Studies of Talent Markets

by

Marko Terviö

M.Soc.Sc., University of Helsinki, 1997

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2003

Author .....
Department of Economics
May 2, 2003

Certified by .....
Bengt Holmström
Paul A. Samuelson Professor of Economics
Thesis Supervisor

Certified by .....
David Autor
Pentti J.K. Kouri Career Development Assistant Professor of Economics
Thesis Supervisor

Accepted by .....
Peter Temin
Elisha Gray II Professor of Economics
Chairman, Departmental Committee on Graduate Studies

Studies of Talent Markets

by

Marko Terviö

Submitted to the Department of Economics
on May 2, 2003, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This thesis consists of three studies of labor markets where differences in talent are associated with very large differences in income. The unifying theoretical feature is the view that the analysis of such labor markets should take into account the scarcity of jobs, which is a natural consequence of the combination of finite demand and positive production costs. In Chapter 1 we propose a model where an industry-specific talent can only be revealed on the job and publicly. Individual inability to commit to long-term contracts leaves firms with insufficient incentives to hire novices, inducing them to bid excessively for the pool of revealed talent instead. This causes the market to be plagued with too many mediocre workers, inefficiently low output levels, and excessive rents for the known high talents. We argue that high wages in professions such as entertainment and entrepreneurship may be explained by the nature of the talent revelation process in those markets, and suggest potential natural experiments for estimating the welfare loss and the excessive talent rents predicted by the model. Chapter 2 is an analysis of the labor market of CEOs. We present an assignment model of managers and firms of heterogeneous talent and scale, and show how the value of underlying ability differences can be distinguished from scale effects using the observed joint distribution of CEO pay and market value. The empirical results suggest that the observed size-pay relation in the US is mainly due to differences in firm characteristics rather than differences in managerial ability. Chapter 3 uses a combination of simple versions of the models of the first two chapters to analyze the role of transfer fees in professional sports. There workers are able to commit to long-term wage contracts, and a transfer fee is the price of a remaining contract. We show that the abolition of transfer fees would reallocate playing time towards older players and increase salaries by more than the current transfer fees. All clubs, including the bigger clubs that are the current net payers of transfer fees, would lose out in the reform.

Thesis Supervisor: Bengt Holmström
Title: Paul A. Samuelson Professor of Economics

Thesis Supervisor: David Autor
Title: Pentti J.K. Kouri Career Development Assistant Professor of Economics

Acknowledgments

The earliest piece of advice on my road to writing a Ph.D. thesis was given by a Chinese fortune cookie that I got from the MIT Economic Theory lunch. *"Your path is arduous but will be amply rewarding,"* read the snippet on my first official day as a thesis writer. The accuracy of this foresight has already been confirmed on both counts.

The guidance provided by my advisors has also been invaluable. Numerous discussions with Bengt Holmström and David Autor helped both with the immediate goal of the thesis and with my development as a researcher. Their respective viewpoints from contract theory and empirical labor economics proved to be complementary (and without wasteful overlap in reading recommendations). On many occasions a discussion would lead me to explore a point that I had left as a footnote and expand it into a section, and equally often, to compress a bloated section or a paragraph into a mere sentence. They both suffered, with good grace, through several drafts of the lead chapter on short notice before the job market deadline. I also benefited from many discussions with Abhijit Banerjee who was quick to see the core structure of any model, and who initially recommended looking more carefully into the transfer fee system, an exploration which ended up as its own chapter.

I am also thankful to many others who helped with the thesis, including Daron Acemoglu, Glenn Ellison, Frank Fisher, Bob Gibbons, Simon Johnson, Eric Van den Steen and Scott Stern. Substantial comments and positive peer effects were provided, among others, by Kenichi Amaya, Shawn Cole, Kevin Cowan, Jonathan Kearns, Mark Lewis, Tim Mueller, Stephanie Plancich and Thomas Philippon. In particular, the coffee club with Saku Aura and Emek Basker served as an instant testing ground for many random ideas. Lakshmi Iyer, cubicle-mate and a fellow loser of the 2002 student office space lottery, provided useful comments and good humor at the basement of E51 (aka the Dungeon) all the way through the crucial marketing stages of the theses.

Thanks go out also to my undergraduate advisors at the University of Helsinki, Seppo Honkapohja and Erkki Koskela, who helped me to make it to MIT and to obtain the necessary funding. That funding was provided by the Yrjö Jahnsson Foundation and the Finnish Cultural Foundation. I am indebted to both foundations, as well as to the wretched Finnish taxpayers who paid for my undergraduate (and earlier) education.

Finally, I dedicate this thesis to the memory of my grandfather Yrjö "Pappa" Terviö.

Contents

List of Figures

Summary

This thesis consists of three studies of labor markets where differences in talent are associated with very large differences in income. The unifying theoretical feature is the view that the analysis of such labor markets should take into account the scarcity of jobs, which is a natural consequence of the combination of finite demand and positive production costs.

In Chapter 1, a model of a labor market is proposed where the level of individual ability can only be revealed on the job and publicly. Individuals' inability to commit to long-term contracts leaves firms with insufficient incentives to hire novices of uncertain talent, causing them to bid excessively for the pool of revealed talent instead. This causes the market to be plagued with too many mediocre workers and inefficiently low output levels, while simultaneously raising the wages for high talents. This problem is expected to be most severe where prior information about talent is imprecise but revealed relatively quickly on the job, and where individual performance is highly observable and relevant for competing firms. We argue that high wages in professions such as entertainment, team sports, and entrepreneurship, may at least partly be explained by the nature of the talent revelation process in those markets. This explanation is distinct from superstar economics, scale effects, and human capital models. We suggest potential natural experiments for identifying and estimating the welfare loss and excessive talent rents predicted by the model.

Chapter 2 is a study of the labor market of CEOs. The predominant fact about executive pay is that large firms pay more to their CEOs. This may be partly due to the higher value of ability in larger firms, and partly to the matching of higher talent with larger scale. In a matching market where talents as well as managerial positions are scarce, talent rents are determined by the full distribution of talents and firms. An assignment model is presented in which the equilibrium division of rents is solved in a manner analogous to screening models. The capital level of firms is allowed to depend on the characteristics of the firm

and the CEO. Assuming that management technology is multiplicatively separable from production technology, the value of underlying ability differences can be distinguished from scale effects using the observed joint distribution of CEO pay and market value. Based on CompuStat data on the largest 1000 US companies in 1999 we estimate the value of managerial ability. We find that the economic value of scarce CEO ability is about $25-37 billion in 1999, of which the CEOs received $5 billion. This suggests that the observed size-pay relation is mainly due to differences in firm characteristics rather than differences in ability.

Chapter 3 is a study of the role of transfer fees in professional sports, where players can commit to binding long-term wage contracts. A transfer fee is the price of a remaining contract which a new club has to pay to the current holder of the contract. In the EU there has been recent pressure to end transfer fees as an illegal restriction on players' freedom, and possibly replacing them with some compensation for documented training costs. We present a model where the crucial role of transfer fees comes from the allocation of scarce playing time between players of varying levels of ability and potential. Abolition of transfer fees would reallocate playing time towards older players with less upside potential, eventually reducing the quality of players. We also show that the increase in player salaries can be expected to exceed the current transfer fees for each level of talent, so that all clubs—not just the current net recipients of transfer fees—would lose out in the reform.

Chapter 1

Mediocrity in Talent Markets

1.1 Introduction

This chapter presents a model of a talent market where industry-specific talent can only be revealed on the job and publicly. The two crucial features of the model are that individuals have finite lives and that output has finite demand. This results in a scarcity of both revealed talent and of job slots. The market price of output has an important role in determining wages: it must adjust to accommodate the hiring of novices, without which the industry would run out of workers. This gives a new twist to the old economic problem of joint production of output and information about worker quality. It is shown that when revelation of talent is relatively quick and observed performance relevant for competing employers, then there is too little exit from the industry and too many mediocre workers are employed. This leads not only to inefficiently low levels of talent and output in the industry but also to excessively high and skewed pay for top talents.

When individuals cannot commit to long-term wage contracts, the value of information accrues as rents to those who turn out to be high talents and see their wages bid up, while the firms that hired them do not get rewarded for the discovery. Unless individuals are able to pay for the opportunity to reveal their talent, firms will only take into account their expected talent for the near term and ignore the upside potential of previously untried individuals. Firms then prefer to hire someone who is known to be even slightly above the population average to hiring a novice of average talent in expectation. There is thus inefficiently low level of exit from the industry, especially by relatively inexperienced workers. If talent is revealed relatively quickly, then most of the active workforce may consist of "mediocre"

types who would exit the industry in the efficient solution. Instead they stay in the industry, producing output that crowds out entry by novices. The industry as a whole has what is in effect an up-or-out rule, but this rule is unduly lenient.[1]

If individuals were risk neutral and had deep pockets, then previously untried individuals would be able to pay for the chance to find out their talent level, up to the expected value of their lifetime talent rents. This would lead to an efficient solution, where even relatively high talents exit the industry if their job slots have higher social value in trying to discover even higher talent. With uncertain return to such talent, i.e. when talent rents mostly accrue to a minority of very successful individuals, the willingness of young individuals to pay for future rents can be much below the expected value. Credit constraints or even moderate levels of risk aversion can cause the market outcome to deviate considerably from full efficiency.

Perhaps surprisingly, the opportunity to save aggravates the inefficiency caused by a credit constraint. Saving by "has-been" individuals who perform well early in their career but who fall below population average in expected talent later allows them to outbid credit constrained novices. Their incentive to pay for job slots is the chance of more talent rents in the future: since talent is only revealed over time, the has-beens still retain some upside potential, albeit less than the novices. However, after sufficiently bad performance even the has-beens exit, regardless of their savings.

At one level this study just provides another explanation for high and skewed wages that is arguably plausible for many industries that appear to have high talent rents. As an explanation it is complementary to theories based on scale effects[2] (see e.g. Lucas 1978 and Rosen 1982) and superstar economics (Rosen 1981), even though less benign in the sense that it is associated with possibly dramatic inefficiencies. These papers are concerned with the efficient allocation of capital (and consumers) to known talent, whereas the focus here is on the discovery process of talent. For example, we might wonder why some alternative manager wouldn't be nearly as good as the current CEO with his exorbitant compensation, scale effects notwithstanding. This chapter shows how the supply of talent, as observed in the market, can be very scarce even when it is not so in the population; and more

[1] Efficient up-or-out rules are possible when information is match-specific, see e.g. O'Flaherty and Siow (1995). For a signalling perspective to up-or-out rules see Waldman (1990) or Kahn and Huberman (1988).

[2] With a scale effect or "scale-of-operations effect" differences in talent are accentuated when higher talent is matched with more productive complementary resources, such as capital.

importantly, revealed talent can be much more scarce than it need be due to the twin imperfection of spot contracts and credit constrained (or risk averse) individuals.

At another level, this study provides predictions about what kind of talents and industries could be expected to exhibit high and uneven wages. Inasmuch as a talent market fits the assumptions of the model, it can be expected to be flooded by too many mediocre workers. Such a market would react to certain exogenous changes, particularly to individual commitment ability, switching costs, credit constraints, or publicness of information, in ways that could be used to identify and quantify the inefficiencies described in the model. The benefit from relaxing constraints (to commitment ability or access to credit) comes through higher exit rates for young workers, a prediction at odds with standard training and human capital models. Higher exit rates would in turn show up as increased productivity, lower wages and decreased wage dispersion.

Talent is equated with level of output in this study. Jobs within an industry are homogeneous, as if all workers operated identical "machines." To say that one individual is twice as talented as another means that he produces twice as much output (possibly in expectation, or in quality-adjusted "hedonic" units). The economic value of talent is endogenous because it depends on the market price of output. Under this definition of talent it is meaningful to consider a thought experiment where the distribution of talent is the same in two industries.

Several commonly studied features of labor markets are assumed away in this chapter. There is no on-the-job training or learning-by-doing, so experience per se is not economically valuable. Neither are there hiring or firing costs, nor any organizational structure to speak of. Information is symmetric at all times: there are no effort problems, career concerns, or adverse selection. The homogeneity of job slots rules out any problems with job assignment within the industry.

For a talent market to be well described by this model, there should be substantial uncertainty about the talent levels of inexperienced individuals. Success in school or performance in other industries should not be a very accurate predictor of differences in talent among inexperienced individuals, even though they may be useful as pass/fail type filters in choosing the potential entrants. This uncertainty about talent is known in the entertainment industry as the "nobody knows" property (Caves 2000). It is exceedingly difficult to forecast the success of individuals in the entertainment industry before letting the public

experience the finished product. At the same time, the queuing for entry-level positions and auditions is suggestive of a credit constraint. The chance to show one's talent can turn unknown artists into superstars virtually overnight.

The inefficiency described here could in principle be identified given a suitable natural experiment. An exogenous change in individuals' commitment ability would be ideal. For example, the end of the studio system in the motion picture industry in the 1940s is a change that the model predicts would lead to rehiring of mediocre talent. Under that system young actors were able to commit to long-term contracts with motion picture companies. Available stylized facts of decreased revenue and output, as well as the casual evidence of increased wages, are consistent with the predictions of the model. However, contemporaneous changes, in particular the advent of television, make it difficult to draw strong conclusions.

The joint production problem of output and information about worker quality has been well understood since Johnson (1978) and Jovanovic (1979). The social planner's solution in this type of problems draws on the "bandit" literature, see e.g. the treatise by Gittins (1989). Miller (1984) uses the bandit approach in a multi-sector setting. MacDonald (1988) presents a stochastic version of Rosen's superstars model, where superstars are selected based on earlier performance. These papers solve for the efficient equilibrium; the focus here is on how the market handles the discovery of talent under constraints to credit and contracting. In this way the model is analogous to setups where firms should give training in general skills but don't have sufficient incentives, due to the same imperfections as here. This literature uses additional imperfections, typically asymmetric information (proposed by Greenwald 1986), to give firms incentives to train, see e.g. Acemoglu and Pischke (1998).

The plan of the chapter is as follows. In Section 1.2, a numerical example is used to illustrate the basic ingredients of the model. Section 1.3 presents the basic model of a talent market, with the simplest possible revelation process: individual talent is initially unknown, and then becomes public knowledge after one period on the job. Mediocrity and the loss associated with inefficient hiring are defined in an empirically quantifiable way. Section 1.4 extends the model to many periods, with talent revealed gradually over time. The problem of insufficient exit is shown to get worse when individuals can save. Section 1.5 discusses the relevance of the findings for real-world talent markets, and suggests possible natural experiments to identify and quantify the welfare cost of mediocrity. Section 1.6 concludes the chapter.

1.2 Example: A Simple Talent Market

Consider a competitive widget industry that combines workers with capital (machines for making widgets). There is free entry by firms that each need one worker to operate one machine that has rental cost of $4 million. All widgets are identical, and the number of widgets that a firm produces depends solely on the talent of its worker (later in the chapter it will be more natural to interpret talent as affecting quality, and the market price as being for hedonic "quality-adjusted" units of output). Industry output faces a downward-sloping demand curve, and firms take the market price as given. There is an unlimited supply of potential workers with an outside wage of zero. A novice is equally likely to produce anywhere between zero and one hundred widgets.[3] The talent of a novice widgetmaker is unknown (also to himself), but becomes public knowledge after one period of work. Careers are finite and last at most 16 periods.[4] Finally, workers cannot commit to decline higher outside wage offers in the future.

Listing the assumptions

1. There is free entry by profit-maximizing firms, which each employ one worker and incur a fixed cost of $4 million per period.

2. The number of widgets produced is equal to the talent of the worker.

3. There is a non-binding supply of workers of unknown talent, willing to work at the outside wage of zero, and talent is distributed uniformly in $[0, 100]$, measured in number of widgets produced per period.

4. A "novice" worker's talent becomes public knowledge after one period of work. He can then go on to work at most another 15 periods as a "veteran."

5. Individuals cannot commit to long-term contracts.

6. The number of firms is "large" enough, so that firms take the market price as given and there is no aggregate uncertainty.

7. There is no discounting.

How does this talent market work? That depends crucially on whether aspiring craftsmen can pay for the opportunity to make their first batch of widgets. There are two extreme cases to consider. In the first, individuals are constrained to take a non-negative wage. This

[3] For example, the machine could have a capacity for one hundred widgets per period, and talent could then correspond to the proportion of successfully completed widgets.

[4] All numbers in this example are chosen for convenience.

is the inefficient, but at the same time also the more straightforward case. In the second case individuals are risk neutral and not credit-constrained. As is intuitive, due to the absence of imperfections, this is the efficient benchmark.

The purpose of the example is to compare the distribution of talent and wages in the industry under these two cases. Only the steady state is considered, where the number of entering and exiting workers is constant over time.

Constrained Individuals

In this case all workers who turn out to be above the population average, i.e. those who were able to make 50 widgets or more, will keep making widgets until they retire. These veterans create more revenue than a novice in expectation, so they can always outcompete them for a job in the widget industry.

The market price of widgets must be such that novice-hiring firms break even. Since potential novices are not scarce, they will always be paid zero. A novice is expected to make 50 widgets, so a price of (\$4 million)/(50 widgets)= \$80000 (\$80K) per widget is needed to cover the capital cost. At this price there is no entry or exit of firms from the industry.

Veteran workers are always scarce. Due to free entry, firms cannot make positive profits and will bid up the wages of veteran workers, who get the difference between their revenue-generating capacity and that of a novice as a Ricardian rent. In particular, the highest type makes 50 more widgets than a novice or an average type. Therefore at the price of \$80K per widget, top veterans get $50 \times \$80K = \$4$ million per period. The average wage of veterans is \$2 million (since talent is uniformly distributed).

Since the production cost per worker is fixed, the efficiency at which the demand for widgets is satisfied depends solely on the average talent of workers in the industry. The average output by veterans is 75 widgets; the average for the whole industry must be lower since it includes the novices (it is in fact 72).[5] A novice has a fifty-fifty chance of being retained in the industry, in which case he will make in expectation the average veteran wage of \$2 million for 15 periods; hence the expected lifetime rents are $0.5 \times 15 \times \$2$ million= \$15 million.

[5]The formula that relates the fraction of novices to the rehiring threshold and the length of career is derived in the next section.

Unconstrained Individuals

Now suppose that aspiring widgetmakers are risk neutral and have access to unconstrained credit. They are then willing to bid for the opportunity to make their first batch of widgets, up to the expected value of future talent rents. The inability to commit to long-term contracts does not cause any problems when individuals can in effect buy the firm. I will now show that this will increase the exit/retention threshold and the average talent of workers in the industry up to the efficient level, while dramatically decreasing the talent rents.

Start by simply assuming that novices are offering \$1.5 million to firms for the chance to work (we will see shortly that this is in fact the unique equilibrium). Then at the widget price $\$P$, a novice-hiring firm will in expectation generate $50 \times \$P$ in revenue, and have a net cost of \$2.5 million. The net cost subtracts what is in effect a negative novice wage of \$1.5 million from the capital cost of \$4 million. For firms to break even, the equilibrium price of widgets must then be $\$P = (\$2.5 \text{ million})/(50 \text{ widgets}) = \$50K/\text{widget}$.

When novices pay to work, then veterans of average talent will not be hired into the industry. They have no incentive to pay for a job, because they have no chance of getting higher wages in the future. The lowest type veteran to work will do so at the outside wage of zero.

The lowest types to stay in the industry, i.e. the threshold types, are those making 80 widgets per period. They generate enough more revenue than novices in expectation to just offset the novice payment of \$1.5 million. To see this, notice that at the price $\$50K$ per widget the additional 30 widgets that they make are worth exactly \$1.5 million.

Veterans who are better than the threshold type (i.e. the 80-widget type) get rents, again by their advantage over the threshold type. For example, the highest type makes 20 widgets more than the threshold type who is available at zero wage; therefore at the widget price $\$50K$ the very best craftsmen get a rent of $20 \times \$50K = \$1$ million per period. The average wage of veterans is \$0.5 million (again by the uniformity assumption).

Finally, to show that this is the equilibrium, calculate the expected lifetime rents. A novice has a 20% chance of turning out to be above the 80 widget threshold, in which case his expected wage is the average veteran wage of \$0.5 million for the last 15 periods. Expected lifetime rents are then $0.2 \times 15 \times \$0.5$ million= \$1.5 million, which was the assumption we

started from. This is also the unique equilibrium, because higher offers by novices increase the exit threshold and thus decrease the expected rents.

The average output of veteran workers is 90 widgets (because veteran talent is uniform between 80 and 100). The industry average is lower, because some workers are novices; in fact it must be exactly 80 widgets per worker. That the optimal (i.e. maximal) average talent level of workers is the same as the optimal exit threshold is a general result (in the absence of discounting). Intuitively, if at the optimum some level of talent gets discarded from the industry then it must be pulling down the industry average, while a talent that is retained must be increasing it.

Table 1. Summary of the example.

	Constrained	Unconstrained
Output price	$\$80K$	$\$50K$
Threshold talent[6]	50 widgets	80 widgets
Average talent	72 widgets	80 widgets
Top wages	\$4 million	\$1 million

Comparison

When novices cannot pay the expected value of future talent rents, then two things happen. First, the exit threshold in the industry is too low. As a result, many job slots are taken over by mediocrities who reduce average talent in the industry, compared to if their job slots were used to discover new talents. Here the workers who make between 50 and 80 widgets per period are mediocrities in this sense; in fact most workers in the industry fall under this category. Second, the rents to talent are higher; here the top wage goes up from \$1 million to \$4 million. The talent rents accrue to the advantage in output that veterans have over the threshold type, so a reduction in the threshold increases the rents of all retained types. The inability of novices to pay for the job increases the price of output, because it must be high enough to cover the cost of production at novice-hiring firms. This increased price further magnifies the rents to retained talent.

[6]The exit rate of novices is threshold divided by 100.

1.3 The Basic Model

This section introduces the basic model of a competitive talent market. The main interest is in comparing the distribution of talent, wages and tenure in the industry with and without the ability of individuals to pay to enter the industry.

Assumptions

1. Each firm employs one individual per period, has production cost c, and output equal to the talent of the worker, θ.

2. There is an unlimited supply of individuals with unknown talent, willing to work at outside wage $\underline{w}$.

3. An individual's talent level becomes public after one period of work in the industry. He can then work in the industry up to T more periods ($1 + T$ periods in total).

4. Talent is drawn from a distribution with a continuous and strictly increasing CDF, $F(\theta)$, with support $[\theta_{\min}, \theta_{\max}]$.

5. There is no discounting. Firms are infinitely lived and maximize average per-period profits.

6. There is free entry by firms. The number of firms I is treated as a continuous variable (measure).

7. Industry output faces a downward-sloping demand function $Q^d(P)$.

The first assumption describes the technology. The firms are identical; all differences in output are caused by the talent of the worker. In other words, the level of talent *is* the level of output.

Assumptions 2 and 3 describe the information structure. That all uncertainty about talent is resolved after one period of work is a simplification of the idea that information about the talent of a novice is much less precise than that of experienced individuals. Information is symmetric at all points in time: firms (as well as individuals themselves) view all novices as identical, so they are expected to have the mean talent level $\bar{\theta}$. After one period of work, an individual's output (i.e. his talent) becomes public.

Assumptions 4 and 5 are not essential and are made purely for technical convenience. Assumption 6 results in a competitive industry that is "large" in the sense that there is no uncertainty about the realization of the distribution of talent. Firms take the price of output as given, and free entry makes sure that expected profits are zero.

The additional assumptions which leads to inefficiency on this talent market are:

A. Individuals cannot commit to long-term contracts.

B. Individuals cannot take a wage below $\underline{w}$, and cannot borrow against future earnings.

In terms of missing markets, the assumptions are that individuals can neither sell their future labor nor insure their unknown talent. The case with both two assumptions will be referred to as the case of "constrained individuals." Assumption B approximates the idea that the ability of novices to pay for future talent rents is small compared to their expected value (this could also be due to risk aversion). In the absence of assumption B individuals are assumed to be risk neutral and have access to unconstrained credit. The case with assumption A but without B will be referred to as the case of "unconstrained individuals."

The absence of either assumption A or B allows the industry to operate efficiently. Without A, i.e. with unhindered contracting, the equilibrium is socially efficient. Since novices are not scarce, they sign up to lifelong contracts to work at outside wage $\underline{w}$, while firms retain the right to fire the worker without further compensation. In this case firms would choose the efficient hiring/firing policy.

Preliminaries

The equilibrating variable here is the exit threshold θ^*; it will be shown later that the measure of jobs I will be determined mechanically given the threshold. Those who turn out to have a talent level below the threshold leave the industry after just one period. Vacancies left by novices who were not good enough to make the grade and by those retiring must be filled by new novices. In this preliminary section I derive the relation of the exit threshold, the equilibrium fraction of novices, and the average level of talent in the industry. With a given exit threshold, this is just a matter of equating the flows of entry and exit.

Denote the fraction of novices by i. When dealing with the distribution of talent, we can think of the industry as consisting of a unit mass of jobs without a loss of generality. Consider only the steady state, where i is constant over time. Each period, the talents of i new workers are revealed, and of these a fraction $F(\theta^*)$ exit. The remaining $1-i$ jobs

in the industry must be held by veterans; of these the oldest cohort, a fraction $\frac{1}{T}$ of all veterans, retires each period. Equating exit and entry yields

$$\begin{aligned} iF(\theta^*) + \frac{1}{T}(1-i) = i \Longrightarrow \\ i(\theta^*) &= \frac{1}{1+T(1-F(\theta^*))}. \end{aligned} \tag{1.1}$$

The distribution of workers by tenure (experience) in the industry is then $i(\theta^*)$ for $t = 1$ and $\frac{1}{T}(1 - i(\theta^*))$ for $t = 2, \ldots, T+1$.

The exit threshold further determines the average talent of workers in the industry. Denote the average talent in the industry by

$$A \equiv i\bar{\theta} + (1-i)\mathrm{E}[\theta|\theta > \theta^*]. \tag{1.2}$$

Clearly A will be above the population average $\bar{\theta}$, because types above the threshold will work for longer than the below-threshold types. Only if there were no filtering at all, would the industry average be equal to the population average. Substituting the equilibrium fraction of novices (1.1) into (1.2) yields the industry average as a function of the exit threshold.

$$A(\theta^*) = \frac{1}{1+T(1-F(\theta^*))}\bar{\theta} + \frac{T(1-F(\theta^*))}{1+T(1-F(\theta^*))}\mathrm{E}[\theta|\theta > \theta^*] \tag{1.3}$$

Note that, regardless of the number of jobs, the total value of output is proportional to the average level of talent in the industry.

