, Joseph E. "Incentives and Risk Sharing in Sharecropping." Rev.Econ. Studies 41 (April 1974): 219-55.

- Udell, G. "The Anatomy of the Franchise Contract," The Cornell Hotel and Restaurant Quarterly 13, no. 3 August (1972) 13-21.

- Vaughn, C.L. Franchising Lexington Books, Lexington Mass. (1974).

- Williamson. O.E. "Transaction Cost Economics: the Governance of Contractual Relations," The Journal of Law and Economics , 22, Oct. (1979) 223-261.

- Winter, R.A. "Vertical Control and Price versus Non-price Competition" The Quarterly Journal of Economics Vol. 108, No. 1 (Feb., 1993), pp. 61-76.