Social Planner's Problem

Consider the problem of maximizing social surplus

$$S(I, \theta^*) = \int_0^{IA(\theta^*)} P^d(q)\mathrm{d}q - I(\underline{w} + c), \tag{1.4}$$

where $P^d(q)$ is the inverse of the demand function $Q^d(P)$ and I the level of employment (measure of jobs). The social surplus is the consumer surplus from total output, i.e. employment times average talent, minus the opportunity cost of the factors of production. The problems of choosing the efficient exit threshold and the socially optimal level of employment

have separate first-order conditions. First of all, the threshold θ^* should be chosen to maximize the average level of talent A in the industry. This will minimize average costs, because cost per job is constant $\underline{w} + c$. Second, the level of employment should be chosen to equate total output with demand at the minimized average cost, so that $P^d(IA) = (\underline{w} + c)/A$.

The choice of the exit threshold is essentially the choice of what fraction of jobs should be allocated to novices; it does not matter how large the industry is. The industry as a whole has constant returns to scale: to double the output, the amount of novices hired and total costs would both be doubled; this would (eventually) double the number of veterans as well.

To maximize the average talent (1.3) take the first order condition.

$$\begin{aligned}
& \frac{\partial}{\partial \theta^*} A(\theta^*) \;=\; \frac{\partial}{\partial \theta^*} \left(\frac{1}{1 + T\left(1 - F\left(\theta^*\right)\right)} \left\{ \bar{\theta} + T \int_{\theta^*}^{\theta_{\max}} a f(a) \mathrm{d}a \right\} \right) \;= 0 \\
\Longrightarrow \quad & \frac{T\, f(\theta^*)}{\left(1 + T\left(1 - F\left(\theta^*\right)\right)\right)^2} \left\{ \bar{\theta} + T\left(1 - F(\theta^*)\right) \mathrm{E}[\theta | \theta > \theta^*] \right\} \\
& \qquad\qquad - \frac{T\, \theta^* f(\theta^*)}{1 + T\left(1 - F\left(\theta^*\right)\right)} \;=\; 0 \\
\Longrightarrow \quad & \bar{\theta} + T\left(1 - F(\theta^*)\right) \mathrm{E}[\theta | \theta > \theta^*] \;=\; \theta^* \left(1 + T\left(1 - F\left(\theta^*\right)\right)\right) \qquad (1.5)
\end{aligned}$$

The first order condition (1.5) can be rearranged to yield the following condition:

$$\theta^* - \bar{\theta} = T\left(1 - F(\theta^*)\right) \left(\mathrm{E}[\theta | \theta > \theta^*] - \theta^* \right). \qquad (1.6)$$

To interpret (1.6), think of the decision to hire a novice over a veteran of above-average talent as an investment. The LHS gives the immediate loss in expected output from hiring a novice instead of the threshold veteran. The RHS shows the expected future gain, assuming that θ^* is also kept as the rehiring threshold in the future.[7]

It is useful to notice that the maximizer of (1.3) is also its unique fixed point in the support of θ.

Proposition 1 $\max_\theta A(\theta) = \arg\max_\theta A(\theta) > \bar{\theta}$.

Proof First, to see that the solution to (1.6) is a fixed point of A, solve the linear term for θ^* and then divide both sides by $1 + T\left(1 - F\left(\theta^*\right)\right)$. This reproduces the objective

[7] With discounting the optimal amount of this investment (into experimentation) would be lower, resulting in a lower rehiring threshold.

function (1.3). Second, to see that the solution exists, is unique, and strictly greater than $\bar{\theta}$, notice that the LHS of (1.6) is strictly increasing, and equal to zero at $\theta^* = \bar{\theta}$. The RHS is decreasing, with slope $-T(1 - F(\theta^*))$; it starts from positive $T(\bar{\theta} - \theta_{\min})$ at $\theta^* = \theta_{\min}$ and reaches zero at $\theta^* = \theta_{\max}$. □

In other words, the optimal exit threshold is also the maximum attainable average level of talent in the industry: $A^* = A(A^*)$. Intuitively, discarding a worker above the optimal threshold must decrease the average, as must retaining a worker below the threshold. Denote this fixed point henceforth by A^*. The optimal level of employment equates supply IA^* with demand at average cost $(\underline{w} + c)/A^*$, so $I^* \equiv \frac{1}{A^*} Q^d(\frac{\underline{w}+c}{A^*})$.

Definition 1 **Mediocre types***: $\theta \in (\bar{\theta}, A^*)$. These are the talent levels above the population average, but below the maximal attainable industry average.*

In other words, "mediocrities" are people who are retained in the equilibrium with a credit-constraint, but are not be retained in the social optimum.

Market Equilibrium

Like the social planner's allocation, market equilibrium can also essentially be described by the exit threshold θ^*. The level of equilibrium threshold will depend on whether individuals are credit-constrained or not, but for a given threshold we can now deduce the equilibrium wages, output price, and employment. In either case, the individual inability to commit to long-term contracts means that wages are determined on a spot market. Equilibrium wages must therefore keep firms indifferent between hiring any worker in the industry for the next period. This means that (known) differences in talent translate into corresponding differences in wages, and into Ricardian rents for inframarginal talents. At the same time, the price of output must adjust to allow the hiring of novices into the industry, while free entry keeps profits at zero.

Proposition 2 $w(\theta^*) = \underline{w}$.

Proof Since veterans have no future payoffs to think about, their decision to stay depends solely on whether the wage they can get inside the industry is more than the outside wage. The lowest type veteran to work in the industry is indifferent and therefore paid exactly the outside wage. □

Proposition 3 *Given an equilibrium exit threshold θ^*, the price of output is $P = (\underline{w} + c)/\theta^*$.*

Proof Due to free entry firms must make zero profits. In particular, a firm employing a veteran of the threshold type θ^* gets revenue $P\theta^*$ and has costs $\underline{w} + c$. The equilibrium price sets these equal. □

The combination of free entry by firms and a binding outside wage for veterans of threshold type pins down the price of output.[8]

Proposition 4 *Given an equilibrium exit threshold θ^*, wages are*

$$w(\theta) = (\underline{w} + c)\left(\frac{\theta}{\theta^*} - 1\right) + \underline{w}. \tag{1.7}$$

Proof For firms to be indifferent between a threshold type θ^* and any other talent θ, the difference in wages must just offset the difference in revenue generated. Hence for any θ

$$w(\theta) - w(\theta^*) = P(\theta - \theta^*) = \left(\frac{\underline{w} + c}{\theta^*}\right)(\theta - \theta^*). \tag{1.8}$$

Combining this with Proposition 2 completes the proof. □

Proposition 5 *Given an equilibrium threshold θ^*, employment is*

$$I(\theta^*) = \frac{1}{A(\theta^*)} Q^d\left(\frac{\underline{w} + c}{\theta^*}\right). \tag{1.9}$$

Proof With threshold θ^* and employment I the supply of output is $IA(\theta^*)$, (measure of workers times average talent, i.e. average output). Set the supply equal to demand $Q^d(P)$, substituting in the output price from Proposition 3, and solve for I. □

Constrained Individuals Notice that Proposition 4 must also apply to novice wages: from the firms' point of view, novices are just like veterans with talent equal to population average $\bar{\theta}$. In the constrained case, novices must always get paid exactly $\underline{w}$. They cannot get more, since they are not scarce, and they cannot subsist on less by assumption. Everyone who is revealed to be better than the population average will make rents as a veteran and

[8] If market demand had a choke price below $(\underline{w} + c)/\theta^*$, then this industry could not operate; clearly this is not the interesting case.

has no reason to exit. The wages are then given by equation (1.7) with the exit threshold at $\theta^* = \bar{\theta}$. The population average is an inefficiently low rehiring threshold: average talent in the industry is not maximized at $A(\bar{\theta}) < A^*$ (by Proposition 1).

If a firm hires a novice that turns out to be above average, his wage will be bid up by other firms. Therefore firms only care about the expected ability of a worker for the current period. They fail to take into account the upside potential of young individuals, who themselves are not able to pay for it due to the credit constraint.

Definition 2 The curse of mediocrity *is* $A^* - A(\bar{\theta})$.

The curse of mediocrity is the efficiency loss in terms of output per worker. It could be observed in the data if there were a shift from one case to the other, for example if the commitment ability of individuals were suddenly removed, or if novices were to lose access to credit. The curse of mediocrity is also associated with a lower exit rate for novices, by $F(A^*) - F(\bar{\theta})$, and a lower proportion of novices in the workforce, by $i(A^*) - i(\bar{\theta})$. In reality, changes to imperfections are unlikely to be of the all-or-nothing type, but the direction of the effect of more limited changes should be clear. Some potential episodes for measuring the curse of mediocrity are discussed in section 1.5.

Unconstrained Individuals If individuals are risk neutral and have sufficient funds to pay for the expected value of talent rents, then the only remaining "imperfection" is the inability to commit to long-term contracts. That alone is not a problem because now individuals would have no trouble in "buying the firm." This must result in the same efficient hiring threshold and level of output as was seen in the social planner's solution. In addition, the distribution of wages is now also determined.[9] Wages are given by equation (1.7), with the exit threshold at $\theta^* = A^*$.

With constrained individuals the problem was that firms did not have the right incentives to hire novices, and too many mediocre veterans were employed as a result. By offering to pay for the chance to work in the industry, unconstrained novices provide firms with the right hiring and firing incentives. Being risk neutral, they will pay the full expected value of future talent rents. This makes firms prefer novices to mediocre veterans who have no

[9] While risk neutral workers are indifferent between any gambles of the same expected value, it seems reasonable to use the solution that is the unique limit of vanishingly small risk aversion. Note that there are no match-specific rents and no scope for bargaining in the model.

incentives to offer such payments—whatever wage they could get in one period, they will get for the rest of their career.

The efficient exit threshold can also be derived by solving for the market equilibrium in the unconstrained case. In equilibrium, each novice and firm takes the output price and the exit threshold as given. Since veterans of threshold type are available at the outside wage, novices have to pay $P(\theta^* - \bar{\theta})$ for their first period job slot.[10] This payment exactly compensates a novice-hiring firm for the one-period revenue loss that it takes compared to if it hired the threshold type. At the same time, the novice payment must be equal to expected lifetime rents: with threshold θ^*, a novice has a probability $1 - F(\theta^*)$ of being retained, in which case he gets the excess revenue $P(\theta - \theta^*)$ as a rent on each of the T remaining periods of his career. This equality is the market equilibrium condition:

$$P(\theta^* - \bar{\theta}) = (1 - F(\theta^*))\, TP\, (\mathrm{E}[\theta|\theta > \theta^*] - \theta^*)\,. \tag{1.10}$$

The market price P factors out of the equilibrium condition, which is therefore just the first-order condition (1.6) in the social planner's problem and yields the optimal threshold A^* as a solution. Recalling proposition 3, the price of output is therefore equal to average cost $P^* = (\underline{w} + c)\,/A^*$.

It is intuitive that the payment by risk neutral individuals with unconstrained credit raises the exit threshold to the efficient level: no imperfection, no problem. In the absence of a credit-constraint, less than infinite risk aversion would make novices willing to pay something for the job and thus make them more desirable to hire than revealed average types. This would raise the exit threshold above the population mean and reduce the inefficiency. This effect could be quite limited, because future income is risky (and would be more so with a right-skewed distribution of talent). Even with unconstrained borrowing, young individuals must take into account that they have to pay back the loan even if they will not be retained.

To summarize, the main effect of the constraint on novice ability to pay is that the standard of performance required for an individual to continue working in the industry is too low. The proportion of young workers is too low, while older workers are not as talented as they would be if the job slots were used more efficiently in discovering talent.

[10] So novice wage is $\underline{w} - P(\theta^* - \bar{\theta})$, which could be negative.

This coincides with higher talent rents, the level of which depends on two factors. First, rents accrue to the difference in units of output that an individual makes compared to the threshold type; this is always higher in the constrained case since the exit threshold is lower. Second, the value of this advantage is proportional to the price of output, which is higher in the constrained case: when novices cannot pay for the opportunity to work, it takes a higher output price to allow novice-hiring firms to break even. Since output price is higher, total industry output must be lower in equilibrium. The effect on employment is ambiguous without further assumptions and is analyzed next.

Comparative Statics

Elasticity of Demand and Employment It was shown before that the price of output must be higher in the case with credit-constrained novices. Total output must therefore be lower, unless demand is completely inelastic. However, the effect of inefficient hiring on employment (i.e. measure of jobs and firms) is ambiguous in general and depends on the demand function.

Proposition 6 *Under a constant elasticity of demand η, employment is higher in the credit constrained case if and only if* $\eta < \log\left(A^* - A(\bar{\theta})\right) - \log\left(A^* - \bar{\theta}\right)$.

Proof The condition for the equality of supply and demand is $IA(\theta^*) = \gamma P^{-\eta}$, where $\gamma > 0$ is a parameter. Inserting the equilibrium price $P = (\underline{w} + c)/\theta^*$ from Proposition 3, we see that employment as a function of the exit threshold is $I(\theta^*) = \frac{\gamma}{A(\theta^*)}(\frac{\underline{w}+c}{\theta^*})^{-\eta}$. The proposition follows from solving the inequality $I(\bar{\theta}) > I(A^*)$ for η. □

Intuitively, since the average talent of workers is lower in the constrained case, then more workers are needed to produce the same output. If demand is sufficiently inelastic, then an inefficiently low exit threshold coincides with inefficiently high employment in the industry.[11]

The welfare loss caused by the curse of mediocrity depends on consumer preferences. Unless demand is completely elastic, then some of the consumer surplus gets transferred to increased talent rents through the higher output price. For high elasticity of demand the loss is small: consumers shift their expenses towards other products without much loss in

[11] This is in contrast to Frank and Cook (1995), who argue that talent markets attract too many hopefuls. In their story, only the highest talent ends up contributing to output.

consumer surplus, and workers and productive resources shift to other sectors. If demand is very inelastic, then the welfare loss is worsened by the increase in jobs, because they waste the opportunity cost of the resources being attracted into the sector. Most of the social loss from the curse of mediocrity could come in the form of excess employment in the profession in question.

One implication of the credit constraint case is that a monopoly would serve the consumers better than a competitive industry if demand is sufficiently elastic. Suppose that the industry could merge into one firm, which could act as a monopsonist on the talent market. It would then have the incentives to enforce the socially optimal exit threshold. By being able to maximize the average level of talent in the industry, a monopolist would therefore also minimize the average cost of production. With sufficiently elastic demand, this would be enough for the monopoly price to be below the competitive price. For example, with constant elasticity of demand η, we know that a profit-maximizing monopoly marks up its price by a factor of $\eta/(\eta-1)$. Since monopolists' average cost is $(\underline{w}+c)/A^*$, and the competitive output price is $(\underline{w}+c)/\bar{\theta}$, it follows that the output price would be lower under a monopolist if $\eta > A^*/\left(A^*-\bar{\theta}\right)$.[12]

Production Costs Consider two otherwise identical talent markets with different cost of production. Individuals in the high-cost industry will be earning higher rents, by the wage equation (1.7), because the slope includes the cost c. For unconstrained novices, higher costs would only make the wages more risky, and increase the required novice payment. In the constrained case, a higher cost increases expected rents. Higher production costs (as well as higher outside wage) increase rents by increasing the output price, which determines the dollar value of talent differences.

The inefficiently low talent levels are not caused by a high cost of production, even though that may seem intuitive at first. Some positive production costs, whether from opportunity cost of the worker or other inputs, are needed for equilibrium price to be positive in the model. However, the distribution of talent in the industry is independent of production cost. For example, consider the extreme case $c = 0$, so that the industry consists of self-employed entrepreneurs. Their sole cost of production is the opportunity

[12]This does not hinge on there being a horizontal supply curve of labor. Since the monopolist can make consumers strictly better off, the results would also go through with some monopsonist's distortion to labor demand.

cost of labor, i.e. lost wages in some other industry. The problem is not that individuals cannot pay for the cost of production that would reveal their talent level. The problem is that if they do so and turn out to be mediocre, they will stick around in the industry. The masses of mediocre individuals supply output that keeps the price at a level that deters more entry by novices. In the equilibrium with unconstrained individuals, the entrants would suffer an expected first-period loss equal to the expected lifetime rents. This would drive down the output price to the efficient level, inducing mediocre veterans to exit the industry.

Speed of Revelation A higher number of "veteran periods" T corresponds to quicker revelation of talent. The parameter T can be interpreted as the ratio of veteran time to novice time, with the latter normalized at one. For any given exit threshold, a longer veteran time means that there must be fewer novices in the industry. The average level of talent is therefore increased for the mechanical reason that the same set of discarded talents spends less time in the industry. In the constrained case this is the only benefit: the exit threshold is always $\bar{\theta}$; with higher T the average talent in the industry gets closer to $\mathrm{E}[\theta|\theta > \bar{\theta}]$ because the below-average types get filtered out faster.[13]

When revealed types stay around for longer, the social return to the investment of hiring a novice is higher. In the efficient solution, a quicker revelation therefore increases the exit threshold. This can also be seen by using the fixed point result (1) in reverse: since the optimal rehiring threshold is always equal to the maximal attainable average talent level, and the latter is increasing in T, then so must be the former.[14] At the limit where talent is revealed instantaneously there would be no need to accept anyone except the highest possible types. For (1.6) to hold at each T, the optimum $A^*(T)$ must go from $\bar{\theta}$ to $\theta_{\max}$ as T varies from 0 to ∞.[15] Inserting these into (1.1) shows that the fraction of novices goes from 1 to 0.

[13] There is a discontinuity at the limiting case of $T = \infty$, because this corresponds to instant revelation (perfect information) case where only the highest types $\theta_{\max}$ would ever be hired.

[14] Totally differentiating the equilibrium condition (1.6) with respect to θ^* and T, and using the envelope theorem, yields

$$\frac{\partial A^*(T)}{\partial T} = -\frac{(1 - F(A^*))}{1 + T\,(1 - F(A^*))}\,\{\mathrm{E}[\theta|\theta > A^*] - A^*\} < 0 \tag{1.11}$$

[15] At the limit $T = 0$ the solution is not defined: there can be no meaningful hiring policy since no information about talent is ever revealed.

When revelation is slow and the veteran time short, filtering is not as crucial and the average worker talent cannot be much improved compared to the population average. When revelation is quick, then veteran time is long and the optimal filtering picky. The contrast between the efficient and inefficient case is higher when the revelation is quicker, whereas for slower revelation the problem of mediocrity is less significant.

1.4 Gradual Learning and the Phenomenon of Has-Beens

This section extends the model by allowing information about talent to be revealed over time. While the optimal solution is analogous to that of the basic model, the case of credit constrained individuals is altered by the opportunity to save. I will show that, instead of mitigating the inefficiency caused by a credit constraint, saving will actually make things worse. It lowers exit rates even further below optimal, because some veterans of below-average talent will linger in the industry.

In the basic model, individuals have essentially two-period careers, with the relative length of the second "veteran" period described by T. All uncertainty about an individual's talent is resolved at a single point in time, so the only variable of choice is the exit threshold at that point. When new information about talent arrives at several points in time, then the decision to continue must take into account the option of exiting at a later time. Without further constraints, this would be a standard optimal stopping problem, introduced into the theory of labor markets by Jovanovic (1979). In this section I explore the general implications of a similar problem, but when job slots are scarce, learning is public and industry-specific, and individuals have finite careers and cannot commit to long-term contracts.

Assumptions

1. Each firm employs one worker per period, and the value of output is $y_t = \theta + \varepsilon_t$, where θ is the worker's talent and ε_t an i.i.d. error term.

2. There is an unlimited supply of individuals willing to work at outside wage $\underline{w}$.

3. Individual careers last up to $1 + T$ periods.

4. The CDFs of θ and ε are strictly increasing, continuous, and yield finite moments.

5. $\{\hat{\theta}_t, t\}$ is a sufficient statistic for $\hat{\theta}_{t+1}$, where

$$\hat{\theta}_t \equiv E[\theta|y_1, \ldots, y_t] \tag{1.12}$$

is the expected level of talent at tenure t (i.e. after t periods of work).

The expectation $\hat{\theta}_t$ is taken with respect to the known distributions f_θ and f_ε. For the novices, no output has yet been observed, so $\hat{\theta}_0 = \bar{\theta}$ for all of them. Since predictions are unbiased by definition, $\mathrm{E}[\hat{\theta}_{t+s}|\hat{\theta}_t] = \hat{\theta}_t$ for any $s = 1, \ldots, T-t$. Period t perceived talent $\hat{\theta}_t$ will often be simply referred to as talent. A crucial implication of assumptions 1 and 4 is that the distribution of prediction errors does not become degenerate in finite time: there is always some chance that the individual is better than he is expected to be.

Going forward in time, information about the talent of any particular worker becomes more precise; it gets in expectation closer to the true value and moves about less. However, it never becomes known for sure. In terms of a whole cohort, the distribution $f_{\hat{\theta}_t}$ starts as a degenerate distribution at $\bar{\theta}$, and then becomes more spread out. Without filtering it would become more like the true distribution f_θ; with filtering, more of the lower types, as well as some unlucky higher types, get discarded as time goes by.

Other assumptions

6. Firms maximize average per-period profits.
7. There is a unit measure of firms.
8. Price of output is normalized to one.

Assumption 6 means that there is no discounting. Assumptions 6 to 8 are made in order to simplify the notation. The extension of results from allowing free entry and an endogenous output price is straightforward in light of section 1.3, it would cause a further magnification of talent rents in the credit constrained case.

Social Planner's Problem

The variable of choice is a stopping (exit) policy $\boldsymbol{\theta}^* = \{\theta_1^*, \ldots, \theta_T^*\}$, which consists of T separate exit thresholds. Analogously to the single exit threshold of the basic model, this policy states that an individual with talent $\hat{\theta}_t < \theta_t^*$ will exit the industry. Since everyone looks identical at $t = 0$, there is no meaningful choice for θ_0^* (besides $\bar{\theta}$). On the other hand, after $T+1$ periods, the individual will retire anyway, and the updating of $\hat{\theta}$ based on y_{T+1}

is useless. Hence there are in total T points in time where a decision to continue or stop has to be made. A decision to exit is final, because after the exit no new information will ever arrive that could change the decision.

The average level of talent in the industry depends on the whole stopping policy. In line with earlier notation, denote the maximal solution by $A^* \equiv \max_{\boldsymbol{\theta}^*} A(\boldsymbol{\theta}^*)$. This would be the object of a surplus-maximizing social planner, as well as of a firm who could keep individuals at a fixed wage. The optimal solution must adhere to the following variant of the fixed point result in Proposition 1.

Proposition 7 *In the optimal solution,* $\theta_T^* = A^*$.

Proof We can assume without loss of generality that there is no turnover between job slots, since they are homogeneous. Consider then the problem of maximizing the long-run average talent at just one job slot, at a time when the job is currently held by a tenure T veteran of talent $\hat{\theta}_T$. If the veteran is rehired for his last period, then expected talent is $\hat{\theta}_T$ for the next period, after which a novice is necessarily hired. The long-run average from then on is A^*. If however, the novice is discarded, then the long-run average is A^* from this period on. At the optimal rehiring threshold these two courses of action lead to the same outcome, therefore $\theta_T^* = A^*$. □

Intuitively, given an individual with just one period left, the optimal decision of whether to retain him or not depends solely on his expected talent; there is no value for any further information about him. Thus he should be retained if and only if he contributes positively to the average talent in the industry.

Market Equilibrium

As in the basic model, equilibrium wages are determined on the spot market, taking into account that only individuals on their last period before retirement have no investment decision to make—it's all about the wages.

Proposition 8 *Wages are* $w(\hat{\theta}) = \hat{\theta} - \theta_T^* + \underline{w}$.

Proof is a combination of two observations. First, for individuals at tenure T, the value of continuing in the industry is simply $w(\hat{\theta}) - \underline{w}$. The lowest type to stay is one who gets exactly the outside wage $\underline{w}$ by doing so, hence $w(\theta_T^*) = \underline{w}$ regardless of the value

of θ_T^*. Second, firms must be indifferent between hiring type θ_T^* and any worker in the industry. The difference in current period wage between any two workers must be equal to the difference in their expected talent. □

In Jovanovic (1979, p. 976), workers have infinite lives, and this "*assumption justifies the exclusion of age as an explicit argument from the wage function.*" Here that exclusion follows from the existence of a spot market for talent. The market price of talent is constant in steady state: the economy is infinitely lived, even though individuals are not.

Proposition 9 *Given any $\theta_T^* \geq \bar{\theta}$, the optimal exit policy for risk-neutral individuals is strictly increasing in tenure: $\theta_t^* < \theta_{t+1}^*$.*

Proof by backwards induction. First consider an individual of type $\hat{\theta}_T = \hat{\theta}$ at tenure T. His payoff or "value function" is

$$V_T(\hat{\theta}) = \max\{0, \hat{\theta} - \theta_T^*\}. \tag{1.13}$$

The value function gives the excess expected utility from continuing as opposed to exiting, and zero if that difference is negative.

Next consider an individual of type $\hat{\theta}_{T-1} = \hat{\theta}$ at tenure $T-1$. If he decides to continue, he gets lifetime expected utility

$$\tilde{V}_{T-1}(\hat{\theta}) = \hat{\theta} - \theta_T^* + \mathrm{E}[V_T(a)|\{\hat{\theta}, T-1\}]. \tag{1.14}$$

The expected utility is taken with respect to $f_{\hat{\theta}_T|\hat{\theta}_{T-1}}(a|\hat{\theta})$. Since the expectation is increasing in the prior, also (1.14) is strictly increasing in $\hat{\theta}$. Since the distribution functions were assumed to be continuous, this is also continuous in $\hat{\theta}$. The optimal exit threshold θ_{T-1}^* is defined by $\tilde{V}_{T-1}(\theta_{T-1}^*) = 0$.

To see that $\theta_{T-1}^* < \theta_T^*$, notice that $\tilde{V}_{T-1}(\theta_T^*) > 0$, because the expectation is strictly positive at $\hat{\theta} = \theta_T^*$ (recall that the distribution of prediction errors does not become degenerate in finite time). Denote by V (without tilde) the value function that incorporates the current period optimal exit policy:

$$V_{T-1}(\hat{\theta}) = \max\{0, \hat{\theta} - \theta_T^* + \mathrm{E}[V_T(a)|\{\hat{\theta}, T-1\}]\}. \tag{1.15}$$

This is zero for $\hat{\theta} \leq \theta^*_{T-1}$, and strictly increasing for $\hat{\theta} > \theta^*_{T-1}$.

Completing the induction backwards in time is straightforward. The value function V_t is always zero below θ^*_t, where there is a kink, and then has positive slope above. Hence $\tilde{V}_{t-1}(\theta^*_t) > 0$ and $\theta^*_{t-1} < \theta^*_t$. □

Intuitively, of two workers of the same expected ability, the younger one has always more upside potential, because the prediction about his talent is less precise. The standards for hiring should therefore be tougher for older workers. In terms of the market equilibrium, the willingness to pay for a job slot is higher for a younger individual: paying for continuation today includes the option to continue tomorrow, and other things equal, an option on an asset with higher variance is more valuable.

Unconstrained Individuals

If individuals are risk neutral and not credit constrained, then market equilibrium must be efficient so that $\theta^*_T = A^*$. Again, the inability to commit to long-term contracts is inconsequential when individuals can pay the expected value of future rents to which that initial job opportunity may lead them. Competition from novices forces incumbent individuals to follow the socially optimal exit policy. This policy is illustrated in Figure 1-1 as the increasing graph from $\{0, \bar{\theta}\}$ to $\{T, A^*\}$. All possible individual paths for $\hat{\theta}$ must start at $\bar{\theta}$; an individual stays in the industry until retirement if and only if the path stays above the optimal exit policy throughout. At each point in time, wages are described by the vertical difference with the horizontal line at A^*, where they are equal to the outside wage.

With many potential points in career for exiting, the breakdown of the workforce by tenure can no longer be captured by the fraction of novices. As in the basic model, more novices are hired in the efficient case, but the exit rates (hazard rates of exit) are in general difficult to solve.

Constrained Individuals

Proposition 10 *If individuals are credit constrained, then $\theta^*_T = \bar{\theta}$, and no one will exit while $\hat{\theta}_t > \bar{\theta}$.*

Proof Novices are not scarce, so they cannot get more than outside wage $\underline{w}$. By assumption of being constrained, they cannot get less either. Therefore $w(\bar{\theta}) = \underline{w}$. This

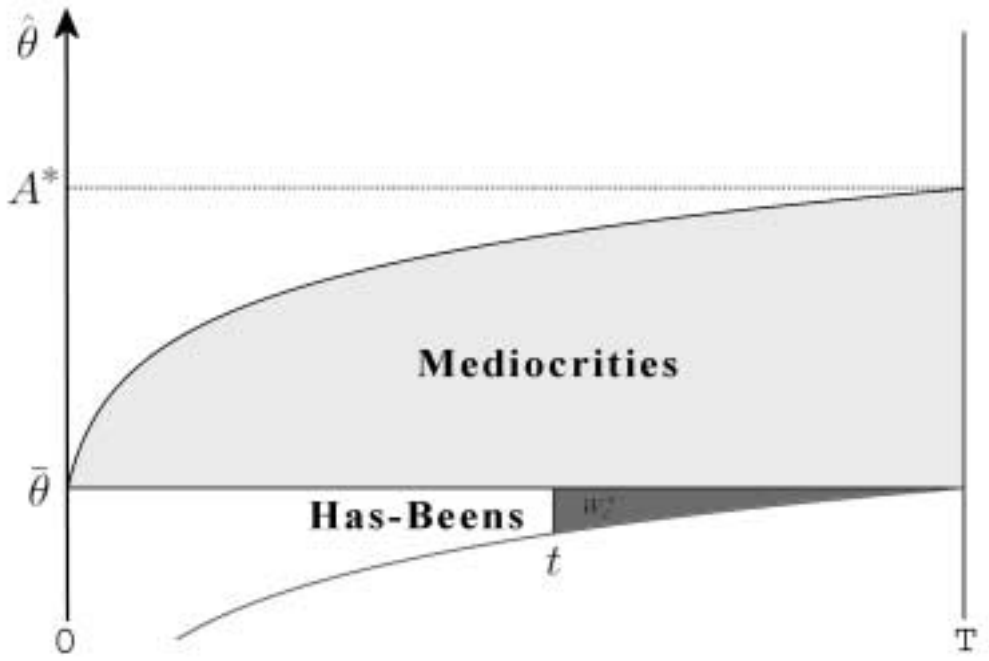

Figure 1-1: Mediocre types and has-beens.

must also be the wage of type $\hat{\theta}_T = \theta_T^*$, by Proposition 8. Therefore $\theta_T^* = \bar{\theta}$. But then everyone with expected talent above the population average is making rents, and will not want to exit. □

In contrast to the basic model, here the definition of mediocrity is age-dependent. A mediocre individual is above the population average, but below the optimal exit threshold for his tenure. As in the basic model, there is too little exit when individuals are credit-constrained. Mediocre individuals stay in the industry, even though their job slots would be in better use with novices.

The mediocre talents are illustrated by the light shaded region in Figure 1-1. They are the types with expected talent above population average but below the socially optimal exit policy. Now the wage is equal to $\underline{w}$ at the (solid) horizontal line, and talent rents at any point in time are given by the vertical distance from it.

If constrained individuals require at least the outside wage $\underline{w}$ regardless of past earnings, then the actual exit policy is $\theta_t^* = \bar{\theta}$ at all t. This behavior would arise if there was no saving, e.g. if individuals were extremely risk averse or impatient.

The Phenomenon of Has-Beens It seems reasonable to assume that individuals could save at least some of their rents. In this case the actual exit decision becomes path-dependent. Novices still cannot pay for job slots, because they have had no opportunity to accumulate savings, and this pins down wages and tenure-T exit threshold by propositions 8 and 10. Individuals would want to follow the optimal exit policy, which takes the last-period threshold $\theta_T^* = \bar{\theta}$ as given. Whether an individual with $\hat{\theta}_t \in [\theta_t^*, \bar{\theta}]$ can continue in the industry depends on his wealth.

Proposition 11 *If individuals are credit constrained, risk neutral, and are able to save, then some veterans of tenure $t = 2, \ldots, T-1$ will not exit even if $\hat{\theta}_t < \bar{\theta}$.*

Proof Consider an individual with $\hat{\theta}_{T-1} = \theta_T^* - \epsilon > \theta_{T-1}^*$ with a history $\{\hat{\theta}_1, \ldots, \hat{\theta}_{T-1}\}$ such that his savings are at least $\epsilon > 0$. The value of continuing is then, using the proof of Proposition 9, $V_{T-1}(\theta_T^* - \epsilon)$. This is strictly positive by definition of θ_{T-1}^* from (1.15): the individual will not exit.[16] $\square$

The optimal exit policy, with $\theta_T^* = \bar{\theta}$, is only *privately* optimal. Anyone who can, will follow the privately optimal exit policy. However, some individuals—including novices and everyone whose talent drops below population average after first period of work—are not able to do so because of the credit constraint.

By assumption, even risk neutral individuals need to consume at least $\underline{w}$ (possibly zero) every period to survive, but they do not mind saving all of the excess until retirement. Savings are useful by making it possible to follow the individually optimal exit policy in the future, should the individual's talent dip below the population average, but not so much as to go below θ_t^*. Previously successful veterans, or "has-beens", are able to compete against novices for scarce job slots, who would have a higher willingness to pay were it not for their credit constraint.

Denote the savings of an individual with history $\hat{\boldsymbol{\theta}}_{t-1} \equiv \{\hat{\theta}_1, \ldots, \hat{\theta}_{t-1}\}$ by $W_t(\hat{\boldsymbol{\theta}}_{t-1}) \equiv \sum_{s=1}^{t-1} \left(\hat{\theta}_t - \bar{\theta}\right)$.

Proposition 12 *For an individual with $\hat{\theta}_t = \theta_t^*$ to not exit, it is sufficient that*

$$W_t(\hat{\boldsymbol{\theta}}_{t-1}) \geq \sum_{s=t}^{T} \left(\bar{\theta} - \theta_s^*\right) \equiv W_t^*. \tag{1.16}$$

[16] Risk neutrality is assumed for simplicity, it is sufficient for individuals to be less than infinitely risk averse.

At the threshold, the required funds are equal to the spending under the worst case scenario, where the expected talent evolves along the stopping policy.

Proof The individual knows that on all possible paths in the future he will not be forced to exit due to the credit-constraint. He will therefore follow the optimal exit policy of an unconstrained individual. □

For individuals above the exit threshold $\hat{\theta}_t > \theta_t^*$, the required wealth for continuing is strictly less than W_t^*. This follows already from the fact that the current required payment for continuation is lower for a higher talent, and future prospects are at least as good. Of course, for $\hat{\theta}_t \geq \bar{\theta}$ the requirement is zero.

Notice that just having enough funds to pay for the next period's job is in general not enough for continuation to be worthwhile. This is because the expected benefits of continuation may mostly come from possible paths where the individual gets positive rents only several periods from now. As a result, some exiting individuals have positive savings.

Definition 3 Has-beens. *Individuals with* $\hat{\theta}_t \in [\theta_t^*, \bar{\theta}]$ *and* $W_t \geq W_t^*(\hat{\theta}_t)$, *where* $1 < t < T$ *and* θ_t^* *is part of the privately optimal exit policy that takes* $w(\bar{\theta}) = \underline{w}$ *as given.*

A has-been must have once been successful enough to have sufficient funds to continue.[17] He is currently below population average by expectation, but above the privately optimal exit policy. The presence of any has-beens in the workforce means that the efficiency loss in terms of average talent in the industry is now at least $A^* - A(\bar{\theta}\mathbf{1})$.

Where could we observe "has-beens" in the sense defined here? In the movie industry, a has-been could be an actor who used to be a star and made large talent rents, but has flopped more recently. He then uses savings from earlier rents to participate in the financing of a movie, which makes negative profits in expectation, but offers him a role and a chance at a resurrection. For him, this gamble has a positive expected value, because a successful comeback would generate more talent rents in the future. However, it is socially inefficient, because novices would have an even higher willingness to pay for the role, but they do not have the funds to bid for it.

Interpretation as One-Sided Long-Term Contracts Suppose firms can commit to a lifetime wage-policy, including severance payment policy, but individuals cannot commit

[17] MacDonald (2001) studies has-beens with a vintage human capital model under uncertainty about future technological change.

to contracts that require them to make payments to the firm at any time in the future. Now the accumulated wealth W_t would be analogous to money "in escrow" at the firm, which must always be nonnegative. In the simplest contract individuals would get $\underline{w}$ until they exit, upon which the firm pays out W_t. Some separations result from insufficient funds in escrow, but even they can involve severance payments by the firm. When the escrow is full, i.e. $W_t \geq W_t^*$, exit is voluntary: the worker quits to stop the bleeding of the escrow because he has fallen below the privately optimal exit policy θ_t^*.

More interestingly, the contract could also include wages above $\underline{w}$ before separation or retirement. For sufficiently good histories, the escrow balance can reach a point where no amount of bad news in the future could ever cause the individual to be fired due to insufficient funds. In terms of the spot contract world, the wealth constraint can no longer become binding until T because $W_t \geq W_t^*$, even though the privately optimal exit policy can still become binding after sufficiently bad performance. This allows the firm to start unloading the account with payments above $\underline{w}$, up to the point where the remaining balance is W_t^*.

The result that a worker's escrow can reach a firing-proof level W_t^* is reminiscent of the "tenure standard" of Harris and Weiss (1984), but this is a different phenomenon. In their paper, for a sufficiently good history of performance, the expected marginal product of a worker reaches a level where the firm knows it can never again fall below the outside wage. A crucial assumption there is that output consists of successes that arrive as a Poisson process and that failures are not possible: the magnitude of the worst possible news, i.e. lack of news for the rest of career, is therefore bounded below.

Firms' ability to commit to long-term contracts does not improve efficiency here, it merely allows a different interpretation of the equilibrium. Addition of unobservable effort would let the escrow serve the useful purpose of (imperfectly) mimicking an up-front performance bond, as in Akerlof and Katz (1989), something which credit-constrained individuals are unable to post. Also, if individuals were both credit-constrained and risk averse, then one-sided long-term contracts would allow firms to provide insurance, as in Harris and Holmström (1982). The difference here is that wage insurance against type realizations below θ_T^* is provided by the outside wage (turning out to be a bad actor does not diminish one's prospects as a dishwasher). Note that there is no scope for wage insurance in the basic model, because workers do not face downside risk: novices know that they will make

at least their current wage in the future, and veterans know that their type (and wages) will stay constant until retirement.

1.5 Applications

The prototypical and most high-profile talent markets are found in the entertainment industry. There job performance is almost entirely publicly observable and success of young talents hard to predict. Neither formal education nor on-the-job training seem to play a large role in explaining wage differences in these industries. The chance to reveal one's talent in a real job is precious: this, and the presence of credit constraints, is suggested by the queuing for positions and auditions. Testing and other artificial ways to assess talent seem to have limited usefulness beyond reducing the number of candidates for any entry-level position. There simply is no good substitute for observing the success of actual end-products. Richard Caves (2000) has dubbed this uncertainty the "nobody knows" property, as the first on a list of distinctive and pervasive characteristics of the entertainment industry. It could be said that in the entertainment trades finding out about someone's talent is largely finding out about the tastes of the public, but this distinction is not operational for analytical purposes.

For a talent market to be analyzable with this model, it should exhibit certain broad features. There should be relatively high exit rates early on (this is true without long-term commitment, although more so with it). The level of talent should be imprecisely known at the entry level, and then become known relatively quickly once in the industry. This would appear as a quick increase in within-cohort income dispersion among the "survivors" in the industry (under long-term contracting, only among free agents). Observed performance in one firm should be a good predictor of performance at other firms, i.e. match-specificity should not be too important. If these conditions hold, and if there is reason to believe that novice-hiring firms are not compensated for the talent they discover, then this would suggest the potential for inefficiencies and excess talent rents described in the model.

There are many models to describe markets for talent that are consistent with stylized facts about entertainment industry, such as high and skewed income distribution. Just observing a talent market under one set of institutions does not allow one to show the existence, not to mention estimate the magnitude, of any inefficiencies. Besides comparing

models by the plausibility of their assumptions, it would of course be desirable to try to identify and quantify "the curse of mediocrity" proposed in this chapter. This would require an exogenous change in one of the imperfections behind the inefficiency—a natural experiment. The ideal experiment would be a surprise legal change from full individual commitment ability to none or vice versa. Such a change would also allow the quantification of the economic value of commitment ability, and its impact on within-profession income inequality. While the model of this chapter is not sufficiently rich to allow a careful empirical analysis of such natural experiments, it can be used to shed light on stylized facts and to suggest potential empirical applications.[18]

1.5.1 Motion Pictures

The motion picture industry in Hollywood operated under the so-called studio system from 1920s to late 1940s. In this system, artists and other inputs were assembled together within a studio under long-term relationships. As a part of the system, entering actors made exclusive seven-year contracts with movie studios.[19] This kept their compensation at moderate levels until the initial seven years came to an end, even if they became big stars meanwhile. This allowed studios to capture much of any increase in an artists' worth during the contract. The studios could rent the artist to other studios on "loan-outs" (for which they charged a premium from the renter), but the artist had no right to quit or refuse roles. The contracts did *not* provide insurance. Even though wages were specified for the whole contract period (typically including moderate increases), the studios had the right to terminate the contract every six or twelve months.

A successful lawsuit by actress Olivia de Havilland, resolved in 1945, made a crucial part of these long-term contracts unenforceable. She had been hired by WB (Warner Bros.) in 1935, having been an unknown protagonist of a college theater play. She quickly proved very popular with both audiences and critics and won her first Oscar nomination four years into the contract, which she then attempted to renegotiate. She refused roles offered by WB, and as a result did not work for six months. At the end of the contract WB claimed that the skipped six months should be added to the contractual obligation, since the original

[18] A careful empirical analysis would require—besides good data of course—a dynamic model that takes into account how a market adjusts from one steady state to another (which can in principle take a lifetime), and how it reacts to demand shocks. This depends on features that do not make a difference to the steady state, most importantly, whether previously exited individuals can come back to the industry later.

[19] The seven-year limitation on personal service contracts dates back to 1890s.

contract required her to actually work for seven years.[20] WB lost, and the "De Havilland decision" made long term contracts less useful, as it gave more renegotiating power to artists who turn out to be big stars.

At the same time, the studio system came under fire from the Justice Department, which filed an anti-trust lawsuit against Paramount Pictures in 1938. This suit accused the eight major studios, which among them produced 95% of movies, of monopolizing the motion picture industry by restraining trade and fixing prices. The main thrust of the suit was aimed at the vertical integration of movie theaters and studios. The Supreme court decision in 1948 forced the studios to divest from movie theaters, which is commonly thought to have ended the studio system. Whatever the reason, the system of long-term contracting ended in the 1940s. After the change, movies have been produced as one-time affairs, where an entrepreneur-producer assembles a line of talents and other inputs for one movie only.[21]

According to the model, the end of long term contracting should have led to insufficient exit of mediocre entertainers, showing up as substitution from unexperienced actors to experienced (but relatively less paid) actors, to higher and more uneven incomes for veteran actors, and to lower total revenue. The wages of star actors on their initial contract during the studio system can be expected to be lower for obvious reasons. More interestingly, the contractual situation of free agents (those past the initial seven years) under the studio system is comparable to actors with the same amount of experience under spot contracting. During the studio system, there should have been a higher supply of talent due to better use of movie roles in discovering talent, moderating also the wages of star free agents. After the change, the share of less experienced actors should have gone down, but the special nature of the product makes predictions about the age structure less clear-cut: actors of different ages are not well substitutable, as the actor's age must be matched with that of the character in the script. Yet to the degree that it is feasible, the end of the studio system should have led to actors being older than their roles.

A major difficulty of the episode is that other things were far from equal. The advent of television in the late 40s and early 50s is a major technological shock, affecting both demand

[20]Sources: Screen Actors Guild History Page, *www.sag.org*, and Capellon & McCann trial lawyers, *www.cappellomccann.com*.

[21]It has also been suggested that the system unraveled because of 90% personal income tax rates during World War II. This caused individuals to set up their own production companies to shift taxable income toward dividends (also complicating any empirical analysis), which were taxed at 60%. See Stanley (1978), Chapter 3. Presumably too frequent dealing with the same studio would have exposed the tax dodge. However, the return of lower tax rates did not bring back the studio system.

for actors and movie tickets. The motion picture industry was left with a comparative advantage in high quality (e.g. color film), but the market for actors was probably integrated across these two media, something which a serious empirical analysis should also take into account.

Unfortunately, the wage data for actors is lacking. According to Caves (2000, p. 389), "*no systematic data have been assembled on whether the studios' disintegration brought more rents into the stars' hands, but casual evidence suggests that it did.*" There is more concrete evidence of a post-war decline in revenue and output at movie studios. The number of movies made was down 48% from the 1940 level in 1956, while revenues declined by 19%; again, this fact is difficult to interpret without quantifying the impact of television.[22] Interestingly though, in terms of quality, the era from the 1920s to the 1940s is often referred to as the golden age of Hollywood movies. For example, according to film director Peter Bogdanovich "*It was a whole system that found actors who were unusual, not necessarily versatile in the way we think of versatile actors today, but actors who had a personality, who had a certain quality ... there was a whole system to that, and it was extraordinary and produced the greatest array of star actors in the history of the world.*"[23]

1.5.2 Record Deals

Exclusive record deals, by which musicians agree to make a certain number of albums for the same record company, are a form of long-term commitment similar to what used to be possible in the motion picture industry. This arrangement is possible in the record industry, because record deals are exempt from the seven year limitation on the length of personal service contracts. Challenges similar to the De Havilland case were forestalled by the California legislature in 1987, when it was decreed that record companies retain rights to the agreed number of albums by an artist, even if seven years has passed since the signing of contract.[24]

The music industry is very competitive at the entry level, where upstart bands and artists are free agents, but agree to exclusive contracts in exchange of production, distribution, and promotion by the record company. The production cost alone for a typical record is from

[22] Average costs (available for two studios) roughly doubled at the same time, but I have not found data on the share of wage costs. Figures are from Conant (1960), Chapter 5.

[23] MacNeil/Lehrer NewsHour, PBS, July 3, 1997.

[24] This amendment is Subsection B of California Labor Code Section 2855.

$100,000 up,[25] but the biggest cost may be the opportunity cost of promoting one band rather than another. The scarcity of attention of programming directors for radio stations and people looking for new music for record shops means that a record by its mere existence has little chance of becoming known. The prospects of which artist will become a big seller are very uncertain. About 80-90% of records by new artists end up making a loss—this must be compensated by the small number of very profitable hits. For the record companies, the most profitable hits are those by artists still on their initial low-paying contracts.

However, the efficacy of the system is constantly threatened by attempts to renege or renegotiate by those who turn out to be big stars and end up getting paid much less than their current "market price" (high-profile cases include Prince and George Michael). The quality of the product is obviously not contractible, and artists can fulfill contractual requirements (or try to force a renegotiation) with a substandard product, though at a reputational cost to themselves. Furthermore, there is currently a lobbying battle in the Congress involving RIAA (Recording Industry Association of America) and AFTRA (American Federation of Television and Radio Artists) about the continued application of the seven-album amendment. Were the current system of record deals to break down, the proportion of new artists and new releases can be expected to be reduced, while the proportion of new artists breaking even and making a second record should go up. A reduced proportion of "failed artists" would be a sign of reduced experimentation and lower efficiency.

1.5.3 Professional Team Sports

Professional team sports in North America have very unusual labor markets, mainly because the firms are organized into leagues that are close to natural monopsonies in their specialized labor markets. The leagues have devised rules that restrict firms from competing for each others' employees. In particular, potential novice players ("rookies") are each assigned to a single firm, which then has the sole right to negotiate with that particular player (the allocation of these monopsony rights is known as the "draft"). Under the "reserve clause" system, players cannot leave for other firms at will, but employers can always sell the player's contract to another firm. This system was upheld by a U.S. supreme court ruling, Flood v Kuhn (1972), against a challenge by baseball player Curt Flood who had been traded against his will.

[25] Vogel (2001).

Players have responded to owners' monopsony power by unionization, leading to occasional strikes.[26] Baseball players achieved some concessions through collective bargaining in 1975, after which players reaching six years of league experience became eligible for free agency, where all teams are free to bid for their services. This change seems to have been anticipated, and 1975 was more like a culmination of gradual unraveling than a sudden shift. The change is only applicable to a minority of players however, since slightly more than half of careers do not last long enough for a player to get a contract as a free agent.

The exit (hazard) rates of major league baseball players indicate that a major shift took place in the 1950s. In the first half of the century, more than half (52.8%) of players exited after no more than three seasons, and over two thirds (68.2%) by the end of their sixth year.[27] From 1960 to 1990 these rates were down to 33% and 50.1% respectively, without a significant break at 1975. For rookies the exit rate was 35.7% before 1950, and 17.2% after 1960. Meanwhile the average age of new players has stayed at 24 years, while the number of teams and players has been growing. Further investigation would be necessary to establish the cause of the shift in exit rates, but based on the model in this chapter, increasing (re)negotiating power of players is a prime candidate.

The accuracy of information about novice talent in professional sports remains an open question under the reserve clause. The draft makes it nearly impossible to evaluate the economic value of expected talent differences between novice players.[28] If prior information is very inaccurate, then the draft should not make much difference to wages.[29] On the other hand, if the rookies also differ from each other substantially by the expected value of their talent, then the reserve clause is both rent extraction (the draft) and remedy to the curse of mediocrity (enforced long term commitment) bundled in one. However, instead of being just a transfer of rents from owners to players, as usually claimed by pundits, an implication of the model is that complete free agency could be expected to cause a welfare loss. It would lead to lower exit rates for young players, lower average quality of players and lower total revenue. In total, players gain less than the owners lose.

[26] First collective bargaining agreement is from 1968; there have been five strikes and three lockouts in major league baseball since then.

[27] Based on data from Sean Lahman's website "The Baseball Archive," *www.baseball1.com*.

[28] Occasional barter between teams, where draft numbers are traded for free agents, could allow some inference.

[29] According to Rottenberg (1956), "the process by which players are brought to the major leagues can be likened to that by which paying oil wells are brought in or patentable inventions discovered."

A similar but potentially much stronger natural experiment may be about to start in Europe, where the system of transfer fees in professional soccer is under scrutiny by EU labor regulators. There young players start as free agents but have the right to commit to binding long-term contracts, the length of which can be negotiated.[30] Casual evidence suggests that entry level information about talent is very inaccurate compared to what is known 4-5 years later. If transferable contracts become unenforceable, then players can be expected to gain more than will be the loss to owners and consumers; at the same time, the age distribution of players should move upwards.

1.5.4 Entrepreneurship

It may be useful to think of the market where entrepreneurs and venture capitalists meet as a talent market. This market would exhibit the curse of mediocrity if two conditions are met. First, the success of a new firm should depend on the talent of its founding entrepreneur, of which relatively little is known until after his first project is financed. Second, entrepreneurs should be able to go on to found new companies later in their career, and the profits of these new firms cannot be claimed by the financiers of previous firms. In this case, much of the expected value of financing a start-up by a novice entrepreneur is not contractible, because it will accrue to the entrepreneur through profits of future projects. As a result, the investment decisions of venture capitalists do not take into account the value of information produced about the abilities of the entrepreneur, only the expected profits from the current project. There is too little investment into projects of inexperienced entrepreneurs, while too many mediocre entrepreneurs go on to found more companies. The mediocrities' new companies are profitable by expectation, but they are not as profitable as is the expected lifetime profitability of novice entrepreneurs' projects, taking into account that unsuccessful entrepreneurs will be filtered out of the market. Known entrepreneurial talent is artificially scarce, leading to excessive incomes for incumbent entrepreneurs. Under these circumstances, we could also expect to see has-been entrepreneurs using their own wealth from previous projects to try to bounce back into talent rents.

[30] In some European countries the contract length became freely negotiable only after the 1995 "Bosman decision," until which a player's old team could require a transfer fee from the new team, even at the end of the contract.

1.6 Conclusion

This chapter has presented a model of a labor market where individual talent can only be revealed on the job. Firms face a joint production problem of output and information about talent, the optimal solution to which has been well understood at least since Jovanovic (1979). In this study it is assumed that individuals cannot commit to long-term contracts and that learning about talent is public, so wages are determined on a spot market for talent. The distinguishing feature in this study is the scarcity of job slots and the focus on the case where individuals are credit constrained. The problem is that firms do not have incentives to hire novices, who themselves cannot pay to be hired. As a result, there are too many mediocre workers, who are better than novices by expectation, but not talented enough to justify them taking up scarce job slots that could also be used to try discover higher talent. Wages are always higher in the inefficient case because revealed talent is more scarce than it need be, and because the price of output (and so the value of talent) is higher.

The model presents a rather bleak picture of talent markets. Most labor markets are markets for ex-ante unobservable talent to some extent. The markets for lawyers, copywriters, and college professors are among potential cases not explored in this chapter. Whether a labor market exhibits the inefficiency and excess rents described in this chapter, and whether these are economically significant, is of course an empirical question. Estimation would require an exogenous change to one of the imperfections behind the inefficiency, but such changes are rare. This chapter suggested potential natural experiments from the entertainment industry for detecting and quantifying these problems, but a careful investigation is left for future research. If the differences in talent that are only discovered on the job are indeed economically significant, then much of observed superstar wages could be a symptom of potentially large inefficiencies, resulting from limitations to contracting.

Occupational licensing is a straightforward solution to the problem of insufficient exit. If workers are required, after a certain amount of experience, to have shown a certain level of talent to stay in the industry, then setting this requirement at the efficient threshold would result in the best achievable distribution of talent in the industry. Equivalently, if all firms had a large number of jobs, then it would be sufficient to require them to set aside a certain proportion of jobs to novices. Instead of relying on a benevolent regulator,

this arrangement could be achieved as a collusive outcome between firms, and might even be allowed by anti-trust legislation, unlike the more simple solution of firms agreeing not to bid for each others workers. The result is the same: less retaining of mediocre talents, and lower wages. Under free entry this "industry standard" would end up benefiting the consumers by lowering the price of output.

The frictionless and informationally transparent spot market assumed in this study all but precludes private solutions to the problem. Firms have no incentives to invest into organizational capital, in the sense of Prescott and Visscher (1980). If learning were not completely public, or if there were other frictions in the labor market (such as switching costs), then firms would get a part of the rents from the talent they discover and would have some incentives to hire novices over mediocrities. For example, in the signalling model of Waldman (1984), competing employers only observe the job category of workers, but can infer the average quality of workers in each category. This leads to underpromotion of workers to the high-productivity job, because promotion must involve a discontinuous wage increase. In light of this study, it seems that models that yield underpromotion in equilibrium could result in overpromotion from the point of view of social efficiency, if they were augmented with a credit constraint and scarce job slots. A further extension into heterogeneous jobs would connect this model with the literature on how organizations use different types of job slots to "breed" high-ability workers, possibly with strategic interactions between firms. The most relevant models would be Guasch and Sobel (1983) and Demougin and Siow (1994, 1996).

The most pressing theoretical extension to this model is the addition of heterogeneous jobs. This could be done by using an assignment model approach presented by Sattinger (1979, 1993), or a model where individuals form teams, as in Kremer (1993), or even with a model that integrates not only learning and job assignment, but also on-the-job training, along the lines of Gibbons and Waldman (1999). In each case, this extension would definitely overturn one apparent conclusion of this chapter, namely that frictions are always good for efficiency. This "feature" is an artifact of assumed job homogeneity and is not robust to a within-industry matching problem. With job heterogeneity, the efficient matching of workers and jobs is subject to change over the lifetime of a cohort as new information becomes available. A limited amount of "poaching" can then be beneficial because it allows workers to move to jobs where their talent has higher productivity. When individuals have

limited commitment ability, this mobility comes at the cost of worsened incentives for the hiring of novices. Reaching full efficiency gets harder when jobs are heterogeneous: it is not enough for individuals to be able to commit to work for a particular firm; they should be able to commit to a contract that can be sold to other firms.

Chapter 2

The Difference That CEOs Make: An Assignment Model Approach

2.1 Introduction

It is a well known fact that larger firms pay more to their CEOs. The elasticity of CEO pay to firm size has been estimated at about 0.3 across industries and time with various measures of firm size.[1] The literature on executive compensation has mainly focused on the structure of incentive pay, while the level of pay has received much less attention. Most studies end up attributing the differences in pay levels to different optimal effort or risk levels, for which the essentially homogenous CEOs must be compensated. In this study the distribution of CEO pay is analyzed as the outcome of a competitive equilibrium in a market where heterogeneous firms and individuals match. The goal is to disentangle scale effects from inherent ability differences in explaining the observed pay differences and to estimate the social value of scarce executive ability.

It seems fairly intuitive that the observed strong relation of firm size and CEO pay levels is a manifestation of scarce executive ability being worth more to larger firms, because the economic impact of a manager's decisions depends on the amount of resources under his control. That this relation results in high levels and skewed distribution of income for CEOs was proposed by Mayer (1960) who used the term "scale-of-operations" effect. In similar spirit, Manne (1965) argued that a major benefit of corporate mergers and takeovers is to

[1] See for example Kostiuk (1999), whose data goes back to 1930's, and the survey by Murphy (1999).

allow the allocation of the control of resources to be adjusted to managerial abilities. Lucas (1978) invoked Manne's suggestion to devise a theory of firm size distribution based on the allocation of capital to a population of potential managers of heterogeneous ability. Rosen (1982) presented a related model with a focus on the division of labor into managers and workers and the allocation of subordinate labor between managers. In these models all size differences between firms are due to differences in managerial ability, although better economies of scale increase the skewness of the distributions of firm size and managerial pay.

In assessing the value of CEO ability it must be taken into account that each firm has only one CEO, and that each individual can work in only one firm at a time. If not just individuals but also firms have important indivisible characteristics then this simple fact has far-reaching implications for the understanding of CEO pay levels. An assignment model is called for; for early assignment models see Koopmans and Beckmann (1957) and Tinbergen (1956, 1957). The assignment model used here builds on the "differential rents" model of Sattinger (1979) by introducing adjustable capital that can be freely allocated between the matched pairs of firms and managers. Sattinger's setup has a continuous distribution of workers and firms which rules out match-specific rents and therefore any need for bargaining, and a complementary production function which guarantees positive assortative matching (here meaning the matching of the best managers with the largest firms).[2]

This chapter presents a new approach to assignment models by describing distributions in terms of their inverse distribution functions or "profiles." The crucial variable describing individuals is now their quantile in the distribution of ability, and not the level of ability which typically lacks a natural scale of measurement. The distributions of factor incomes are solved in a manner analogous to the standard method of solving screening models. I believe this quantile approach to be more intuitive and tractable than working with density functions, especially when considering empirical applications. In particular, this approach makes it clear how the rents accruing to the ability of an inframarginal individual are equal to her marginal productivity, defined with respect to total industry output while taking into account the resulting reassignment of firms and other individuals if she were to exit the industry.

[2]See also the survey by Sattinger (1993) which includes a detailed exposition of the "differential rents" model.

The basic assumption of this chapter is that there is a competitive and frictionless labor market for executive ability which is equally applicable in all companies, but is more productive at larger companies.[3] Even though all firms would rather hire the most able individual for the job, it is the companies with the highest absolute value for that ability that will pay the most for it and therefore attract the best individuals. In equilibrium each firm must prefer hiring its CEO at her equilibrium pay level to hiring any other company's CEO at their pay level. At the same time, the levels of adjustable capital must equalize its marginal product across firms at the market rate of return. The pay levels of individuals depend on the distributions of firm size and CEO ability in the economy.

A general lesson of assignment models is that the income distributions of cooperating factors must be analyzed together. A regression approach, such as estimating an earnings function, could be wildly misleading. In the empirical section of this chapter, I use the assignment model to analyze the dependence of the pay distribution on the distributions of individual and firm characteristics. Using the observed joint distribution of CEO pay and shareholder income, the model can be used to answer various quantitative questions about the effects of CEO ability on social surplus and CEO pay. These questions necessarily take the form of counterfactuals about one distribution of factor quality while holding the other constant. Furthermore, here the sensitivity of some results to the assumed value of a free parameter in the model (the elasticity of output with respect to capital) means that only the counterfactuals about the distribution of ability yield meaningful results. Many other questions cannot even conceivably be answered if the model is taken seriously.[4]

This study has the polar opposite approach to CEO pay of most of the literature because the structure of pay is not considered, only the level. Differences in required effort or risk-bearing are in effect assumed away as possible explanations for the variation in pay levels in favor of differences in individual ability and firm-specific usefulness for that ability.[5] Undoubtedly incentive pay is needed to align the interests of managers with those of the shareholders in a firm of any size. In a perfect contractual world the effective ability

[3]Parrino (1997) provides evidence for positive assortative matching in the market for CEOs (as well as for substantial frictions). He shows that succesful CEOs that switch firms are more likely to move to a larger company.

[4]For example, due to the assumed lack of frictions, any movements of CEOs between firms would only reflect changes in information about their ability and could not be used to identify the value of their ability; although this can be sensible within other models (e.g. Hayes and Schaefer 1999).

[5]Assumed away are also explanations based on dishonesty such as the skimming explanation in Bertrand and Mullainathan (2001).

that managers provide would be higher for each inherent ability level. In this sense the model hides all incentive problems under the levels of ability, and the expected cost of a CEO's compensation is simply the market price of her managerial ability. A somewhat similarly motivated paper within the incentive literature by Baker and Hall (2002) explores the relation of incentives and firm size while assuming away differences in ability. In their model, effort and firm size are allowed to be complementary, so the optimal level of effort and sensitivity of compensation to market value depend on firm size. Using cross-sectional data on the structure of CEO pay and firm size, they find evidence for a substantial complementarity: the estimated elasticity of the marginal productivity of CEO effort with respect to firm size is about 0.4.

For an assignment model to be tested and estimated, one would need to observe individual and firm characteristics that well capture the variation in management ability and firm size but have not been affected by the characteristics of one's matching partner. Furthermore, these would have to be observed over time and with sufficient exogenous variation in the shapes of the distributions. This is too much to ask for in the case of CEOs and firms, but the model can still be used to answer quantitative questions by assuming a functional form for the relation of output to unobservable ability and firm size. The value of within-sample differences in managerial ability can be estimated under the assumption that effects of managerial ability are multiplicatively separable from production technology and unobserved firm characteristics.

For the empirical implementation I use CompuStat data on the 1000 largest publicly traded companies in the US in 1999 (companies with market value above $614 million). The main quantity of interest is the difference that their CEOs make to total economic surplus, compared to the counterfactual case where they would only have the same undetermined baseline ability as the lowest type CEO in the sample. This value was about $25-37 billion in 1999, of which the CEOs' received $5 billion. Another counterfactual is random matching within the sample: the implied social value of sorting the top 1000 CEOs by company size is estimated at $11-15 billion. Finally, the difference in pay levels between the CEOs of smallest and largest sample companies decreases by about a factor of ten in the counterfactual case where all firms are similar to the current 1000th largest firm. It could be said that the bulk of the observed size-pay relationship is explained by the exogenous differences in firm size (rather than by differences in managerial ability).

The chapter is divided into two parts. In the theoretical part (Section 2) an assignment model is presented, where firms and individuals of exogenous qualities match, and where the level of capital can be adjusted according to the quality of the match. It is shown how the division of surplus into factor incomes depends on the distributions of individual and firm characteristics. The nonstandard intuition of assignment models is discussed and clarified. In Section 3 the model is applied to CEO pay. It is shown how the ability profile of individuals can be inferred, up to a constant of proportionality, from the observed joint distribution of CEO pay and market value of firms. The model is used to estimate the value of CEO ability in the largest U.S. companies, under various assumptions about the elasticity of output with respect to capital. The chapter is concluded with a discussion of the results.

2.2 An Assignment Model of Pay

In an assignment model productive resources are embedded in indivisible units and these units must be combined in fixed numbers to produce output. Here the units are individuals and firms, and they are matched one with one. A production function describes the resulting output from any individual with any firm as a function of their characteristics. I make three simplifying assumptions about the production function: one-dimensionality of inputs, continuity, and complementarity. The first two assumptions are made for analytical convenience, while the complementarity assumption is central to the whole approach. Further assumptions are symmetric information and risk neutral firms.

The first assumption is that individual and firm characteristics affecting output can both be summed up by one number. If these characteristics are denoted by $\mathbf{x}$ and $\mathbf{z}$ respectively, then the output from matching individual i and firm j can be written as $y_{ij} = Y(\mathbf{x}_i, \mathbf{z}_j) = Y(a(\mathbf{x}_i), b(\mathbf{z}_j))$. In other words, units of input have one-dimensional sufficient statistics with respect to output; these statistics will be referred to simply as "ability" and "size", denoted by a and b respectively. The units of measurement of ability and size are only defined up to a positive monotonic transformation, but there is an unambiguous ranking of individuals and firms by their productivity that is independent of who they are being matched with. Note that different individuals can have different "strengths", i.e. different components of $\mathbf{x}$ contributing to their ability to affect output. These different components

can be complements as well as substitutes, only the "aggregate" qualities of the factors must be complementary to guarantee positive sorting.

Second, it is assumed that the production function is continuous and strictly increasing in both of its arguments, and that there is a unit mass of individuals and firms with "smoothly" distributed characteristics. The distributions of a and b have continuous finite supports and no atoms; the resulting distributions of output and factor incomes will inherit these properties. Dispensing with this assumption would only complicate the notation without bringing more insights.

The substantive assumption is that of complementarity. When the production function has a positive cross-partial, then efficiency requires positive assortative matching: the best individual must be matched with the largest firm, the second best with the second largest etc. If the sorting were not perfect, then total output could be increased by shifting some individuals between firms.[6] The individuals and firms are thus matched in the simplest possible way in equilibrium. The determination of output is very straightforward, its division into factor incomes is what requires further analysis.

It will be convenient to refer to distributions by their inverse distribution functions or "profiles". Think of the individuals as ordered by their ability on the unit interval, so that $a[i]$ is the ability of the i:th quantile of individuals and $a'[i] > 0$. In general, when the mass of "observations" is normalized at one, then the profile of any positively sorted variable is also its inverse distribution function. Denoting the distribution function by F_a, the profile of a is defined by

$$a[i] = a \text{ st. } F_a(a) = i. \tag{2.1}$$

The slope of the profile is the inverse of the density:

$$a'[i] = \frac{1}{f_a(a)} \text{ st. } F_a(a) = i. \tag{2.2}$$

$$a''[i] = -\frac{f_a'(a)}{f_a(a)^3} \text{ st. } F_a(a) = i \tag{2.3}$$

If there were atoms in the distribution of a they would correspond to flat parts in the profile,

[6] Positive assortative matching ("positive sorting") maximizes the output from matching $a_1 \leq a_2$ and $b_1 \leq b_2$ if $Y(a_1, b_1) + Y(a_2, b_2) \geq Y(a_1, b_2) + Y(a_2, b_1)$. Rearranging this inequality to $Y(a_2, b_2) - Y(a_1, b_2) \geq Y(a_2, b_1) - Y(a_1, b_1)$ illustrates the fact that complementarity can also be defined as "increasing differences" in the production function.

while gaps in the support of a would appear as jumps.

2.2.1 The Determination of Factor Incomes

In a competitive equilibrium, the profiles of factor incomes must support the efficient matching of individuals and firms, which we know involves perfect sorting. Two types of conditions must hold in competitive equilibrium. First, there are the sorting constraints: all firms must prefer hiring their efficient match at the equilibrium wage to hiring any other individual at their equilibrium wage. Second, there are the participation constraints: all firms and individuals must be earning at least their outside income. Notice that the sorting constraints look like incentive compatibility conditions in a typical nonlinear pricing problem.

$$\begin{array}{lll} Y(a[i],b[i]) - & w[i] \geq Y(a[j],b[i]) - w[j] & \forall i,j \in [0,1] \quad \text{SC}(i,j) \\ Y(a[i],b[i]) - & w[i] \geq \pi_0 & \forall i \in [0,1] \quad \text{PC-}b[i] \\ & w[i] \geq w_0 & \forall i \in [0,1] \quad \text{PC-}a[i] \end{array} \tag{2.4}$$

The outside opportunities are assumed to be the same for all units of a given factor.[7] The unit mass should be thought of as a normalization of the mass of pairs of individuals and firms that are active in equilibrium. The lowest active firm-individual pair is the one that just breaks even with the outside opportunity:[8] $Y(a[0],b[0]) = \pi_0 + w_0$. The firms are not residual claimants in any sense: the equilibrium conditions could equivalently be stated in terms of individuals hiring firms.

As in the mathematically analogous nonlinear pricing problem, the amount of constraints can be reduced drastically by noticing that for any $i \geq j \geq k$, the sum of two adjacent sorting conditions SC(i,j)+SC(j,k) implies SC(i,k). The binding constraints are the marginal sorting constraints that keep firms from wanting to hire the next best individual, and the participation constraints of the lowest types. Regrouping the sorting constraint SC$(i,i-\varepsilon)$ and dividing it by ε gives

$$\frac{Y\left(a[i],b[i]\right) - Y(a[i-\varepsilon],b[i])}{\varepsilon} \geq \frac{w[i] - w[i-\varepsilon]}{\varepsilon}. \tag{2.5}$$

[7] A weaker assumption would do here, namely that the outside opportunities increase slower along the profile than the equilibrium wage does.

[8] If the amount of available factor units were binding, so that even the lowest types more than break even, then the lowest types of the binding factor would get a positive rent.

This becomes an equality as $\varepsilon \to 0$ and, via the definition of the (partial) derivative, yields the slope of the wage profile.

$$w'[i] = Y_a(a[i], b[i])a'[i] \tag{2.6}$$

The wage profile itself is then obtained by integrating the slope and adding in the binding participation constraint $w[0] = w_0$.

$$w[i] = w_0 + \int_0^i Y_a(a[j], b[j])a'[j]\mathrm{d}j \tag{2.7}$$

Analogously, or from $y = \pi + w$, the profile of profits is

$$\pi[i] = \pi_0 + \int_0^i Y_b(a[j], b[j])b'[j]\mathrm{d}j. \tag{2.8}$$

All inframarginal pairs produce a surplus over the sum of their outside opportunities, and the division of this surplus depends on the distributions of factor quality. At any given point in the profile, i.e. at any given quantile, the increase in surplus is shared between the factors in proportion to their contributions to the increase at that quantile.

Because of the continuity assumptions, the factor owners don't earn rents over their next best opportunity within the industry. In a discrete model there would be a match-specific rent left for bargaining, as the difference in the pay of two "neighboring" individuals could be anywhere between the differences of their firms' willingness to pay for the ability difference between them. In a continuous model there is nothing to be bargained over because all units have arbitrarily close competitors (there would be match-specific rents only if both factor profiles had a jump at the exact same quantile).

2.2.2 Adjustable Factors

The distributions of factor qualities are exogenous in the model, in which the factor incomes are determined in a spot market, and depend on the distributions of factor qualities. The model allows adjustable factors, but they don't have to show up explicitly, if their levels are assumed to be chosen optimally so that they are just functions of the exogenous factor qualities. The output as defined so far is then really the surplus from the match or output net of the cost of adjustable factors. Denoting the adjustable factors by $\mathbf{k}$, the (net) production

function is

$$Y(a,b) = \max_{\mathbf{k} \geq \mathbf{0}} \left\{ \tilde{Y}(a,b,\mathbf{k}) - c(a,b,\mathbf{k}) \right\}, \tag{2.9}$$

where $\tilde{Y}$ is gross output. For the production function to be complementary, it is sufficient that all factors, including the adjustable ones, are complements in the surplus maximization problem. Factors that are purchased at constant unit cost must have decreasing returns to scale for a finite maximizer to exist.

The adjustable factors could further be divided into those that can be adjusted instantly or in the short run, such as effort, and those that must be chosen before the matching takes place and can only be adjusted in the long run, such as education. However, this distinction is not necessary when there is no uncertainty about the matching partner, because then everyone just chooses the optimal level of investment into long-run adjustable factors. There is no room for strategic behavior, such as firms or individuals threatening to invest suboptimally, because there are no match-specific rents.[9]

Some factors that can be adjusted in the short run may be "third party" inputs, such as raw materials and labor, that are purchased on the spot market. Other adjustable factors are embedded in one of the factor units, making it likely to confound the cost of these adjustable factors with the true compensation or rent of the fixed factor quality. This confusion is the reason for having to think about them. For example, education is clearly inseparable from a particular individual. Disentangling the cost of education from a rent to talent is of course a classic identification problem in economics.

For a simple case, which will be used later in the empirical part of the chapter, suppose that firms can rent capital at constant unit cost r and that the gross production function is $\tilde{Y}(a,b,k) = abk^{\theta}$. The (net) production function gives the surplus that will be divided between the fixed factors according to the assignment model.

$$Y(a,b) \equiv \max_{k} \left\{ abk^{\theta} - rk \right\}. \tag{2.10}$$

The optimal capital level is $k^*(a,b) = \left(\frac{\theta}{r}ab\right)^{\frac{1}{1-\theta}}$, which results in the following multiplica-

[9] With uncertainty about the quality of the match, the optimal pre-matching investment would depend on the whole distribution of possible matches.

tively separable closed-form production function

$$Y(a,b) = (1-\theta)\left(\frac{\theta}{r}\right)^{\frac{\theta}{1-\theta}}(ab)^{\frac{1}{1-\theta}}. \tag{2.11}$$

Adjustable capital is complementary with both a and b, so the production function has the complementarity property. A solution is guaranteed by decreasing returns, $\theta \in [0,1)$. The factor incomes of a and b are again determined from equations (2.7) and (2.8), with (2.11) serving as the production function. The payment to adjustable capital is $rk^*(a,b)$, and is likely to be mixed with the "factor income" of b, which could be an economic profit or a payment for previously sunk capital (more about this in the next section).

For the division of surplus, it does not matter into which factor the adjustable factors are physically or legally embedded in. For example, if higher effort levels increase output mainly by increasing the marginal productivity of b (if $\tilde{Y}_{be} >> \tilde{Y}_{ae}$), and there is sufficient variation in b, then the economic return from higher effort levels goes mainly to firms, while individuals are mostly just reimbursed for the cost of effort (i.e. its "reservation price"). The possibility to adjust the levels of some factors affects the division of surplus by affecting the productivity of the fixed factors, regardless of whether these are apparently "controlled" by one party or another.

It is straightforward to add more variables, as long as the complementarity assumption is not violated. Costly effort (or education) e can be added to the maximization problem, here in a way that interacts the cost with ability.

$$Y(a,b) \equiv \max_{e,k\geq 0}\left\{\tilde{Y}(a,b,e,k) - c(e,a) - rk\right\}. \tag{2.12}$$

The payments to the owners of the firm include the cost of adjustable capital. Similarly, a part of the wage is a compensation for the cost of effort $c(e^*(a,b),a)$, and the remainder is a payment to a scarce ability or talent.

The effort-cost function can serve as a simple reduced-form way to incorporate the ramifications of informational asymmetries between owners and managers. As is well known, the surplus-maximizing actions by the manager are not achievable under a wide range of circumstances. A partial alleviation of this problem is possible but costly, requiring individual compensation to be made in a form that the individual values below its market

price. This is a waste of money, compared to a perfect information, complete contract world, and should be included in the effort cost. An innovation in contracting technology could correspond to a downward shift in $c(\cdot, a)$: with better contracts, individuals of any given ability supply more effort at any given cost (or more realistically, effort is more accurately directed towards increased surplus). The distributional effects of such an innovation are not a priori clear, but depend on the strength of complementarities between effort, ability and firm size.

2.2.3 The Determinants of Firm Size Differences

The inalienability of ability is very natural, because it so palpably can not be moved from one person to another. But what are the fixed firm-specific characteristics that can't be chopped into pieces and shuffled between firms? For there to be many exogenously heterogeneous firms, there must also be some productive resources that are indivisible in nature and of heterogeneous quality, as well as some limitations for combining these resources under the management of a single individual. In other words, decreasing returns to scale are required on two fronts: in allocating more adjustable (divisible) resources into one firm, and in merging more fixed (indivisible) resources under the management of one individual. Without decreasing returns all resources complementary to talent should be allocated to the most talented individual.

There are inescapable decreasing returns to management ability stemming from the scarce attention of individuals, who must specialize their ability to some extent. Different firms do different things. Each firm operates in a slightly different market niche, and one individual can only be up-to-date about a limited number of niches at a time, at least sufficiently enough to manage a firm that operates in them. The abilities of the managers are of course themselves determinants of the size of the market where the firms operate; by size of niche I mean the exogenous component in the determinants of size, inherent in consumers' preferences and technology. Even if all managers were exactly equal in ability, there would be vast size differences between firms. Manufacturing of wide-body aircraft is going to be a bigger business than building yachts, and probably separate from it, under most circumstances.

The exogenous component in the size of the firm is the fixed firm-specific variable with which managerial ability is complementary. Managerial ability makes a larger absolute

difference to surplus in a firm that occupies a larger niche. These two factors together result in a scale-of-operations effect and differential rents. The possibility to adjust the levels of other variables, such as capital and labor, may further enhance this effect.[10]

As the most extreme case of loss-of-focus, any individual could only specialize in managing operations in a market for only one particular variety of goods at any given time, and the goods differ by their demand curves. At the other extreme, it could be that product varieties of equal demand shares are divided into bundles of markets, within which the same specialized managerial knowledge applies. A large niche then stands for a large bundle of varieties that are feasible to combine efficiently under one management. Either way, firms inhabiting the largest niches (i.e. those with largest b) hire the best managers. There appears to be a rent for being a firm in an attractive high-demand niche, but this expected rent should have been dissipated back when it was decided who got to enter that niche (perhaps in a patent race, or through premature entry, or as a rent to a talented founder).

The explicit inclusion of adjustable factors adds this assignment model some of the flavor of the models of Lucas (1978) and Rosen (1982), where one heterogeneous factor, namely managers, gives rise to a distribution of firm size and rents to management ability. In these models all differences between firms arise from the ability of their manager. The manager in effect sets up the firm, by renting the capital and hiring the subordinates, pays their market price, and then claims the residual as his own compensation. The size distribution of firms is solely a reflection of the economy's solution to allocating productive resources to different managers: if all individuals were equally apt, then all firms would also be identical. The determination of CEO pay in these models is akin to the previous case with adjustable capital, but with a degenerate distribution of b.

2.2.4 Understanding the Assignment Model

A central feature to understand about the assignment model is that fixed unit-specific characteristics are essentially ordinal. Any increasing transformation of "the scale of measurement" for a factor quality, combined with the inverse change in the functional form of the production function, changes nothing of substance in the model. This means, for example, that using a Cobb-Douglas form $Y(a,b) = Aa^{\theta}b^{1-\theta}$, as opposed to a simple mul-

[10] A bull market could be interpreted as an across-the-board increase in the exogeneous components of firm size (which is measured in market value). CEO pay levels should then be procyclical (and in apparent defiance of relative performance evaluation), as has been pointed out by Himmelberg and Hubbard (2000).

tiplicative $y = ab$, would be superfluous, or even misleading if it causes one to believe that the income shares should have a tendency to be related to the exponents. This is a special case of a more general mistake of assuming that factors are paid their marginal products, in a situation where the amounts of factors can not be shifted across different units of production. This transferability of factors across units is what pins down the linear scale of measurement for a factor quality in the usual case: the sum of the factors in the whole economy must adhere to some budget constraint. In an assignment model there is less flexibility. The collection of factor units is what it is, and the economic problem is how to combine these factor units into units of production.

The quality of factor units can not be measured in dollars or units of output, because there is another fixed factor that the units must be combined with for there to be any output.[11] What can be defined in dollars is the difference that two factor units make, when matched with a particular type of a counterpart. For example, $Y(a_2, b) - Y(a_1, b)$ is the value of the ability difference between individuals of abilities a_2 and a_1, if matched with a firm of type b. This difference is increasing in b for $a_2 > a_1$ by the assumption of complementarity.

The Quantile Scale Since the qualities of fixed factors are essentially ordinal variables, any increasing scale of measurement for them can be chosen by adjusting the production function accordingly. With the distributions of factor qualities fixed, the most convenient scale is obtained by using the quantiles as the measures of factor quality. This amounts to using the CDFs of the qualities as the positive monotone transformation allowed by the model. Defined this way, the production function $Y(i, j)$ gives the output from matching an individual in the ith quantile with a firm in the jth quantile. This choice of units is well suited for illustrating the effects of changes in technology.

The unit square in Figure 2-1 covers all possible matches, with the quantiles of the two factors as the coordinates. The production function defines a surface over the quantile pairs, the height of which is the level of output. This height is $w_0 + \pi_0$ at the origin, when the size of the industry is limited by the participation constraint of the lowest types.[12] The

[11] The only exception is the trivial case when the effects of different factors do not interact, i.e. when they are additive. In this case there would be no reason to expect assortative matching or correlation of factor incomes.

[12] There could be inactive types at negative "quantiles" outside the figure, for which the output with the equilibrium match does not cover the opportunity costs.

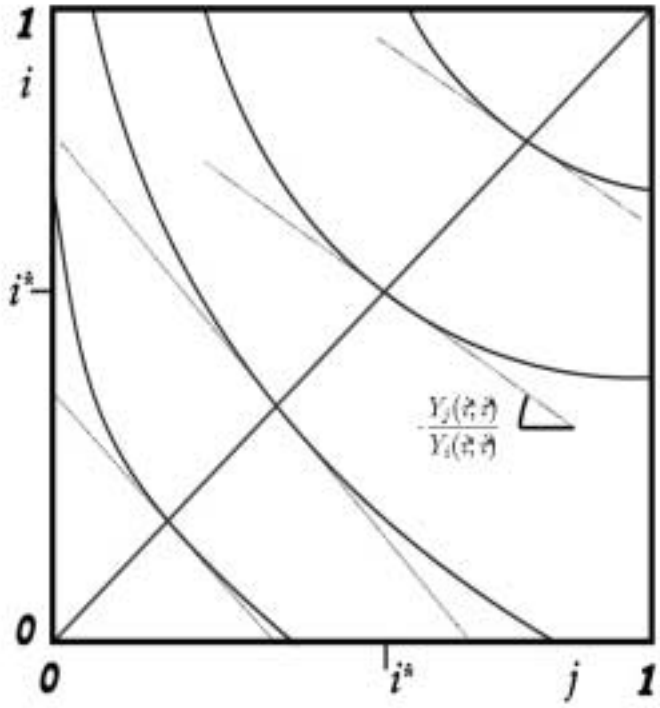

Figure 2-1: The isoquants of the production function on a quantile scale.

isoquants (contour lines) represent the combinations of individual and firm quantiles that would result in the same level of output. With this choice of units, both factor qualities are distributed uniformly in $[0, 1]$. The pairs that match in equilibrium are located on the 45-degree line, and there is a uniform mass of pairs on every point on the line.

The division of surplus at any quantile—at any point on the matching line—depends on the slopes of all the isoquants at the points of intersection below. The slope of the wage profile (2.6) is now simply $w'[i] = \frac{\partial}{\partial i} Y(i,j)|_{j=i} \equiv Y_i(i,i)$. The proportion of wages of the increase in output along the profile is

$$\frac{w'[i]}{y'[i]} = \frac{Y_i(i,i)}{Y_i(i,i) + Y_j(i,i)} = \frac{1}{1 + \frac{Y_j(i,i)}{Y_i(i,i)}}. \tag{2.13}$$

The ratio in the denominator is the slope of the isoquant, evaluated at the matching line. It is the marginal rate of substitution between quantiles of individuals and firms in the following sense. Consider a pair (i, j) that would produce an output $Y(i, j)$ if matched. If you wanted to form a pair that produces the same output as the pair (i, j), but with a

lower ranked individual, then it would need a higher-ranked firm to match with, and this MRS tells you the local trade-off in terms of quantiles for attaining the same level of output. With individuals on the vertical axes, steeper sloped isoquants mean that individuals make a smaller difference to output at the margin. In other words, individuals are more substitutable for other individuals at the margin, so the MRS (defined with Y_i in the denominator) is lower. This results in a larger denominator for (2.13), and a share w'/y' that is closer to zero.

The marginal rate of substitution between firms and individuals determines the division of surplus, but it only matters along the graph of equilibrium matching. Even large changes in (potential) output off the equilibrium matching graph have no effect on factor incomes, unless they become large enough to break the complementarity. On the contrary, small changes in technology that change the slopes of the isoquants near the 45-degree line can have a large cumulative impact on the division of surplus above, even if equilibrium output is unaffected (picture the isoquants rotating around the 45-degree line).

The Multiplicative case A simple and intuitively comprehendible specification that is very suitable for graphical illustration is the multiplicatively separable production function, of which Cobb-Douglas is a special case. The graphical convenience of multiplicativity comes from the simple fact that the level of output from matching a and b is the rectangle between ab and the origin. Therefore a basically three-dimensional problem can be illustrated in two dimensions. The matching graph $a = \varphi(b)$, defined by $\{(a, b) \text{ st } F_a(a) = F_b(b)\}$, can be any strictly increasing curve. Its slope is

$$\varphi'(b) = a'[F_b(b)]f_b(b) = \left.\frac{a'[i]}{b'[i]}\right|_{i=F_b(b)}. \tag{2.14}$$

Note, however, that the matching graph alone does not tell how the mass of pairs is distributed on top of it, only that there is a positive mass. Changes in the distributions of factor qualities appear as changes in the shape of the matching graph. For example, if individuals become more able, in the sense of first-order stochastic dominance, then the matching graph shifts up.

The area of the smaller rectangle in Figure 2-2 is the break-even level of output, $y[0] = a[0]b[0]$, that just covers the reservation prices of the factors. The division of this minimum

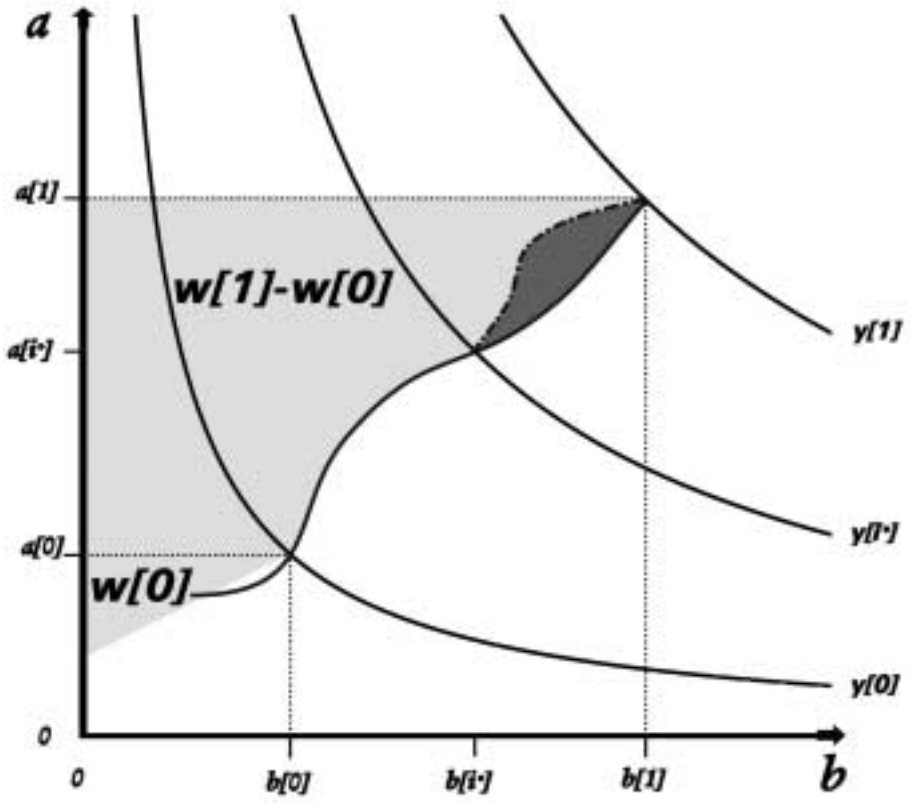

Figure 2-2: The multiplicative case.

output is exogenous; here the shaded triangle represents the reservation pay of individuals $w[0]$. Inframarginal types $i > 0$ create a surplus $y[i] - y[0]$, whose division depends on the distributions of a and b in a very simple way: the surplus of individual i is represented by the area between $a[i]$ and $a[0]$ and to the left of the matching graph. While moving up the matching graph (so also up the profile), the size of the rectangle representing output increases. The contribution of a higher a to this increase is proportional to the horizontal side of the rectangle, which is $b = \varphi^{-1}(a)$.[13] Conversely, the marginal productivity of b is $\varphi(b)$.

[13]The division of surplus can also be deduced by changing the variable of integration in the wage equation (2.7) from quantile j to ability a. Then $j(a) = F_a(a)$, $\frac{\mathrm{d}j}{\mathrm{d}a} = \frac{1}{a'[j]}$ and $b[j(a)] = \varphi^{-1}(a)$. This results in

$$w[i] - w[0] = \int_0^i a'[j]b[j]\mathrm{d}j = \int_{a[0]}^{a[i]} \varphi^{-1}(a)\mathrm{d}a, \tag{2.15}$$

which is indeed the area of the shaded region between $a[i]$ and $a[0]$.

The wage of any type of an individual can be read off the graph in a similar matter; the entire shaded region represents the wage of the highest type $a[1]$. Note again that wage is not merely a function of ability, but depends on the whole distribution of a and b below. The distributions of ability and wages are not seen in the figure: the matching graph gives a relation between distributions of a and b, but it is consistent with many combinations of distributions. One could imagine another dimension above the graph describing the density of individual-firm pairs; no restrictions are required on the shape of this density.

To illustrate the model's nonstandard implications, it is useful to do a comparative statics exercise with the distribution of abilities as the variable. Suppose that the ability of individuals between quantiles i^* and 1 is increased, while the qualities of firms and other individuals are unchanged. The new matching graph is shown by the line above the smaller shaded region in the figure. The distribution of b is not changed, so the quantiles move vertically: the ith quantile is matching exactly above where it used to. It can be seen that the pay levels of highest type individuals must go down, even though the amount of surplus produced is up at every firm in $(i^*, 1)$, and unchanged elsewhere. The loss in the pay of the very highest type $a[1]$ is the entire dark shaded region.

The income of the lower range of the improved quantiles goes up as a result of the change, as might be expected. Individuals gain from increased productivity, but can also lose due to tougher competition from other individuals below. The income for all ability levels $a[i]$ for $i > i^*$ is reduced by the amount of the dark shaded region between $a[i]$ and $a[i^*]$, the effect of tougher competition from below, and gain by the amount of the light shaded region between $a[i]$ and $a[i^*]$, the effect of increased ability. From this it already follows that the highest types must be worse off, since all they get is this loss. Of course, if everyone's ability is increased sufficiently, then everyone can also be better off. For this it would be necessary for the highest level of ability to increase enough to retain a sufficient relative advantage over its lower-ability competitors. Inspection of the wage equation (2.7) reveals that a sufficient condition for everyone's pay to increase is that the slope of the ability profile should increase at every quantile.

From the point of view of the other factor, the gains are unambiguous: all firms of types $b[i^*]$ or above are better off than before. The dark region to the left of $b[i]$, for any $i \in (i^*, 1]$, is the resulting gain for firm i. The converse result holds if a section of firms became inherently "more productive", i.e. experienced an increase in b. Individuals of lower

ability would feel no "trickle-down" effect from increased productivity at the higher-level firms, but instead there would be a trickle-up effect. High-ability individuals gain whether the level of output at their firm is increased or not, because the value of ability at lower-ranked firms has been increased, shifting the division of surplus to individual's favor.[14]

These comparative statics results can be summed up in terms of first order stochastic dominance in an interval that excludes the maximum. If the new distribution of ability dominates its old distribution, and the distribution of firm size is held fixed, then the new distribution of pay levels does *not* dominate its old distribution, whereas the new distribution of profits does dominate its old distribution (vice versa for a change in firm size distribution).

Finally, suppose that there is a general increase in productivity, as if a multiplicative parameter in the production function were to increase.[15] This merely changes the labels on the isoquants, but has no effect on the matching graph or the slopes of the isoquants, and therefore can not affect the division of surplus, except by potentially changing the size of the industry. If the quality of either of the factors increased proportionally all across the distribution, then the gains from this improvement would be shared by both factors in proportion to their current shares of the surplus. To be exact, this neutrality would require the outside opportunities to be increased in the same proportion as productivity within the industry; if not, then the size of the exogenously divided break-even output is decreased in the figure and some previously inactive firm-individual pairs will enter. The division of their surplus depends on the shape of the matching graph inside the old break-even rectangle.[16]

This example also illustrates why merely studying an earnings function can be misleading. Equilibrium relations such as $w(a)$, or equivalently $w(b)$, depend on the distributions

[14]If the price of output (so far normalized to one) was decreasing in total industry output, then lower-ranked individuals would actually be worse off, and the industry would contract in size (employment) as the least profitable firms could no longer break even at the reduced output price.

[15]In the multiplicatively separable case, so including Cobb-Douglas, an increase in general productivity is, by its effects, indistinguishable from a factor augmenting advance in technology: neither changes the substitutability between factor units. Both types of changes cancel out from (2.13).

[16]For a completely different example, suppose that the factors of production are a population of breeding studs and mares and that variables a and b fully describe their genetic qualities. Surmise then that each beast can only be used in a limited number of breedings each period where a foal of expected present value ab is produced (to be used in horse-racing or something). It would make no sense to say that either the studs or the mares contribute more to the price of foals, since they are both needed for there to be any offspring at all. In the price of any foal the contribution of the stud and the mare can only be evaluated in relation to what the prices would have been had they been coupled with other studs and mares in the population. In general, if one factor of production is very homogenous in quality then most of the variation in difference made to output and in factor incomes occurs among the heterogenous factor.

of a and b. Even if ability and the earnings function were observed directly, it would give the wrong predictions about (even the signs of) the changes in earnings, if more than a zero measure of individuals were to change in ability. It is arguably more sensible to think of the distributions of factor incomes as the primary variables of interest, rather than a functional relationship between factor incomes and factor qualities, since the latter are measured on arbitrary scale.

Marginal Productivity Redefined It is often claimed that wages are not equal to marginal productivity when the economy faces an assignment problem.[17] It would indeed be misleading to say that factors earn their marginal productivity by the usual definition of marginal productivity. The marginal productivity of the ability of an individual of ability a is $Y_a(a, \varphi^{-1}(a))$. This is the marginal increase in output if he were to increase in ability, while still matched with a firm of type $b = \varphi^{-1}(a)$. But if he were to increase in ability, then he would also move up in the ranking and be matched with a higher b. Moreover, whenever someone moves up in the rankings, someone else must move down and experience a decrease in productivity. The assignment model is needed to keep account of the changes that are caused by the rearrangement of individuals, when distributions of ability or firm size change. Interestingly it seems to have escaped the attention that the "differential rents" assignment models (including our model) satisfy "the No-Surplus Condition" of Ostroy (1980, 1984) which is an alternative definition for a perfectly competitive equilibrium. This means that individuals actually do receive their marginal product, if the margin is defined with respect to individuals.

Wage is not in general equal to the marginal productivity of ability because changing someone's ability is not a true economic margin—ability can not conceivably be extracted from one individual and poured into another. The true margin in a market with an assignment problem is whether a given individual (or other factor unit) will participate in the industry or not. By this definition, factors do indeed earn their marginal product as rents over their outside opportunity. Equation (2.7) gives the decrease in total industry output, taking into account the resulting reassignment of firms and other individuals, if individual i were to leave the industry. In the absence of individual i, all firms in $[0, i]$ would then have

[17] For example, according to Sattinger (1993, p. 848) "Because of the fixed proportions technology, in which one worker can only be used in combination with one machine, the marginal products of workers and machines are not defined."

to match with a marginally lower talent than before.

To see this explicitly, suppose that individuals in quantiles $[i-\varepsilon, i]$ were to leave the industry. What would be the resulting loss in output per individual? Individuals below the quantile $i-\varepsilon$ would move up and match with a firm that is ε quantiles higher ranked than their previous match. The total change in output, divided by the mass of lost individuals, is then

$$\frac{\Delta Y}{\varepsilon} = \frac{1}{\varepsilon}\int_0^i \left(Y(a[j-\varepsilon], b[j]) - Y(a[j], b[j])\right) \mathrm{d}j. \tag{2.16}$$

The marginal product of a single individual is obtained by letting the mass of "disappearing" individuals go to zero.[18] This is mathematically the same derivation that was earlier used to obtain the wage profile (2.7). If a single individual is forced to quit the industry for some reason, then the resulting social loss is the difference between his wage and his outside income. This definition of marginal product could be used to derive the factor incomes in an assignment model to begin with, given that the efficient matching has first been solved for. There is also a distributional effect—a pecuniary externality—from the disappearance of individual i. It shifts the division of surplus to individuals' favor.

One striking feature of this model industry is that factor owners are only affected by changes in the quality of those below them in the rankings. Mathematically this is obvious from the fact that the equations for factor income profiles take the form of integrals over the profiles below. Intuitively, the binding constraint on any factor owner is the quality and price of their next best competitor. If the next best competing factor unit becomes less competitive, then one can raise the price a little bit, and this price increase spills over along the whole profile above by shifting the division of surplus.

2.3 Applying the Assignment Model

When should the assignment model be used? The basic requirement is that there is a market where some factors of production are embedded in discrete units, and where the units of one factor could in principle match with any unit of another factor. If the observed matching is consistent with assortative matching, then a particularly simple type of an assignment

[18] Whether the lowest firms in $[0, \varepsilon]$ hire the best individuals outside the industry (with a negative quantile $[-\varepsilon, 0]$), or exit the industry, won't matter at the limit. Technically, the new level of output should be $\max\{w_0 + \pi_0, Y(a[j-\varepsilon], b[j])\}$, where the outside option will be higher for the very lowest pairs.

model can be applied. In the absence of assortative matching a more general assignment model could still be relevant, though probably less instructive.[19]

The strong positive relation of CEO pay and (any definition of) firm size suggests that something like assortative matching may be going on. The CEOs and companies are matched one-for-one, so if scarce management ability tends to be more valuable in larger firms and firm size differences are not purely explained by differences in managerial ability, then the level of pay can not be expected to equal marginal productivity in the standard sense. Yet, like any model, this assignment model captures only some aspects of reality. The assumption of perfect sorting leaves out frictions and idiosyncratic characteristics of individuals and firms, which are surely important in practice.

The properly defined marginal product implicit in the pay level provides an answer to a very simple counterfactual: what is the effect of this one person on industry output. So one could say that in a competitive market CEO pay levels only tell us what is their marginal product, end of story. More interestingly, we can use factor income data to assess the effects of changing a whole distribution. The model makes it possible to run thought experiments with changes in exogenous variables, and data can be used to bring actual numbers into these experiments. However, doing this requires making some further assumptions. By making a functional form assumption about the relation of ability and other variables, and plugging in real data, we can get rough answers to several interesting quantitative questions.

While the rent accruing to a single individual is equal to the extra value that he brings to the industry, the sum of the rents of all individuals understates the value that they all add together. The pay and the marginal product of an individual are defined holding the characteristics of other all individuals as given; if more than one individual leaves the industry, then the possibilities to counter some of this loss by reassigning the other individuals are diminished, and total output is reduced even more.

If we ask what is the value of ability of all current managers in the economy, then we are actually considering replacing the existing distribution of management ability with some other distribution. Replacing the CEOs by no one at all is not a sensible counterfactual; even if a company had no one by that title, someone would still have to make those decisions.

The distribution of abilities of the replacement CEOs should be somewhere below the

[19]The seminal assignment model of Koopmans and Beckmann (1957) considers a general problem for matching plants and locations in a linear programming framework. See also Sattinger (1984).

current lowest type manager. Since there is no way to estimate the relative abilities of out-of-sample CEOs, I will use the lowest type sample CEO as the hypothetical replacement type. The counterfactual from the title of the chapter, the difference that CEOs make, refers to the social loss from replacing all CEOs in the largest 1000 firms by a type expected to be found at the 1000th largest. In terms of the model, this difference is $\int_0^1 \left(Y(a[i], b[i]) - Y\left(a[0], b[i]\right)\right)\mathrm{d}i$.

On a more positive note, we can consider a change where all CEOs become as good as the current highest ability CEO, i.e. we can use $a[1]$ for all i as the counterfactual distribution. A third simple counterfactual is random matching of CEOs and firms; in this case the expected ability of managers would be equal to the current average ability at every firm. The loss in total output that would be caused by a switch to random matching is the social value of assortative matching among the top 1000 firms.

2.3.1 Inferring Factor Profiles

The assignment model shows how the distributions of factor qualities determine the equilibrium distributions of output and factor incomes. If the profiles of both factor incomes are observed, then the model can be used reversely to infer the profiles of both factor qualities, up to a constant. Estimation of $Y(a, b)$ is pretty much out of the reach in the case of CEOs, but we can proceed less ambitiously by assuming a plausible and often used functional form and by using the observed factor incomes as the data. The distributions of factor quality alone are meaningless—for reasons discussed before—but their differences together with the production function will allow us to answer various counterfactuals about the effects of CEO ability on social surplus and the distribution of CEO pay.

The basic idea for inferring factor profiles comes from the observation that the slopes of the equilibrium factor income profiles and the break-even level of output form a system of two differential equations with a boundary condition.

$$\begin{aligned} w'[i] &= Y_a(a[i], b[i])a'[i] \\ \pi'[i] &= Y_b(a[i], b[i])b'[i] \qquad i \in [0, 1] \\ w[0] + \pi[0] &= Y(a[0], b[0]) \end{aligned} \tag{2.17}$$

Using the observed factor income profiles $w[i]$ and $\pi[i]$, the profiles of factor qualities can be solved from (2.17), up to constants of integration. In general, it includes a pair of

nonlinear differential equations without a closed-form solution, so numerical methods may be required. However, assuming a multiplicatively separable production function allows for the profiles of a and b to be solved directly, up to multiplicative constants. These constants conveniently wash out of the predicted economic effects of hypothetical rearrangements or changes in the qualities of individuals and firms.

The production function that will be used here allows for the level of adjustable capital to depend on the ability of the CEO, as was discussed before in section 2.2.2. Restated at its most general form, this production function is

$$Y(a,b) = \max_k \left\{ cg(a)h(b)k^\theta - rk \right\}, \tag{2.18}$$

where c is a positive constant, and g and h are positive increasing functions, to be chosen at convenience. This choice will of course affect the scale of measurement for a and b, but these are just nuisance parameters anyway. We are not interested in relations like $w(a)$, but in economic counterfactuals like $Y(a[i], b[k]) - Y(a[j], b[k])$, and these are invariant to the choice of c,g and h.

The closed-form solution of (2.18) is a multiplicative production function, as seen before in (2.11). The most convenient choice is to set $c^{-1} = \left(\frac{\theta}{r}\right)^\theta (1-\theta)^{\frac{1}{1-\theta}}$, $g(a) = a^{1-\theta}$, and $h(b) = b^{1-\theta}$. This choice yields the simple multiplicative production function $Y(a,b) = ab$, as the solution of (2.18). The parameters θ and r can affect the results through their effect on the interpretation of data. The elasticity of gross output to capital, θ, is the share of adjustable capital in gross output. The division of a firm's (shareholders') income into a rent to the fixed factor b and a cost of adjustable capital is implicit in the assumed θ. It is straightforward to do the calculations for different assumptions of $\theta \in [0, 1)$; for any question this gives two bounds for the economic effects of CEO ability.[20] Fortunately these bounds don't turn out to be too wide to be informative. On the other hand, questions about the effects of changes on the rent to b are not tenable, because these answers are completely sensitive to the assumed value of θ.

With this simplification, it is now easy to infer the distributions of relative levels of

[20]The share of adjustable capital θ cannot be arbitrarily close to one, because observed CEO pay is the lower bound of the surplus for the fixed factors. It is not sensible to assume a θ higher than the lowest observed share of shareholder income in gross surplus, this share is about 0.96 in the data.

factor quality from observed factor incomes. The system (2.17) becomes

$$\begin{aligned} w'[i] &= a'[i]b[i] \\ \pi'[i] &= a[i]b'[i] \qquad i \in [0,1]. \\ w[0] + \pi[0] &= a[0]b[0] \end{aligned} \tag{2.19}$$

Dividing the slope of the pay profile by the profile of surplus, $y[i] = w[i] + \pi[i]$, gives the rate of increase of the ability profile.

$$\frac{w'[i]}{y[i]} = \frac{a'[i]b[i]}{a[i]b[i]} = \frac{a'[i]}{a[i]}. \tag{2.20}$$

The other unobserved factor b cancels out, and relative abilities can be solved by integrating the resulting equation. This leaves a multiplicative constant of integration, which is the undetermined baseline ability level $a[0]$.

$$\tilde{a}[i] \equiv \frac{a[i]}{a[0]} = \exp\left\{\int_0^i \frac{w'[j]}{y[j]}\mathrm{d}j\right\} \tag{2.21}$$

The relative ability of two individuals, $a[k]/a[j] = \exp\{\int_j^k \frac{w'[j]}{y[j]}\mathrm{d}j\}$, gives the ratio of surplus in any given firm in case it was matched with an individual ranked k or j respectively. The other factor profile, $\tilde{b}[i] \equiv b[i]/b[0]$, can be recovered in the same way, again leaving an undetermined multiplicative constant.

The most general type of a counterfactual we can answer is, how much difference does it make to economic surplus at firm of quantile k if it were managed by an individual of quantile i as opposed to quantile j?

$$Y(a[i], b[k]) - Y(a[j], b[k]) = b[k]\left(a[i] - a[j]\right) = y[0]\tilde{b}[k]\left(\tilde{a}[i] - \tilde{a}[j]\right). \tag{2.22}$$

The last form can be calculated because it includes only the inferred relative factor qualities and the observed baseline output. The counterfactuals discussed above can be constructed from special cases of the form (2.22).

2.3.2 Data

The sample comprises the 1000 largest publicly traded US companies in the ExecuComp database in 1999, provided by CompuStat. The variable for executive pay is taken from

CompuStat, where the options are priced using the standard Black-Scholes formula. If firms are risk neutral, then they are willing to pay in expectation the competitive price for an individual's expected ability. However, individuals value risky contingent pay substantially below its cost, as is well known. The difference between the market cost of compensation and its value to the CEO is a part of the cost of effort (defined in a broad sense) that could be avoided in a perfect information world. Here this cost is subsumed under the level of CEO pay.

It is crucial to distinguish between financial and economic returns to CEO ability. While a more able executive is expected to produce more economic surplus with given resources, there can be no excess return to securities of companies that in equilibrium employ better executives. The effects of superior CEO ability must be included in the current market value. If the current year CEO of firm i turned out to be worse than expected, say of baseline ability, then $Y(a[i], b[i]) - Y(a[0], b[i])$ would be the expected social loss to be borne by the current shareholders, and possibly partly by the CEO himself through contingent compensation. A CEO of expected ability is just expected to maintain the market value on its expected path.

The output from a matched pair is the expected joint income, i.e. the combined income of the CEO and the shareholders. In one year, this income, *gross* of the cost of adjustable capital, is $w + rv$, where w is the expected cost of CEO pay, v is market value, and r is the expected rate of return. Shareholder income includes both the CEOs effect on current profits and on discounted future profits. Any income going to other parties, such as employees and suppliers, has already been deducted at this point. After paying for the adjustable capital, the surplus to be shared between the fixed factors is $(1-\theta)(rv+w)$, which leaves $\pi = (1-\theta)rv - \theta w$ as the factor income of b. This implies that b could be interpreted as a type of unduplicatable (possibly intangible or sunk) capital k_b, defined as the "hidden capital" from $v - k^* = k_b$, where k^* is the optimal level of adjustable capital.

The prerequisite for the whole exercise to make sense is that CEO pay is increasing in firm size. The relationship of CEO compensation and ranking by market value is shown in Figure 2-3. It bears out the well known fact that larger companies pay more to their CEOs. Sample correlations and descriptive statistics are listed in Tables 2.2 and 2.3. Market value is the best explanatory variable for CEO pay in the data. The highest correlations of CEO pay are with market value, 0.51 and 0.56 for logs and ranks respectively. The profile of

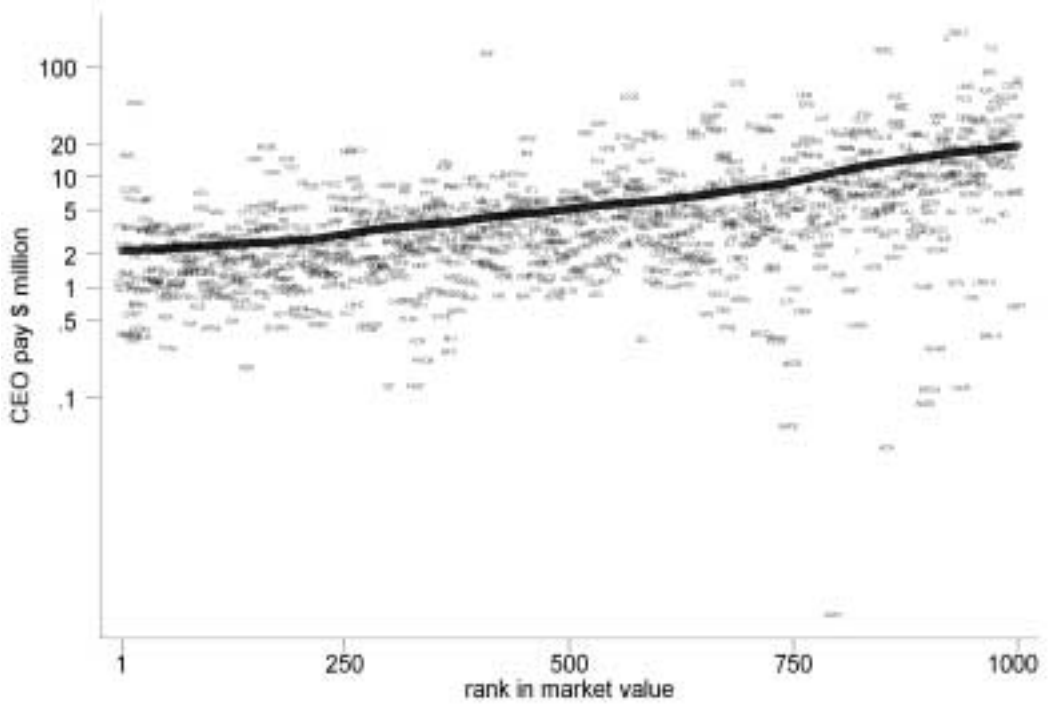

Figure 2-3: The relation of market value and CEO pay among the largest 1000 publicly traded US companies in 1999. The pay profile is *not* smoothed through the log function, so it appears upward biased in the figure with log-scale.

market values is shown in Figure 2-4.

In reality, the magnitude of the potential economic impact of CEO ability in any given firm depends on several factors, and can not be expected to be perfectly rank correlated with market value. These other factors, as well as stochastic factors affecting contingent pay and the deviation of actual ability from what was expected at the time of matching, are reflected in the variation of realized pay levels. To be used in the model, the observed factor incomes must be fitted over some common order i, in which both fitted profiles are strictly increasing. The order i should be chosen to maximize some criterion involving the goodness-of-fit of the fitted profiles. The trade-off is that by choosing an ordering in which the ordered data of one variable deviates less from its strictly increasing fit, the deviations in the fit of other variables are likely to be increased. I use the order in market value, because it is likely to be better measured. The pay that is recorded in a given year does not necessarily compensate for the services in one calendar year, due to deferred pay and bonuses.

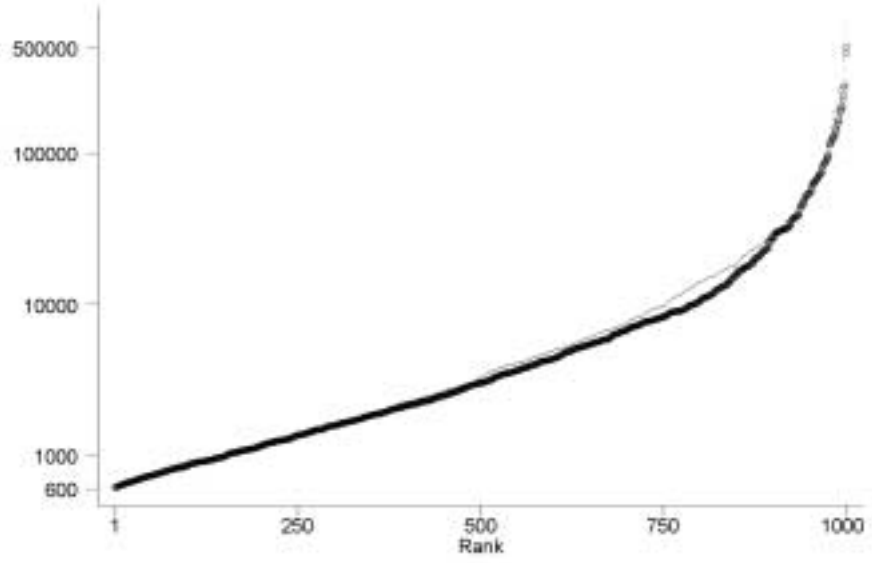

Figure 2-4: Largest U.S. firms by market value (thick line) and by asset value (thin line) in $ millions.

Variables	Mean	Std.Dev.	Min	Max
CEO Pay (m)	7.06	13.8	0	193.8
CEO Pay (a)	6.87	13.3	0	193.8
Market Value	12694	35489	613.8	508329
Assets	16062	52046	625.3	716937
Fitted CEO Pay (m)	6.93	5.14	2.12	21.0
Fitted CEO Pay (a)	6.78	3.86	2.77	16.0

Table 2.1: Descriptive statistics. (m) refers to sample with 1000 largest firms by market value, (a) by assets. The units are in $ millions.

In general, if we know that one variable is better measured than others, then it is better to use the ordering in the better measured variable, because its ordering is likely to be closer to the "true" ordering. Denote by $w(i)$ the observed noisy values of w, when (i) is the ranking of observations by observed π. Taking the average of $w(j)$, for j close to i, gives a better estimate for $w[i]$ then doing the converse. Estimating $\pi[j]$ from the ranking in w would result in an average of observations that are not near to each other in the true ranking, and the slope of the fitted profile of π would be seriously biased towards zero. As much as π is also measured with error, or includes omitted variables, the estimated wage profile will be biased towards a flat profile, causing a downwards bias in the estimated ability differences.

The fitting could be done in many ways. The simplest way to smooth the profiles would

logs						
CEO Pay	1					
Market Value	0.511	1				
Assets	0.471	0.652	1			
Sales	0.443	0.589	0.813	1		
Employees	0.376	0.515	0.648	0.870	1	
Fitted CEO Pay	0.508	0.974	0.645	0.569	0.498	1

Table 2.2: Sample correlations

ranks					
CEO Pay	1				
Market Value	0.559	1			
Assets	0.473	0.645	1		
Sales	0.492	0.608	0.810	1	
Employees	0.424	0.499	0.628	0.862	1

Table 2.3: Sample rank correlations

be to divide the observations into a histogram, and interpolate a function through the bin averages (and extrapolate at the edges). This is similar to what the kernel estimator does, the main difference being that a kernel estimator creates a smooth nonlinear fit.[21] To put it roughly, it takes a weighted moving average of CEO pay along the order by market value, using higher weights for nearby observations. The fitted pay profile is shown as the dark line in Figure 2-3. It begins at a baseline of $2.1 million, and reaches $21 million at the top quantile.

While it is necessary to obtain increasing profiles for the factor incomes, continuity is not essential. The model easily generalizes to a case with a discrete number of factor units. The main difference of the discrete model is a match-specific rent, which adds to a more complicated notation. Using either extreme assumption about the division of match-specific rents has an effect on the results that is within rounding error. The implied match-specific rents caused by the ability-size complementarity are just too small to matter.

2.3.3 Results: The Value of CEO Ability

It was outlined in section 2.3.1 how the relative factor profiles can be inferred from observed factor income profiles. The estimated profile of relative abilities is graphed in Figure 2-5.

[21] The Epanechnikov kernel was used, with window width 150. When observations are ranked by market value, the smallest window width which yields a strictly increasing fit for the pay profile is about 130; the results are virtually the same with window width varying between 130 and 200. When observations are ranked by the value of assets, then a bandwidth of 200 is needed for a strictly increasing fit.

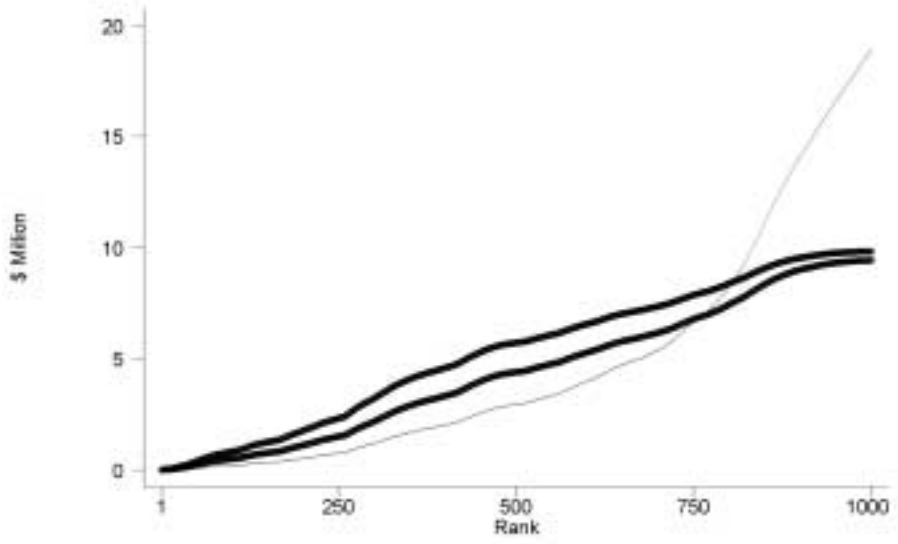

Figure 2-5: The difference that CEOs would make if all matched with a same type of a firm. The thick lines graph this difference, $Y(a[i], b[0.5]) - Y(a[0], b[0.5])$, evaluated at the median firm, for $\theta \in \{0, 0.967\}$. The thin line is the actual difference in pay levels $w[i] - w[0]$..

The question is, how much difference would management ability make, if all firms were of the same type. The dollar value of the answer depends on the size of the firm that is used as the point of evaluation. In the figure, the value of differences is evaluated at the median firm. If all firms were similar to the current median firm, then the pay difference between top and bottom CEOs would be about $10 million, compared to $19 million now. If, instead, ability differences were evaluated at the baseline firm size, then the advantage of the most able CEOs would be less than $2 million. These hypothetical differences in pay still include the scale-of-operations effect rising from the possibility to adjust the level of capital according to the manager's ability, but not the effect from the complementarity between ability and exogenous component of firm size.

Another way to evaluate the differences is to keep the distribution of firm types fixed at what it is, and see what is the value of current levels of ability compared to some counterfactual. The dollar difference in economic surplus from replacing the existing CEOs by some particular type are obtained by plugging in the inferred and counterfactual profiles into equation (2.22) , and summing over i. Table 2.4 lists the results for the conceptually possible values of θ (of which more in a moment). The first row is the motivation for the title of the chapter. It gives the total value of scarce ability of top 1000 CEOs, defined as

$r = 0.10$	$\theta = 0$	$\theta = 0.5$	$\theta = 0.75$	$\theta = 0.9$	$\theta = 0.95$	$\theta = 0.967$
Value of $a - a[0]$	36.9	36.3	35.2	32.3	28.1	24.8
Value of $a[1] - a$	3.6	3.6	3.7	3.8	4.1	4.5
Value of sorting	15.0	14.8	14.5	13.6	12.1	11.0
Value of $b - b[0]$	1212	605	302	120	59.2	39.6

Table 2.4: Results by assumed θ, in billions USD.

the difference that they make compared to the case where they were all replaced by lowest type individuals in the sample, i.e. of the type managing the smallest sample firm. This value is between \$25 and \$37 billion, depending on the assumed value of θ. For comparison, the actual total pay of the CEOs is \$7 billion, of which \$5 billion is "extra" accruing to the scarce ability. Since the baseline CEOs make about \$2 million per year, the cost of pay for all 1000 hypothetical replacement CEOs would be \$2 billion.

The second row is the optimistic counterfactual: the change in total surplus, if all top 1000 CEOs became as good as the current highest ability individuals. In this case the gains are much more modest than the losses in the previous grim scenario: the figure is between 3.6 and 4.5 billion. The gains from increased ability are relatively small, because the largest companies benefit the least: their managers are already nearly top ability. In this hypothetical case all CEOs would get paid only about \$4 million, which is baseline pay plus their excess value at the smallest firm. The increase in overall ability melts most of the pay advantage of the better CEOs. In total, the CEOs earn \$3 billion less under this (socially) optimistic scenario.

The third line gives the difference in total output compared to a case where the matching of individuals and firms is random. Under multiplicative separability, this is equivalent to assuming that all CEOs are replaced by an individual of average type. The value of sorting within the top 1000 firms and CEOs is estimated between \$11-15 billion.

In the fourth scenario the individual types are held fixed, but it is assumed that all firms become the same type as the current smallest sample firm, which has a market value of \$614 million. This is not a very useful counterfactual, because the result is totally sensitive to the assumed value of θ.

The role of the parameter θ is best understood by considering how an otherwise similar economy—with the same distributions of factor qualities—would differ depending on the value of θ. For higher values of this parameter, not only will levels of adjustable capital be higher all around, but higher-ability individuals can increase their advantage more. The

distributions of both factor incomes as well as capital levels are more skewed for higher values of θ. To interpret the effect of assumed θ on the empirical results, turn this idea around. With observed pay levels and shareholder income fixed, the inferred factor quality distributions will have to vary with θ. The higher we assume θ to be, the larger is the role of adjustable capital assumed to be in observed differences in market value and the smaller are the required ability differences needed to explain the observed CEO pay differences. After all, CEOs can only get paid extra for using capital and the firm's unique position in the economy to produce something that is worth more than the cost of that capital.

How could the optimal level of adjustable capital not depend on manager's ability, i.e. how could θ ever be zero? We could think of the production function as having the form

$$Y(a,b) = \max_{k \geq 0} \left\{ a \min\{b,k\} k^{\theta} - rk \right\}. \tag{2.23}$$

When capital is not binding in the Leontief part of the function, then this is exactly the previously seen multiplicative two-factor production function with adjustable capital. However, as $\theta \to 0$, it will start to bind and $k = b$ becomes optimal. In this case the gross production function is $\tilde{y} = ab$, but this is just another multiplicatively separable production function. The point is that even for $\theta = 0$, a large part of firm's income could still be a payment of adjustable capital, but the optimal level of that capital does not depend on the ability of the manager, only on the type of the firm.

What happens as θ is near one? It can not sensibly be interpreted as being arbitrarily close to one, because it is the share of adjustable capital in gross surplus ($y = w + rv$), and this can not be higher than the observed proportion of shareholder income in observed gross surplus. It is conceivable that all firms could be inherently similar after all, i.e. have the same b, but in this case all of the net surplus from the variation in capital levels across firms would have to accrue to the CEOs. All differences in total shareholder income between firms would just reflect the difference in the rental cost of adjustable capital. This would require the ratio of CEO pay to shareholder income to be equal at firms of all sizes, namely $\frac{1-\theta}{\theta}$, and CEO pay should be linear in market value. This is clearly not the case; the share of CEO pay is decreasing in firm size. At assumed values of θ close to one, the model can not sensibly interpret the data. The maximum sensible value of θ is the smallest observed share of shareholder income, which is 0.967 in the sample.

The dollar values are defined in flow terms, because the measured pay and shareholder income are for a period of one year. To convert the value of ability to stock terms (literally), these values must be multiplied by the inverse of the rate of return. This stock value of CEO ability includes the discounted value added by all future CEOs. For example, in the first scenario, the stock value of ability is the drop in market value if it was suddenly found out that all companies would have to match with a CEO of baseline ability from here to eternity.

The effect of assuming a different expected rate of return is small within reasonable limits for r. For example, assuming $r = 0.05$ halves the share of this period's expected shareholder income, and the effect of this period's CEO, in the market values. The relative shareholder incomes between firms are not affected by this, and the difference to relative surpluses is also small because CEO pay is a very small proportion of gross surplus. A replication of Table 2.4 would show that, for $r = 0.05$, the difference that CEOs make is between \$24.1 and \$35.9; and for $r = 0.15$ it ranges from \$24.6 to \$37.2.[22]

2.3.4 What about Estimation and Testing?

Data requirements for doing away with the functional form assumption and actually estimating the model are stringent. In the case of CEOs, it seems unreasonable that observables like education and experience could capture a significant part of what the market considers a good manager. Movements of CEOs between firms are of no help in uncovering magnitudes of unobservable qualities, because in the model these movements should just reflect changes in the ranking by expected ability.

Not only do the factor qualities have to be observable, but having any number of cross-sections alone would be inadequate even if all variables were measured without error. Positive sorting forces factor incomes and factor qualities to be perfectly collinear, on some increasing scale of measurement. In terms of Figure 2-2, all observations $\{a, b\}$ would lay on the matching graph. To estimate the factor income equations, we would need to find out the slopes of the isoquants in the region of equilibrium matching. But if we take the model seriously, then all deviation from perfect linear correlation is due to noise, frictions, and the chosen scales of measurement. To be able to identify the shapes of the isoquants, we would need to observe cross-sections at different points in time with sufficient variation

[22]The corresponding maximum sensible θs are 0.978 and 0.936 respectively.

in the distributions of factor qualities.

The availability of data might be more favorable in some other application. Suppose that we observed vectors of relevant unit-specific characteristics $\mathbf{x}$ and $\mathbf{z}$. The simplifying power of assortative matching relies on the assumption that there are one-dimensional sufficient statistics $a(\mathbf{x})$ and $b(\mathbf{z})$, describing the relative quality of factor units. These should be estimated as the first step in estimating the production function. The goal would be to find functions a and b that maximize some criterion involving the rank correlations between factor qualities and factor incomes. In other words, $a(\mathbf{x}_i) > a(\mathbf{x}_j)$ should be a good predictor that $b(\mathbf{z}_i) > b(\mathbf{z}_j)$, $w_i > w_j$, and $\pi_i > \pi_j$. Note that functions a and b are not restricted to be supermodular or even monotonic functions. Again, these functions are only defined up to a positive increasing transformation; the choice of transformation will be counteracted by the estimated production function.

Having, against all odds, estimated the production function, we could test the model by comparing the predicted factor income profiles with the actual. The model predicts a certain way for the surplus $(y[i] - y[0])$ to be divided between factor owners. The actual division of surplus into factor incomes is not used in the estimation of the production function, because the left hand side variable is their sum $y = w + \pi$. The model could be deemed successful if the predicted division of surplus into factor incomes, based on the observed factor qualities and the estimated production function, was in sufficient agreement with the actual division.

2.4 Conclusion

It goes without saying that the above estimates must be taken with a grain of salt: they don't test the model, they are based on the assumption that it is true. However, the fact that the observed relationship of firm size and CEO compensation is so strong gives hope that this exercise could give some insight into the magnitudes of underlying CEO ability differences and to the extent to which they can explain the observed differences in pay. The total economic value of the scarce ability of the top 1000 CEOs is estimated at about \$25-37 billion, depending on the assumed elasticity of output to capital. This is the difference they make to total economic surplus, compared to the counterfactual where all of them were only of the same ability level as the CEO of the smallest sample firm.

Since the economic difference that CEOs make is much larger than their total pay, most

of the value of top talent apparently goes to the shareholders. In light of the assignment model, this suggests that the differences in talent are a relatively small factor in determining the observed firm size distribution, compared to exogenous or predetermined firm-specific factors. All in all, the implied economic impact of differences in CEO ability seem quite small compared to the differences in the market values of companies.

There is a time-honored tradition in economics to assume that prices are competitive and reflect all available information, at least as the first approach in analyzing market data. In addition to giving a ballpark estimate for the difference that CEOs make, this chapter has explored how assignment models can be used in this spirit to analyze data from markets with positive assortative matching.[23] For future applications that would allow the assignment model to be tested, one would need to find a matching market with assortative matching, where the characteristics that determine the productivity of factor units are directly observed and not contaminated by the characteristics of the matching partner; furthermore the distribution of these characteristics would have to vary over time sufficiently for the production function to be estimated.

The next natural step in using the assignment model to analyze CEO pay would be an inclusion of frictions (such as switching costs), incentive problems, and the structure of pay.[24] A more realistic model would take into account the complementarity of both inherent ability and costly effort with firm size.

[23] See Teulings (1995) for a more ambitious general equilibrium approach.

[24] Shimer (2001) introduces coordination frictions to an assignment model (due to mixed strategies in the job application process); however this approach is probably not well applicable to the CEO market.

Chapter 3

Transfer Fees and Development of Talent

3.1 Introduction

The purpose of this chapter is to evaluate the role of transfer fees in professional sports. The motivation comes from the recent push to regulate, and possibly abolish, the transfer fee system in European football (soccer). A transfer fee is basically the price of a player's remaining contract, which an acquiring club may have to pay to the current employer. Like many economists (but unlike most pundits) who have discussed the system, I find that transfer fees serve an important allocational purpose. However, I believe the standard defense that transfer fees provide compensation for the cost of training is insufficient at best, and seems to be pointing the way to detrimental ways of tempering the ongoing regulation. The main point in this chapter is that transfer fees, far from being just scarcity rents to talent or compensation for training, are needed to efficiently allocate job positions among players of varying levels of talent and potential.

The labor market institutions in professional team sports are quite unusual.[1] In European football, young players start as free agents but are able to make binding long-term contracts. The length of these contracts as well as the salaries are negotiable. The contracts effectively prevent players from working for other football clubs while the contract is in duration. This is where the transfer fees come from: before the end of the contract, a

[1]For a recent overview, see Rosen and Sanderson (2001).

player cannot switch clubs without the consent of the current club, for which the new club typically has to pay a transfer fee. About five out of six transfers results in a payment of a transfer fee (Carmichael et al 1999). The buying club takes on the responsibilities of the old contract, and the contracts are often extended at the time of a transfer with any changes in terms subject to approval by the player. A strong complementarity between talent and club size is a prominent feature of the industry: it makes sense for the best players to play for the clubs with the most fans. As a result, the net flow of talent is from the smaller clubs and leagues to the bigger, and so the net flow of transfer fees is to the opposite direction. For example, in the 1998-99 season in the English Premier league transfer fee payments were £269 million, of which £133 million was paid to foreign leagues and £13 million to English Division One clubs.

The market for football players has recently attracted much attention in Europe in the wake of soaring transfer fees.[2] The transfer fee system is deemed by some to be in breach of the European Union labor regulations guaranteeing workers the right to change employers. The mere idea of trading in people is a cause for lot of indignation, as is well captured by the statement of Viviane Reding, the EU's Sports Commissioner: "I find it scandalous that players are being used as objects of speculation, bought and sold like commodities."[3] The football industry has defended the transfer fees as necessary for compensating clubs that lose talented players they discovered and developed. A quote from Rick Parry, the CEO of Liverpool, sums up the main point of industry leaders: "My great concern is the impact of these proposals on developing young players. How can you protect the investment over the long term?"[4] The international players' union FIFPro has been muted in its comments, possibly due to internal disagreements, and in any case has not been advocating an abolition of transfer fees.

Regardless of any possible advantages of the transfer fee system, it has often been asked why professional sports should be exempted from the usual restrictions of labor law. Why can professional athletes commit to binding wage contracts that are not possible in other industries? Historically the reason why clubs have been able to enforce the transfer fee system probably comes from the special nature of sports industries: no firm can produce

[2]The current record fee of 67 million Euros was paid by Real Madrid for the French midfielder Zinedine Zidane in 2001 for the remaining four years of his contract with Juventus. His salary is not public, but is speculated to be about 4 million Euros per year (www.footballtransfers.info).

[3]Financial Times, August 31, 2000.

[4]Quoted in www.soccernet.com, September 6, 2000.

output without the cooperation of other firms, making it easy to punish cheaters. In the U.S. the exceptionality of sports is supported by several Supreme Court rulings (largely based on dubious competitive parity arguments), but in Europe the situation is more precarious. Some of the exceptionality was already removed with the so-called "Bosman ruling" of the European Court of Justice in 1995, by which clubs could no longer require transfer fees for players moving at the expiry of a contract. It is hard to see any crucial economic role for this strange feature of the old transfer system, and despite some vocal concerns from the industry at the time the only visible effect of this change seems to have been a lengthening of contracts.[5] However, the Bosman case brought the transfer system under the limelight and has started a process that may lead to more serious limitations on player contracts that could make them practically non-tradeable. The only respite is coming from the general understanding that clubs that train young players should be compensated when their players are "poached" by richer clubs. In the opinion of the Bosman decision judge, Advocate General Lenz, transfer fees should be limited by the actual cost of training, and the fee should only be payable for the first transfer from the club that trained the player. Different formulas have been suggested for pinning down "a fair price" for the training that would replace market-determined transfer fees and presumably decrease their general level.

It is easy for economists to see a valuable role for transfer fees. Much of contract theory studies the problems that stem from worker inability to credibly commit to keeping his word. In this sense professional sports look like a positive anomaly, and in any case a very interesting institutional arrangement. The defense of transfer fees has been based on the need for clubs to recoup training costs. However, these seem to be quite small compared to transfer fees.[6] Furthermore, players' market value can increase by orders of magnitude over the course of a year or two, but the largest increases take place while players are already playing professionally—not earlier while they train at youth academies or play for junior teams.

In the model I assume that there is no training. This is an approximation of the idea that the costs of training young players are only a small factor behind transfer fees. A change in a player's market value is due to his development as a player, which is a by-product of

[5]This quirk was only in effect in some countries, notably in Belgium and France. The more substantive part of the Bosman ruling prevented the discrimination of E.U. nationals in sports, which used to be common in the form of maximum quotas for foreign players.

[6]In the view of Morris et al (2000, p. 257) the transfer fees are even "wholly unrelated to the actual training and development costs incurred."

getting to play. I model this development as public learning about the talent of a player; it could also be interpreted as learning-by-doing by the player if the player's capacity to benefit from learning opportunities is defined as the ex ante unknown level of talent. Here not just talent but also the opportunities for learning are scarce. In football this scarcity stems from the scarcity of actual playing time with able co-players and opponents at the professional level, for which no amount of training can substitute. Here the scarcity is for simplicity modeled as an exogenously scarce number of clubs (or viable markets for a professional club).

Another crucial part of the model is the complementarity of player talent with job-specific characteristics, which are for simplicity subsumed under a single characteristic, referred to as "club size." Due to the heterogeneity of clubs by size, the efficient matching of a cohort of players with clubs changes as players develop and new information becomes available. The matching of clubs and players takes place in a competitive market, where the prices of talent (whether transfer fees or salaries) determine the division of rents between buyers and sellers of talent. Heterogeneity of clubs also means that scarcity rents are possible, not just for talented players, but also for clubs that have higher-value use for talent, i.e. "big" clubs. Besides affecting efficiency, the nature of the transfer system can affect the division of these rents.

The model predicts that the abolition of transfer fees causes an increase in the price of talent. In total, players gain less than clubs lose. Jobs are inefficiently reallocated towards reduced experimentation: some positions that should be used to try out new players with upside potential will instead be filled with older players with less potential, but with better expected near-term performance. There is a decline in turnover (in and out of the industry) and an increase in the average age and career length of players. Expected value of players' lifetime incomes goes up, but the model does not lend itself to analysis of player welfare because it does not account for risk aversion and thus ignores the insurance benefit from making long-term contracts. (In the model a credit constraint keeps young players from paying to play).

The outline of the chapter is as follows. First the basic model is presented and the determination and division of surplus is explained. The outcome of the market is analyzed both with and without the possibility of long-term commitment to transferable contracts by the individuals, with focus on the distribution of profits and wages, and turnover. The

effects of non-tradeable firm-specific commitment with the duration set by the regulator, and of fixed transfer fees with the level set by the regulator, are also considered. The chapter is concluded with a discussion of the results.

3.2 The Model

The starting point of the model is that the revenue generated by a player is increasing in his talent and in the size of the home market of the club. The size of the market, or "firm size" for short, is a fixed characteristic stemming from factors like the size of the club's home city and its historically determined fan base. The other central feature is that the level of talent can only be found out by actually playing in one of the scarce positions in the industry. Players have finite careers, so some new talent must be hired every period, at the very least to replace retiring players.

Assumptions

1. The revenue generated by a player is ab, where a is the talent of the player and b is firm size.

2. Every period a unit measure of potential players are born, their talent drawn from a distribution with the mean $\bar{\theta}$ and a continuous and strictly increasing CDF F_θ.

3. Player careers last up to two periods: the talent level of a novice is unknown, but becomes public knowledge after one period in the industry.

4. Players cannot work for less than reservation wage $\underline{w}$.

5. There is a unit measure of jobs in a continuum of risk neutral firms with an exogenous size profile $b[i]$, which is continuous and strictly increasing at all $i \in [0, 1]$

6. $\bar{\theta} b[0] \geq \underline{w}$.

7. There is no discounting. Firms are infinitely lived and maximize long-run average profits.

Assumption 1 defines the simplest possible complementary production technology. Of any two players, the more talented one would generate more revenue at any club, but this

difference is larger the bigger the club.[7] The main consequence of this complementarity is that the efficient matching is positively assortative: the best players should play for the biggest clubs.

Assumptions 2 and 3 describe the simplest possible information structure: at first nothing is known about the talent of novice players, and after one period of work their talent is known exactly. All novice players have therefore the expected talent of the population average $\overline{\theta}$. The ex ante homogeneity of potential players is not a crucial assumption, the point is that information about talent is much more inaccurate for inexperienced players. What is crucial is that important aspects of talent can only be reliably assessed in "real jobs" within the industry.

Assumption 4 is a simplification of the idea that young players are credit constrained (or risk averse) and cannot pay for the opportunity to play. Assumption 5 fixes the number of firms and jobs exogenously. The important feature is that it is impossible for all potential players to get to play and show their talent: there is a unit mass of jobs but a mass 2 (two cohorts of unit mass) of potential players alive each period. The assumption of a continuum of firms means that the market is competitive so that there is no room for bargaining or strategic behavior. Assumption 6 guarantees that even the smallest firm would at least break even if it had no other income besides revenue from the mean type. Assumption 7, no discounting, is made to simplify the notation.

There is no asymmetric information or effort cost and thus no moral hazard, adverse selection or other incentive problems in the model. Neither are there any kinds of frictions or firm-specific learning. Footballing talent is assumed to be general to the whole industry, which leaves out potentially very complicated real-world complementarities and substitutability between different types of players within a club.

The assumption of two-period careers means that long-term commitment is necessarily equated with career-long commitment, which is not observed in reality. In practice the contracts are staggered over the career: players are typically first traded while on initial contract, at which point the new contract is extended beyond the duration of the original contract. At this point, if the player is moving up, the wage is also typically revised upwards. That all players start in the industry as equally promising is also a stark abstraction, but

[7] For a stylized multiplicative example, suppose b is the number of club's potential supporters who come from mutually exclusive populations (i.e. no consumer is a potential supporter for more than one club). Talent a could then be defined as the average revenue per potential supporter.

again not crucial for the results. The model applies when there is significant uncertainty: when there are players who were expected to become stars but fade away and nobodies who just marginally, perhaps through an injury of somebody else, get a chance to play and become stars. This type of uncertainty is arguably very common in sports

In this study talent is defined merely as the capacity to generate revenue, whether through sales of tickets and merchandize, or television rights. Whether it is based on an inborn ability to acquire athletic skills or on the level of internal motivation is immaterial here. Also, revelation of talent is for practical purposes the same as learning-by-doing, if talent is defined as the initially unknown capability to benefit from an opportunity to learn on the job. The crucial factor is that the learning opportunities where stars are separated from mediocrities are only available inside the industry.

It is a crucial assumption here that the audience values player talent for its pure entertainment value on the quality of the game and not just for its effect on winning. If, to the contrary, audiences only cared about seeing their team win, then any investment into level of talent would be socially wasteful. A model where revenue depends on the player's talent exclusively via its rank in the distribution of talent would result in excessive search for talent, along the lines of Frank and Cook (1995).

The effect of competitive parity on industry revenue is a related issue. Total industry revenue could be lowered by having player talent too unevenly spread between the clubs, because too uneven competition lessens the interest in the sport. This can weaken the complementarity between player talent and club size, but cannot take it away. When there is value for spreading the chance of winning a competition, it is still more efficient to have the larger clubs win more often.

Finally, in a somewhat non-standard fashion, it is assumed that firms are inherently heterogeneous by how much revenue any given talent would generate there. What is the non-duplicable exogenous component in club size that is the source of economic rents for firms? There may be free entry into being a professional football club in Manchester, but not into being Manchester United.[8] Fan loyalty - a type of a very high switching cost - acts as a source of rents for a club with a large fan base. An increase in ticket prices may cause

[8]The way to enter the professional football industry in Europe is to buy an amateur or semi-professional low-tier club, and turn it into a fully professional clubs: a fifth division club can in theory be promoted to the top league within five seasons. This mechanism allows for the regional reallocation of major league clubs, which in the U.S. is achieved by moving the franchises.

die-hard fans to stay home more often and just read the newspaper report, but is less likely to turn them into supporters of a cheaper club. A club name with a glorious history and with a home in a large city is a unique asset that can earn rents. These rents should be expected to discounted in the stock price, and perhaps dissipated back when it was decided what club gets to occupy that niche, but that is inconsequential for the contemporaneous division of rents between clubs and players.

The Supply of Talent The distribution of talent in the population is fixed, but the distribution of talent in the workforce depends on how many novices were hired in the previous period. In steady state some proportion i^* of jobs are filled with novices and the remaining $1 - i^*$ jobs are filled with veterans, who were novices last period. Each feasible proportion of novices $\left(i^* \geq \frac{1}{2}\right)$ corresponds to a different threshold level of talent θ^* (which could also be used as the equilibrating variable). Players who turn out to be above the threshold θ^* "make the grade" and get to stay in the industry as veterans, while those below exit after one period. The threshold can not be below the population mean $\bar{\theta}$ in equilibrium because there are always more novices available, who are of the expected type $\bar{\theta}$ and willing to work at the lowest possible wage.

In what follows, the distributions are described in terms of their inverse distribution functions, or "profiles" for short. The profile of talent in the population of potential players, and therefore in any cohort of novices, is denoted by $\theta[i]$, and the profile of expected talent in the industry by $a[i|i^*]$. In $\theta[i]$, i refers to the quantile in the cohort, of whom an endogenous measure i^* gets hired as novices, whereas in $a[i|i^*]$ it refers to the quantile within the workforce, which was normalized to be of measure one (by there being a measure one of jobs).

The lowest i^* types in the industry by expected talent are the novices, who are in expectation of the average type $\overline{\theta}$. Since there is a measure $1 - i^*$ of veterans who are the best of the last period's cohort of novices they must be a proportion $\frac{1-i^*}{i^*}$ of that cohort. The threshold type must therefore be the $\left(1 - \frac{1-i^*}{i^*}\right)$th quantile of the population distribution, giving the relation of the thresholds as[9]

$$\theta^*(i^*) = \theta[2 - \frac{1}{i^*}]. \tag{3.2}$$

[9]Another way to derive this relation is to start by noting that the proportion of novices that get to stay on as veterans is $1 - F_\theta(\theta^*)$. In steady state the measure of veterans must be equal to the measure of novices

The talent profile of the veterans comes from the truncated distribution above θ^*. The combination of these is the profile of expected talent in the industry.

$$a[i|i^*] = \begin{cases} \overline{\theta} & i \in [0, i^*] \\ \theta[1 - \frac{1-i}{i^*}] & i \in (i^*, 1] \end{cases} \tag{3.3}$$

Note that the players in $[0, i^*]$ are actually a random draw from the whole distribution, but since their talent is unknown and firms are risk neutral they can be treated as the mean type $\overline{\theta}$.

The Division of Rents The first question in the model is how the rents are divided between the clubs and the players. This section derives the prices of talent for a given distribution of talent and firm size. The prices are determined in a competitive market where buyers (firms) and sellers of talent (firms or players) meet under symmetric information. All but possibly the least productive match will produce a rent over the sum of their outside opportunities, and the equilibrium prices pin down the division of this rent into "factor incomes" for the owners of talent and firms. The price of talent can consist of a wage, a transfer fee or both, depending on who owns the (contractual rights to) talent.

Due to the complementarity in production, the efficient matching of individuals and firms is simple: the largest firm hires the highest available talent, the second largest hires the next highest talent and so on. Equilibrium prices must be consistent with this efficient matching. The assumptions of a continuum of players and continuous distributions of talent and firm size guarantee that they are also unique, i.e. there will be no match-specific rents left for bargaining. With these assumptions the setup is essentially the same as that in the assignment model of Sattinger (1979).[10]

The profiles of talent and firm size are denoted by $a[i]$ and $b[i]$ respectively, where i is the quantile $i \in [0, 1]$. (Here the proportion of novices i^* is treated as a constant and suppressed from the notation.) The equilibrium price for talent $a[i]$ is denoted by $p[i]$, it is paid by firm i to the owner of talent $a[i]$ and does not include the reservation wage $\underline{w}$,

who turn out to be above the threshold:

$$\begin{aligned} 1 - i^* &= i^*(1 - F_\theta(\theta^*)) \\ \Longrightarrow i^*(\theta^*) &= 1/(2 - F_\theta(\theta^*)), \end{aligned} \tag{3.1}$$

which is the inverse of (3.2).

[10] For a survey of assignment models see Sattinger (1993).

which must be paid to all players regardless of who gets the rents from their talent. The condition for all firms to want to stick to their own match is

$$a[i]b[i] - p[i] \geq a[j]b[i] - p[j] \qquad \forall i, j \in [0,1]. \tag{3.4}$$

Furthermore, the firms have to at least break even and the sellers must get a nonnegative price.

$$a[i]b[i] - p[i] - \underline{w} \;\geq\; 0 \quad \forall i \in [0,1] \tag{3.5}$$

$$p[i] \;\geq\; 0 \quad \forall i \in [0,1] \tag{3.6}$$

Inequalities (3.4) and (3.5) are mathematically analogous to incentive compatibility and participation constraints in a nonlinear pricing problem with quasi-linear utility functions and "types" $b[i]$. The prices that simultaneously fulfill the above criteria for all buyers and sellers can be found using the constraint reduction method familiar from nonlinear pricing problems. The binding constraints are those that prevent firms from wanting to hire the next lowest talent. These binding constraints define the slope of the price profile.[11]

$$p'[i] = a'[i]b[i] \tag{3.8}$$

Finally, by integrating the slope of the price profile, we get the equilibrium prices for talent.

$$p[i] = \int_0^i a'[j]b[j]\mathrm{d}j, \quad i \in [0,1]. \tag{3.9}$$

The intercept $p[0] = 0$ results from the assumption that there are more potential players than there are jobs, so the lowest type hired cannot get any rent.[12] Thus firms capture all of the rent at the bottom, $\pi[0] = a[0]b[0] - \underline{w}$. The buyer's share of the rents is easily recovered from equation (3.9) as the leftover $\pi[i] = a[i]b[i] - p[i] - \underline{w} = \pi[0] + \int_0^i a[j]b'[j]\mathrm{d}j$. Firms

[11] Regrouping the IC constraint (3.4) for $j = i - \varepsilon$ and dividing it by ε gives

$$\frac{p[i] - p[i-\varepsilon]}{\varepsilon} \leq \frac{(a[i]b[i] - a[i-\varepsilon])\, b[i]}{\varepsilon}. \tag{3.7}$$

This holds as an equality as $\varepsilon \to 0$ and, via the definition of the derivative, yields the slope of the price profile.

[12] We could think of the profile of available talent starting at some negative i.

are not residual claimants however, the equilibrium could equally well have been defined starting from sellers' constraints.

The level and dispersion of rents to talent depend on the dispersion of talent levels and firm size. If firms were homogeneous, so that $b[i] \equiv \bar{b} > 0$, then rents to talent would simply be Ricardian rents: $p[i] = (a[i] - a[0])\,\bar{b}$. In any case, for any level of talent, the rents are increasing in the advantage over the marginal talent and in the level of the complementary factor. When firms are heterogeneous then the division of rents at any quantile i depends on the whole distributions of talent and firm size below. The price of a talent of level $a[i]$ in (3.9) is a weighted sum of the "increments" in talent between $a[i]$ and $a[0]$, where the weights are the sizes of the firms matched at each increment. The share of talent of the rent created at firm i, i.e. of $a[i]b[i] - \underline{w}$, is therefore higher when the talent levels of competitors below are closer to $a[0]$, because then the high values of a' are weighted by higher $b[i]$s. It is also higher when the competing buyers are as close as possible to $b[i]$ in size, so that the weights are everywhere as high as possible. As is intuitive, it is best to have one's own competitors to be of low productivity, and to have one's equilibrium match have to compete with many close substitutes.

The prices of talent constitute an equilibrium in a market where both buyers and sellers can make offers. In equilibrium no firm can lower its offer to its efficient match without losing that talent to another firm and no firm would like to hire another firm's match at their equilibrium price. Neither buyers nor sellers can gain by making any other offers to anyone else besides the equilibrium offer to their efficient match.

In this continuous model there is nothing to be bargained over: every buyer's and every seller's opportunity cost inside the market is exactly binding. In a discrete model there would be a relationship-specific rent bounded by the threat values of matching with the next lowest counterpart. For thin-tailed distributions of talent and/or firm size this bargaining residual could be substantial at the highest level, but less so for more ordinary individuals and firms for whom the market is more "liquid" in the sense of there being very similar alternative matches and competitors in the market.

3.3 Equilibrium with Transfer Fees

When individuals can commit to transferable long-term contracts, then novices agree to do so at their reservation wage. They cannot do any better since individuals of unknown talent are not scarce. When a club discovers a high talent it can sell his remaining contract and get the full talent rent as the transfer fee. Since novices are the least talented players by expectation to work in the industry it is the small clubs that employ them, while the big clubs buy their talent from the small clubs. In equilibrium, the threshold firm i^* must be indifferent between hiring a novice and getting the expected transfer fee, and between hiring the threshold talent $\theta^*(i^*)$ for a zero transfer fee. The threshold talent is transferred at a zero fee, because it is also the highest talent to be discarded from the industry. The transfer fees are the prices of talent as determined as in section 3.2, but with the profile of talent $a[i|i^*]$ now dependent on the endogenous proportion of novices.

Since the smallest i^* firms hire novices they don't pay any transfer fees. In terms of the price equation (3.9), the profile of expected talent is flat for the novice-hiring firms: $a'[j|i^*] = 0$ for $j \in [0, i^*)$. Given i^*, the transfer fee paid by firm $i > i^*$ for its match, a talent of level $a[i|i^*]$, is

$$p[i|i^*] = \int_{i^*}^{i} a'[j|i^*]b[j]\mathrm{d}j. \tag{3.10}$$

(And zero for $i < i^*$). Total transfer fees in the industry are

$$P[i^*] = \int_{i^*}^{1} p[i|i^*]\mathrm{d}i = \int_{i^*}^{1}\int_{i^*}^{i} a'[j|i^*]b[j]\mathrm{d}j\mathrm{d}i = \int_{i^*}^{1} (1-i)\, a'[i|i^*]b[i]\mathrm{d}i, \tag{3.11}$$

where the last step involves a partial integration. Since all novice-hiring firms draw their talent from the same distribution they all get the same expected share of these total rents in expectation, namely $\frac{1}{i^*}P[i^*]$. They also get the revenue generated by a novice, who is of the expected type $\overline{\theta}$. The long-run average profits of a novice-hiring firm i are

$$\pi^0[i|i^*] = \overline{\theta}b[i] + \frac{1}{i^*}P[i^*] - \underline{w}. \tag{3.12}$$

Firms that hire veterans get the revenue from their match while paying the corresponding

transfer fee.

$$\pi^1[i|i^*] = a[i|i^*]b[i] - p[i|i^*] - \underline{w}, \quad i \geq i^*. \tag{3.13}$$

Either way, the firms must always pay the current employee $\underline{w}$ to get him to actually work, regardless of his experience or talent level.

The threshold firm must be indifferent between employing novices or veterans so the equilibrium i^* is defined by $\pi^0[i^*|i^*] = \pi^1[i^*|i^*]$. Rearranging this equilibrium condition we get

$$\left(\theta^*(i^*) - \overline{\theta}\right) b[i^*] = \frac{1}{i^*} P[i^*], \tag{3.14}$$

where $a[i^*|i^*] = \theta^*(i^*)$ and $p[i^*|i^*] = 0$ were used. On the left is the opportunity cost of hiring a novice: it is the lost revenue from hiring a novice as opposed to the threshold talent (who would be available at zero transfer fee). On the right is the benefit, the expected transfer fee earned by hiring a novice. Note that the equilibrium rehiring threshold is strictly above the population average: small clubs sacrifice some current revenue in exchange for expected transfer fees in the future.[13] The equilibrium is unique because the left side is strictly increasing in i^* and changes sign at an intermediate value, whereas the right side is positive, strictly decreasing, and reaches zero at $i^* = 1$. Higher i^* means that more firms are trying to sell talent to fewer buyers, so it is intuitive that the expected price goes down.

To verify that the equilibrium is efficient, first note that the opportunity cost of a unit measure of players and firms is fixed, so total surplus only depends on i^* via its effect on the distribution of talent in the industry. Total surplus in the industry is a function of the proportion of novices i^*

$$Y\left(i^*\right) = \overline{\theta} \int_0^{i^*} b[i]\mathrm{d}i + \int_{i^*}^1 a[i|i^*]b[i]\mathrm{d}i - \underline{w}. \tag{3.15}$$

The first order condition is

$$\begin{aligned} Y'\left(i^*\right) &= \overline{\theta} b[i^*] - a[i^*|i^*]b[i^*] + \int_{i^*}^1 \frac{\partial a[i|i^*]}{\partial i^*} b[i]\mathrm{d}i = 0 & (3.16) \\ &= \left(\overline{\theta} - \theta^*\left(i^*\right)\right) b[i^*] + \int_{i^*}^1 \frac{\partial a[i|i^*]}{\partial i^*} b[i]\mathrm{d}i = 0. & (3.17) \end{aligned}$$

[13] With discounting, the right side would be multiplied by the discount factor.

Expanding the integrand by using (3.3) and taking the derivative yields

$$\frac{\partial a[i|i^*]}{\partial i^*} = \frac{\partial}{\partial i^*}\theta[\frac{i+i^*-1}{i^*}] = \frac{1-i}{i^{*2}}\theta'[\frac{i+i^*-1}{i^*}] = \frac{1-i}{i^*}a'[i|i^*]. \tag{3.18}$$

Plugging this back into (3.17) we see that the integral is equal to $\frac{P[i^*]}{i^*}$, so the first-order condition of the total surplus maximization problem is the same as the market equilibrium condition under transfer fees. Of course, this is just what must result from perfect competition and lack of externalities.

Transfer fees would not be needed for efficiency if novices could pay to play. Risk neutral novices with unconstrained credit would be willing to pay for the opportunity to play up to the expected value of second period talent rents. In equilibrium the novice wage would then be

$$\underline{w}^* = \underline{w} - \frac{P[i^*]}{i^*} = \underline{w} - \left(\theta^*(i^*) - \overline{\theta}\right) b[i^*]. \tag{3.19}$$

The novices would in effect buy out older players of below $\theta^*(i^*)$ talent, which could be very costly if the difference between mean talent and threshold talent is worth a lot of money at the threshold club. The veteran wage would be $\underline{w} + p[i|i^*]$, and this payoff could be very risky. For skewed distributions of talent and firm size most of the expected rents come from a small chance of being a superstar, so even moderately risk averse novices with access to unconstrained credit might be willing to pay only a small fraction of the expected rents.

3.4 Equilibrium without Transfer Fees

There can be no transfer fees if it is not possible for individuals to commit to transferable long-term contracts. Without long-term commitment talented players can be "raided" by another club that offers a higher wage. Due to the complementarity in production, it is efficient that the best players should move up and pay with the biggest clubs, but now the clubs that "discovered" them will get no compensation. As before, players that turn out to be below average will not be rehired because novices are abundant and more talented by expectation. The problem is that now all players that turn out to be above the population average, no matter by how little, will be hired by some club.

At the level of the whole industry the basic trade-off is how to allocate the scarce playing

opportunities between novices and veterans. Hiring more novices allows for a larger supply of talented veterans but leaves fewer positions for them to use that talent. Without transfer fees this trade-off does not factor into clubs' hiring decision. When clubs base their hiring decisions only based on the expected revenue at the hiring club they ignore the upside potential of younger players and the higher value of talent at bigger clubs. The upside potential is now inevitably captured by players themselves.

In the absence of transfer fees the novice-hiring clubs only get revenue from output. The right side of the equilibrium condition (3.14) is therefore replaced by zero, so that $\theta^*\left(i^0\right) = \bar{\theta}$ is the only solution. Since a larger fraction of novices get to stay in the industry as veterans, the proportion of jobs held by novices is now smaller: $i^0 < i^*$.[14] That fewer clubs hire novices should be intuitive since it is made less profitable by the elimination of transfer fees. A section of medium-productivity jobs, $[i^0, i^*]$, will be filled with veterans $\theta \in [\bar{\theta}, \theta^*(i^*)]$ instead of novices. These "mediocre" types are more talented than novices by expectation, but would not be employed under the transfer fee system. Notice that the solution i^0 is independent of the firm size profile $b[i]$, reflecting the fact that the gains from trade between clubs are not taken into account.

3.5 The Abolition of Transfer Fees

Before listing the predicted effects of the abolition of transfer fees, let's note one change that does not happen even though it might seem intuitive at first. Even though there is reduced discovery of talent, it is not the case that the distribution of talent in the industry would simply become worse (e.g. in the sense of stochastic dominance). With a reduced proportion of jobs set aside for novices, fewer high talents are indeed discovered. However, there are also fewer low types because novices are (on average) the lowest types to play. Thus there must be are more middle types, i.e. "mediocre" players who are better than (population) average $\bar{\theta}$ but not good enough to have been retained before. Figure 3-1 shows the profile of talent in the industry with and without transfer fees, depicted by dashed and solid lines respectively. The corresponding profile of clubs size is fixed and not shown in the figure.

So what does the model tell us about the effects of abolishing transfer fees? There are

[14] The proportion of novices is now $i^0 \equiv i^*(\overline{\theta}) = 1/(2 - F_\theta(\overline{\theta}))$, see the footnote on page 92.

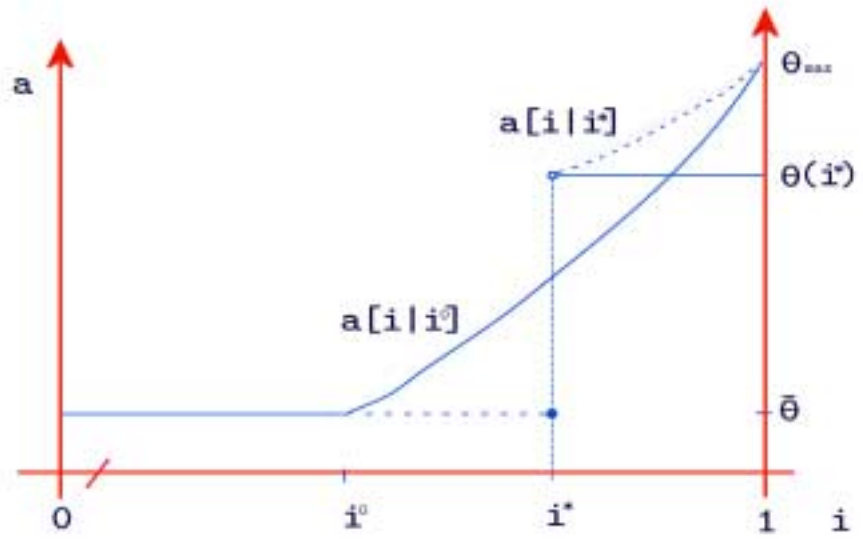

Figure 3-1: The profile of (expected) talent in the industry, with and without transfer fees, depicted by dashed and solid line respectively.

three main implications that are robust to the stark simplifying assumptions of the model.

1. The price of talent is increased for each level of talent $\theta > \bar{\theta}$.

Consider the price of talent before and after the abolition of transfer fees, denoted by $p^*(\theta)$ and $p(\theta)$. First, notice that $p(\theta) > p^*(\theta) = 0$ for $\theta \in (\bar{\theta}, \theta^*(i^*)]$, since these types were discarded and available at zero fee under the transfer fee system, but have a positive price afterwards (since they are above the new threshold type). Second, each level of talent above the old threshold, $\theta > \theta^*(i^*)$, is now matched with a bigger club than they would have been matched with under the transfer fee system; this is just the flip-side of the fact that each club above i^* is now matching with a lower talent than before. Third, the derivative of the price function $p(\theta)$ at any $\theta' > \theta^*(i^*)$ is equal to the size of the club matched with θ'. This can be seen by changing the variable of integration from i to θ in (3.9). Therefore $p(\theta') - p(\theta'') > p^*(\theta') - p^*(\theta'')$ for any $\theta' > \theta'' \geq \theta^*(i^*)$. The increase in the price of talent caused by the abolition of transfer fees is in fact higher for higher talent, so the biggest increase of all goes to the very highest type.

2. All clubs get lower profits.

This is obvious for clubs that hire novices before and after, for them the only change is the loss of transfer fees. Clubs that hire veterans before and after, each of them (except

$i = 1$) now has to match with a lower talent than before, whom they could have hired before at a lower price than now but didn't. Finally, the switching clubs in the middle now get more revenue from output than before, since they employ above-average types instead of novices, but they also lose the transfer fees. The loss must be larger than the gain, since the types that they hire are below the old threshold θ^* (i^*) type who would have been available at zero fee before.

3. Player careers become longer.

The exit rate of novices is lowered simply because the level of talent required to stay in the industry as a veteran is lowered. Thus, even though fewer players ever get the chance to enter the industry, a higher proportion of those who do get to stay for the long term. Of course, an increase in the average age of players is also immediate from the increased proportion of veterans.

To sum up, the elimination of transfer fees would result in fewer young players and highly talented old players and more old mediocre players getting scarce playing time in professional leagues. All clubs are worse off as a result, while salaries of older players increase. Since equilibrium with transfer fees maximizes total surplus, this means that the players gain less than the owners lose. The most clear-cut observable predicted change would be the upward shift in the age distribution of players.

3.6 Other Regulations

The main motivation of the opponents of the transfer fee system seems to be a visceral revulsion to what is viewed as a trade in human beings. While there is widespread understanding that clubs need to be compensated for nurturing young talent, this is not seen as justifying the multimillion Euro fees observed at the top of the market. However, this understanding has lead to suggestions to temper the reform in ways that would still reward investment into young talent while removing the outright trade in players, or at least cutting down the current fee levels.

In this section I will analyze two such modifications. The first is firm-specific commitment, i.e. non-tradeable contracts. The idea is that clubs that discover talent can keep them for some amount of time at the original contractual wage, but they cannot sell them to other clubs. At the end of the contract the players become free agents and can move

to whichever club offers the highest wage. This system gives some rewards to clubs that develop talent while eliminating the hated transfer fees. I will show that it could actually be worse for efficiency than an outright abolition of the transfer fee system. The other proposal is a fixed transfer fee. There the idea is that the regulator sets what is deemed a single reasonable level of compensation for all transfers. In practice this might include some formula that takes into account a player's age and the number of games played both for the club of origin and the acquiring club. I will show that if combined with a mandatory turnover rule then a correctly chosen fixed fee could in principle replicate the efficient allocation of the unregulated transfer fee system, with only redistributive effects to the favor of the (best) players.

It is assumed throughout that any restriction on transfer fees cannot be bypassed, for example by masking the fees as termination penalties. The transfer fee system could in fact be exactly replicated by renaming the fees as termination penalties. These would be set at least as high as the market price of highest talent under the transfer fee system. After the revelation of talent, the novice-hiring club would offer a discount on the maximum penalty, and in equilibrium the penalty actually paid would equal the market price of talent. Termination penalties might be rhetorically acceptable to some opponents of the current system, as individuals are not transparently being "bought and sold like commodities" but just fined on their broken promises. However, in case they were allowed after the abolition of transfer fees it would probably become quite soon obvious to everyone that the penalties are just transfer fees under a different name.

Firm-specific Commitment

Under firm-specific commitment players can commit to work at the same club but the contract cannot be traded to other clubs. If this commitment lasted the whole career than the system would be a virtual hiring autarky, however the regulator may want to limit the length of commitment time to less than the whole career. Limited firm-specific commitment eliminates transfer fees and allows for some reward to clubs that discover talent. The reward for the novice-hiring clubs is that they get to keep any talent they find for a limited amount of time at the original contract wage. At the end of the limited commitment period players become free agents, and move to the club that offers the highest wage. Free agents are therefore always efficiently matched with the biggest clubs, who get a steady supply of

proven high talent without needing to give high-productivity playing time to novices.

To facilitate comparison with the other cases, keep assuming a (maximum) career length of two periods, where it takes one period to for the level of talent to be revealed. Then suppose that periods are divisible and use τ to denote the fraction of players' second period after which they become free agents. The duration of commitment $(1+\tau)$ is chosen by the regulator and clubs take it as a given parameter in their profit-maximization problem.

The long-run average profits of a novice-hiring club i are $\pi[i] = A(\theta^*|\tau)b[i] - \underline{w}$, where $A(\theta^*|\tau)$ is the long-run average level of talent employed. To maximize profits, novice-hiring clubs choose the retaining threshold to maximize the average level of talent at the club. To see what this average is, first note that by using a retaining threshold θ^* a club will be hiring novices a fraction

$$\phi(\theta^*|\tau) = \frac{1}{1+\tau(1-F_\theta(\theta^*))} \tag{3.20}$$

of time, and retaining above-θ^* types the rest.[15] The long-run average level of talent at novice-hiring clubs is then

$$A(\theta^*|\tau) = \phi(\theta^*|\tau)\bar{\theta} + (1-\phi(\theta^*|\tau))\,\mathrm{E}[\theta|\theta > \theta^*]. \tag{3.22}$$

The first order condition is

$$\theta^* - \bar{\theta} = \tau\,(1-F_\theta(\theta^*))\,(\mathrm{E}[\theta|\theta > \theta^*] - \theta^*)\,. \tag{3.23}$$

The maximizer is necessarily the fixed point of this function $A(\cdot|\tau)$, denoted by $A^*(\tau)$. In other words, at the maximum, a club's retaining threshold is equal to the average talent of its employees over time. The intuition for this result is that when the maximal average level of talent employed is A^*, then discarding a talent above A^* would have to decrease that average as would retaining a talent below A^*, meaning that A^* must be the maximizing threshold.

[15] To solve for ϕ, first suppose that retainment also lasted for a full period. Then the equilibrium probability p that a club employing a novice at any given period must satisfy

$$pF_\theta(\theta^*) + (1-p)1 = p \tag{3.21}$$

because not employing a novice this period means that next period the club will employ a novice for sure. This results in $p = 1/(2-F_\theta(\theta^*))$. Since retaining periods only last a fraction τ of novice periods, the proportion of time spent retaining is $\phi = \frac{p}{p+\tau(1-p)}$, which results in (3.20).

The first-order condition (3.23) describes well the nature of the "investment" problem: by hiring a novice instead of a known type θ^* a club reduces its immediate talent by $\theta^* - \bar{\theta}$, but makes the expected gain described by the right side. A longer commitment time will make it more worthwhile to hire novices, increasing the threshold cum average: $A^{*\prime}(\tau) > 0$. For $\tau = 0$ this is just the case without any commitment, so $A^*(0) = \bar{\theta}$. With $\tau = 1$ this becomes the case of full firm-specific commitment, where all firms would operate in hiring autarky. The resulting average $A^*(1)$ is the maximum feasible average level of talent in the whole industry.

Compared to a simple abolition of transfer fees, which is equivalent to setting $\tau = 0$, this system has the benefit that it gives some incentives to fire mediocre players so there will be more experimentation with new players. The trade-off from increasing the commitment τ is illustrated in 3-2, where the dashed line shows a profile of talent in the industry for an intermediate value of τ. The clubs that hired novices under $\tau = 0$ gain because the average talent of their player goes up. Longer retainment makes it more attractive to hire novices, so more clubs start doing it. The downside is that a smaller fraction of the high-talent players are available for the biggest clubs to hire at any time, so they will have to match with a lower type than before: after an increase in τ there is a better supply of talent in the industry, but the matching less efficient. The biggest clubs compete for the free agents, who will receive the market price of talent as wages. Whether total surplus is increased or decreased would depend on the distribution of firm size. Roughly, if the distribution are very concentrated (i.e. if the profile $b[i]$ is quite flat) then the effect of the increase in average talent is more useful, but if the gains from efficient over random matching are very large (e.g. when the largest firms are very large compared to the others) then the loss from inefficient matching is more important. It could therefore be best to simply end all long-term commitment.

On the other hand, career-long commitment ($\tau = 1$) would only be optimal if all firms were identical. To see this, consider starting from $\tau = 1$ and decreasing τ a little bit. This would discontinuously increase the talent available to the largest firm $b[1]$ from $A^*(1)$ to $\theta[1]$, while lowering the average talent at all other firms only slightly. If $b[1]$ is strictly larger than the average firm in $i = [0, 1)$ then this will increase the total surplus $Y = \int_0^1 a[i]b[i]\mathrm{d}i - \underline{w}$.

To summarize, when transfer fees are not allowed and the regulator can set the length of non-tradeable contracts, then the optimal commitment time is decreasing in the hetero-

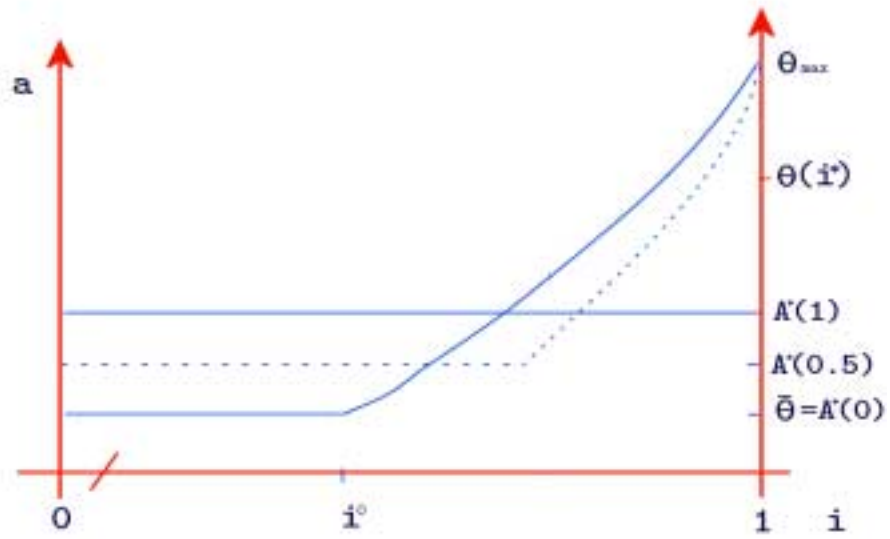

Figure 3-2: Profile of (expected) talent in the industry, for $\tau \in \{0, 0.5, 1\}$.

geneity of the firms by size. A zero commitment time may be optimal, but full commitment is only optimal if all firms are identical.

Fixed Transfer Fees

Following the opinion of Advocate General Lenz, some regulation proposals have called for a fixed industry-wide transfer fee, which would presumably reflect some generally acceptable level of compensation. The motivation for a fixed transfer fee is apparently to end "the speculation in players" while still providing clubs with some incentive to invest into young talent.

A fixed transfer fee can be implemented when the regulator can observe payments and movements of players between clubs. It is not necessary to observe talent levels or payments between clubs and players (which would be harder because they could be nonmonetary). Fixed fees do not disturb the efficient matching of those individuals who are actually traded: clubs would be indifferent between who buys their players, but it seems plausible that players would be sold to whoever offers the highest wage for them, i.e. their efficient match.

One might think that the problem with a fixed fee is that some efficient trades are prevented when the gains from trade are less than the fixed fee. In these cases it would be in the interest of selling clubs to give discounts when their player is not good enough to

be bought at the fixed fee. However, if the fee could be rigged downwards in this matter, then an efficient solution would not be possible. This may be counter-intuitive, but the reason is that the problem with the complete abolition of transfer fees is not that efficient trading of available talent would be prevented but rather that too many mediocre talents are retained, causing the distribution of available talent to be suboptimal. A fixed fee can work as a "retainment tax" which is used to subsidize clubs that hire novices. In this sense discounts would amount to a distortinary tax dodge, allowing mediocre talents to stay in the industry.

For a fixed transfer fee to achieve the efficient solution it is required that clubs are never allowed to retain their own novices. After all, selling clubs can also use talent in-house and might not want to sell the player if the price is too low. Such retaining could be prevented with a mandatory turnover rule stating that players who stay in the industry after their novice period must either switch clubs or exit the industry.

Suppose that players must either be sold or discarded after one period, and that the fixed fee is set at some level $\bar{p}$ that results in a fraction i^* of clubs hiring novices. The profits of a novice-hiring club i are then

$$\bar{\pi}^0[i|i^*] = \bar{\theta}b[i] + \frac{\bar{p}\,(1-i^*)}{i^*} - \underline{w}. \tag{3.24}$$

Here the fixed transfer fee is multiplied by the probability that a novice will turn out to be good enough to be sold; with two-period careers this probability must be the ratio of veterans to novices. For a club that hires veterans the profits are

$$\bar{\pi}^1[i|i^*] = a[i|i^*]b[i^*] - \bar{p} - p[i|i^*] - \underline{w}, \tag{3.25}$$

where the price of talent $p[i|i^*]$ now goes to the player as a wage.

The equality $\bar{\pi}^0[i^*|i^*] = \bar{\pi}^1[i^*|i^*]$ is the indifference condition of the marginal firm, which must hold for i^* to be the equilibrium. Using $a[i^*|i^*] = \theta^*(i^*)$ and $p[i^*|i^*] = 0$ it can be rearranged to

$$\left(\theta^*(i^*) - \bar{\theta}\right) b[i^*] = \frac{\bar{p}}{i^*}. \tag{3.26}$$

This is almost the same as the equilibrium condition of the unregulated market (3.14),

which was shown to be a condition for the maximization of total industry output. Setting $\bar{p}^* = P[i^*]$ will replicate the efficient equilibrium condition. In other words, the optimal fixed transfer fee is equal to the average transfer fee per all jobs in the industry under the unregulated transfer fee system. It is therefore less than the average transfer fee per transaction in the unregulated case, which is $\frac{P[i^*]}{1-i^*}$. As a result of fixing transfer fees at $\bar{p}^*$, all clubs are worse off by $\bar{p}^*$ while players who are good enough to be retained gain the price of talent that used to go to discovering clubs.

This scheme only works because the novice-hiring firms are assumed to not be able to hoard unsalable talent. To see why this requires mandatory turnover after revelation of talent, notice that the marginal firm would strictly prefer to hold on to the threshold type if he were available without a transfer fee. Also crucially, it was assumed that exit of firms is not a problem. This leaves some rents at the bottom and gives leeway in rearranging the rents without affecting the distribution of talent in the industry. However, compared to a complete abolition of transfer fees, the problem of exit by smallest clubs would be smaller here since they at least get some transfer fee income.

To summarize, it is in principal possible to set a fixed transfer fee that would replicate the efficient solution (from market-determined transfer fees), except for a shift in the division of rents to players' favor. This requires that clubs that hire novices are neither allowed to retain their own finds nor to give discounts below the fixed fee. The fixed fee must be thought of a retainment tax that shifts the allocation of jobs in favor of novices, and displaces the mediocre veterans who are inefficiently retained in the absence of transfer fees. Of course, the fixed fee would have to be set at exactly the right level for full efficiency to be attained, which would be unlikely in practice.

3.7 Conclusion

The ongoing discussion on the role of transfer fees in professional sports and on the reform of the transfer system in the European football industry has so far ignored the importance of experimentation and on-the-job learning.[16] There has been much concern over the effect of elimination of transfer fees on clubs' incentives to train young players, but the training costs do not seem to justify the current levels of transfer fees. Actual playing time is scarce

[16]See e.g. Antonioni and Cubin (2000), Feess and Muhlheusser (2002) and Sanderson and Siegfried (1997).

and the most significant investment that a club can make to develop its players may be to let them play. Given that the club will always let someone play, this is a pure opportunity cost that does not show up in any accounting. As much as transfer fees give clubs the right incentives to develop their players, their elimination or replacement with training-cost reimbursements could have dire consequences for the quality of players and the game in general. Reduced incentives for experimentation with new talent would show up as lower turnover and an upward shift in the age distribution of players. For those who get their foot inside the door at professional level it would be easier to stay in, and the increase in wages for each type of player would be higher than are the current transfer fees. Not just the clubs currently selling talent, but also the net buyers—not to mention the consumers—would be worse off as a result.

Bibliography

Acemoglu, Daron and Jörn-Steffen Pischke (1998), "Why Do Firms Train? Theory and Evidence" *Quarterly Journal of Economics*, 113, pp. 79–119.

Akerlof, George A and Lawrence F Katz (1989), "Workers' Trust Funds and the Logic of Wage Profiles" *Quarterly Journal of Economics*, 104, pp. 525–536.

Antonioni, Peter and John Cubbin (2000): "The Bosman Ruling and the Emergence of a Single Market in Soccer Talent." *European Journal of Law and Economics,* 9, pp. 157–173.

Baker, George P and Brian J Hall (forthcoming): "CEO Incentives and Firm Size." *Journal of Labor Economics.*

Baker, George P and Michael C Jensen and Kevin J Murphy (1988): "Compensation and Incentives: practice vs Theory." *Journal of Finance*, pp. 593–616.

Becker, Gary S (1964), *Human Capital.* University of Chicago Press.

Bertrand, Marianne and Sendhil Mullainathan (2001): "Are CEOs Rewarded for Luck? The Ones Without Principles Are." *Quarterly Journal of Economics*, 116, pp. 901–932.

Brown, Charles and James Medoff (1989): "The Employer Size-Wage Effect." *Journal of Political Economy*, 97, pp. 1027–1059.

Carmichael, Fiona, David Forrest and Robert Simmons (1999): "The Labour Market in Association Football: Who Gets Transferred and For How Much?" *Bulletin of Economic Research*, pp. 125-150.

Carmichael, F and D Thomas (1993): "Bargaining in the Transfer Market: Theory and Evidence." *Applied Economics*, 25, pp. 1467-1476.

Caves, Richard E (2000), *Creative Industries: Contracts Between Arts and Commerce.* Harvard University Press.

Conant, Michael (1960), *Antitrust in the Motion Picture Industry: Economic and Legal Analysis.* University of California Press.

Demougin, Dominique and Aloysius Siow (1994), "Careers in Ongoing Hierarchies." *American Economic Review*, 84, pp. 1261–1277.

— (1996), "Managerial Husbandry and the Dynamics of Ongoing Hierarchies." *European Economic Review*, 40, pp. 1483–99.

Feess, Eberhard and Gerd Mühlheusser (2001): "Transfer Fee Regulations in European Football." *Working paper.*

— (2002): "Economic Consequences of Transfer Fee Regulations in European Football." *European Journal of Law and Economics,* 13, pp. 221–237.

Fort, Rodney and James Quirk (1995), "Cross-subsidization, Incentives, and Outcomes in Professional Team Sports Leagues." *Journal of Economic Literature*, 33, pp. 1265–1299.

Frank, Robert H and Philip J Cook (1995), *The Winner-Take-All Society.* The Free Press.

Ghatak, Maitreesh; Massimo Morelli and Tomas Sjöström (2001), "Occupational Choice and Dynamic Incentives." *Review of Economic Studies*, 68, pp. 781–810.

Gibbons, Robert and Michael M Waldman (1984), "Careers in Organizations: Theory and Evidence." in *Handbook of Labor Economics*, Vol 3, Orley Ashenfelter and David Card, eds. North-Holland, pp. 2373–2437.

— (1999), "A Theory of Wage and Promotion Dynamics Inside Firms." *Quarterly Journal of Economics*, 114, pp. 1321–1358.

Gittins, John C (1989), *Multi-Armed Bandit Allocation Indices.* John Wiley & Sons.

Greenwald, Bruce C (1986), "Adverse Selection in the Labour Market." *Review of Economic Studies*, 53, pp. 325-347.

Guasch, J Luis and Joel Sobel (1983), "Breeding and Raiding. A Theory of Strategic Production of Skills." *European Economic Review*, 22, pp. 97–115.

Harris, Milton and Bengt Holmström (1982), "A Theory of Wage Dynamics." *Review of Economic Studies*, 49, pp. 315-333.

Harris, Milton and Yoram Weiss (1984), "Job Matching with Finite Horizon and Risk Aversion." *Journal of Political Economy*, 92, pp. 758–779.

Hayes, Rachel M and Scott Schaefer (1999): "How Much Are Differences in Managerial Ability Worth?" *Journal of Accounting and Economics*, pp. 125–148.

Himmelberg, Charles P and R Glenn Hubbard (2000): "Incentive Pay and the Market for CEOs: An Analysis of Pay-for-Performance Sensitivity." *Working Paper.*

Holmström, Bengt (1992): "Comment." in *Contract Economics*, edited by Werin, Lars and Hans Wijkander, Blackwell, pp. 212–217.

Johnson, William R (1978), "A Theory of Job Shopping." *Quarterly Journal of Economics*, 93, pp. 261–277.

Jovanovic, Boyan (1979), "Job Matching and the Theory of Turnover." *Journal of Political Economy*, 87, pp. 972–990.

Kahn, Charles and Gur Huberman (1988), "Two-Sided Uncertainty and Up-or-Out Contracts." *Journal of Labor Economics*, 6, pp. 423–444.

Koopmans, Tjalling C and Martin Beckmann (1957): "Assignment Problem and the Location of Economic Activities." *Econometrica*, 25, pp. 53–76.

Kostiuk, Peter F (1990): "Firm Size and Executive Compensation." *Journal of Human Resources*, 25, pp. 91–105.

Kremer, Michael (1993): "The O-Ring Theory of Economic Development." *Quarterly Journal of Economics*, 108, pp. 551–575.

Kremer, Michael and Eric Maskin (forthcoming): "Wage Inequality and Segregation by Skill." *Quarterly Journal of Economics*.

Lazear, Edwin P and Sherwin Rosen (1981), "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy*, 89, pp. 841–864.

Lucas, Robert E Jr (1978), "On the Size Distribution of Business Firms." *Bell Journal of Economics*, 9, pp. 508–523.

MacDonald, Glenn M (1988), "The Economics of Rising Stars." *American Economic Review*, 78, pp. 155-166.

— (2001), "The Economics of Has-Beens." *NBER Working Paper* 8464.

Manne, Henry G (1965): "Mergers and the Market for Corporate Control." *Journal of Political Economy*, 73, pp. 110–120.

Mayer, Thomas (1960): "Distribution of Ability and Earnings." *Review of Economics and Statistics*, pp. 189–195.

Miller, Robert A (1984), "Job Matching and Occupational Choice." *Journal of Political Economy*, 92, pp. 1086-1120.

Morris, Philip; Stephen Morrow and Paul Spink (2003): "The New Transfer Fee System in Professional Soccer: An Interdisciplinary Study." *Contemorary Issues in Law*, 5:4, pp. 253–281.

Murphy, Kevin J (1999): "Executive Compensation." in *Handbook of Labor Economics Volume 3*, edited by Orley Ashenfelter and David Card, pp. 2485–2563.

O'Flaherty, Brendan and Aloysius Siow (1995), "Up-or-Out Rules in the Market for Lawyers." *Journal of Labor Economics*, 13, pp. 709–735.

Ostroy, Joseph M (1980): "The No-Surplus Condition as a Characterization of Perfectly Competitive Equilibrium." *Journal of Economic Theory*, 22, pp. 183–207.

— (1984): "A Reformulation of the Marginal Productivity Theory of Distribution." *Econometrica*, 52, pp. 599–630.

Parrino, Robert (1997): "CEO Turnover and Outside Succession: a Cross-Sectional Analysis." *Journal of Financial Economics*, 46, pp. 165–197.

Prescott, Edward C and Michael Visscher (1980), "Organizational Capital." *Journal of Political Economy*, 88, pp. 446-461.

Rosen, Sherwin (1981), "The Economics of Superstars." *American Economic Review*, 71, pp. 845-858.

— (1982), "Authority, Control and the Distribution of Earnings." *Bell Journal of Economics*, 13, pp. 311-323.

— (1992): "Contracts and the Market for Executives." in *Contract Economics*, edited by Werin, Lars and Hans Wijkander, Blackwell, pp. 181–211.

Rosen, Sherwin and Allen Sanderson (2001): "Labour Markets in Professional Sports." *Economic Journal*, 111, pp. F47–F67.

Rottenberg, Simon (1956), "The Baseball Players' Labor Market." *Journal of Political Economy*, 64, pp. 242–258.

Sattinger, Michael (1979), "Differential Rents and the Distribution of Earnings." *Oxford Economic Papers*, 31, pp. 60-71.

— (1984): "Factor Pricing in the Assignment Problem." *Scandinavian Journal of Economics*, 86, pp. 16–34.

— (1993), "Assignment Models of the Distribution of Earnings." *Journal of Economic Literature*, 31, pp. 831-880.

Shimer, Robert (2001): "The Assignment of Workers to Jobs in an Economy with Coordination Frictions." *NBER working paper 8501.*

Simmons, Robert (1997): "Implications of the Bosman Ruling for Football Transfer Market." *Economic Affairs*, 17 (September), pp. 13–18.

Stanley, Robert H (1978), *The Celluloid Empire.* New York. Hastings House.

Szymanski, Stefan and Tim Kuypers (1999), Winners and Losers: The Business Strategy of Football. Penguin Books, London.

Teulings, Coen N (1995): "The Wage Distribution in a Model of the Assignment of Skills

to Jobs." *Journal of Political Economy*, 102, pp. 280–315.

Tinbergen, Jan (1956): "On the Theory of Income Distribution." *Weltwirtschaftliches Archiv*, 77, pp. 155–175.

— (1957): "Some Remarks on the Distribution of Labour Incomes." in *International Economic Papers, no. 1. Translations prepared for the International Economic Association*, edited by Peacock, Alan T et al., Macmillan, pp. 195–207.

Vogel, Harold L (2001), *Entertainment Industry Economics*, 5th ed. Cambridge University Press.

Waldman, Michael (1984), "Job Assignments, Signalling and Efficiency." *Rand Journal of Economics*, 15, pp. 255–267.

– (1990), "Up-or-Out Contracts: A Signaling Perspective." *Journal of Labor Economics*, 8, pp. 230–250.