Managing Careers in Organizations[*]

Preliminary - Please do not circulate without permission.

Rongzhu Ke   Jin Li   Michael Powell

April 14, 2014

Abstract

This paper investigates the optimal personnel policies when careers in organizations are important. Our model extends the classic Shapiro and Stiglitz model to allow for multiple jobs within an organization. Organizations make hiring-, demotion-, promotion-, retention-, and wage-policy decisions. The optimal personnel policies display features of internal labor markets: organizations institute a port of entry and a linear career progression. When promotion opportunities become limited, organizations optimally push out workers at the top. Organizations also become more top heavy. Finally, organizations become less able to respond to outside changes.

1 Introduction

Firms and organizations often attract, motivate, and retain their workforce by offering careers. One problem with this is that the speed of career advancement depends on the opportunities available at the top. When promotion opportunities become limited, workers lower in the firm may get discouraged and leave, and potential hires are reluctant to join. This problem plagues business firms, education institutions, government, military, music groups, sports teams, and any organizations where slot constraints are relevant. Drawing on extensive survey evidence, Cappelli (2008) concludes that "frustration with advancement opportunities is among the most important factors pushing individuals to leave for jobs elsewhere."

Organizations use a variety of personnel policies to address this issue. Mandatory retirement is one way to keep the lines of advancement open. Promotion-from-within is another. Organizations can also add positions at the top or reduce hiring at the bottom to create more opportunities. All these policies, however, incur costs on the organizations. They should therefore be used only when perfect contracting is infeasible. In particular, the use of these policies should be considered jointly as part of the optimal personnel policy in response to imperfect contracting.

This paper investigates the optimal personnel policies under imperfect contracting. Our definition of personnel policies includes hiring, promotion, demotion, retention, and wage policies and will be made more explicit below. By characterizing the optimal personnel policies, the paper examines how the availability of promotion opportunities affect the use of personnel policies, and relatedly, the choice of the production decisions such as the span and the size of the organization.

To address these questions, we develop a simple model by introducing multiple jobs into Shapiro and Stiglitz (1984)'s efficiency-wage model. Homogeneous workers privately choose whether to work or shirk, and the firm can commit to a wage, and therefore a rent, that is tied to the job, coupled with the threat of firing shirking workers in order to provide motivation. Each firm has two types of tasks that have to be performed, and each worker can only perform a single type of task in each period. The two tasks differ in the level of rents that are required to provide motivation, say because one task is more onerous or more difficult to monitor than the other task. We will refer to the task that requires more rents as the **high-rent task** and the other task as the **low-rent task**. The firm's output depends on the number of workers performing each type of task, and the firm maximizes its steady-state profits.

To do so, the firm has to choose the number of positions that will be available for workers performing each task and the wage associated with each task. Further, the firm has to choose how many new workers to hire into each task in each period. At the end of each period, for each

incumbent worker, the firm has to decide whether to retain the worker. If the firm retains the worker, it chooses which task to assign him to in the following period, depending on which task he performed today. The firm's decisions are limited by two key constraints. The firm's decisions must induce workers to exert effort in each task. That is, each worker's incentive-compatibility constraint must be satisfied. Additionally, for the firm to be in steady state, a **flow constraint** must be satisfied: the number of incumbents and new hires that flow into each task must equal the number of workers that flow out of that task in each period.

In steady state, we can think of the future rents promised to a worker as being comprised of two components: the rents the worker will receive if he stays at the firm, which we refer to as the worker's **employment rents**, and the rents the worker would have received if he stayed at the firm but for whatever reason left. We refer to this second component as the worker's **separation rents**. Since workers are motivated by their employment rents and not by their separation rents, worker turnover creates an opportunity. When a worker leaves the firm, the rents promised to him can instead be promised to another worker, in turn providing motivation for that worker. In the Shapiro and Stiglitz model, separation rents are wasted: future rents promised to a departing worker are indeed paid to his replacement, but his replacement comes from the external labor market where the motivational effects of these separation rents cannot be appropriated by the firm.

In our model, separation rents are allocated optimally. The main theoretical result of the paper is that at the optimum, workers performing the low-rent task are motivated solely by the separation rents of the workers performing the high-rent task. In contrast, workers performing the high-rent task are not motivated by separation rents. Any bundle of personnel policies implements a particular allocation of separation rents, so the optimal personnel policy is the bundle of policies that implements the optimal allocation of separation rents. This result delivers many implications regarding the optimal personnel policies, their interactions with the opportunities available, and the interplay of personnel policies with production decisions.

Optimal Personnel Policies   Since workers performing the high-rent task are not motivated by separation rents, they are motivated by an efficiency wage as in Shapiro and Stiglitz. Turnover of these workers creates separation rents, which are probabilistically allocated to incumbent workers performing the low-rent task in the form of random promotions. Since they are allocated only to incumbent workers, this means that the firm does not hire workers into the high-rent task. Rather, workers are only hired into the low-rent task.

As a result, an endogenous hierarchy between the two tasks emerges. The low-rent task is performed in the **bottom job**, which serves as a **port of entry** (Doeringer and Piore, 1971).

Workers remain in the bottom job for a random number of periods after which they are promoted to the **top job**, which involves performing the high-rent task. Separating out the value of the bottom job from the rents that motivate effort for those performing the low-rent task allows the firm to extract all the rents from workers in the bottom job.

Optimally allocated separation rents therefore serve as a glue that binds the jobs in the firm together. The resulting staffing dynamics play the role of workers' trust funds (Akerlof and Katz, QJE 1989) to prevent shirking. Workers effectively post a bond by accepting employment in the bottom job, and their pay is backloaded through a high wage in the top job, which in turn is high enough to motivate effort in the high-rent task. A worker's wages therefore increase upon promotion (Baker, Gibbs, and Holmstrom, 1994). Further, since the wage in the top job serves both as a motivator for those performing the high-rent task as well as a prize that can be extracted from workers in the bottom job through lower wages, the size of the wage increase upon promotion is larger than the difference between the Shapiro-Stiglitz efficiency wages that would be necessary to motivate workers if the jobs were treated independently.

In our model, workers are homogeneous. If workers differ in their ability to perform the high-rent task, the firm may want to hire a high-ability outsider to fill an opening for the top job. The opportunity cost of doing so, however, is that the firm must allocate separation rents to an outsider rather than to an insider, and the firm cannot extract these separation rents from the outsider. Therefore, the firm may be reluctant to hire outsiders unless their contribution relative to insiders exceeds this opportunity cost, giving rise to an "insider-bias" in hiring. (Baker, Gibbs, and Holmstrom, 1994) Moreover, the insider-bias is likely to be particularly severe when the career opportunities are limited.

Opportunities and Personnel Policies   By allocating separation rents to workers inside the firm, the optimal personnel policy connects the career prospects of an individual worker to the careers of his co-workers. In particular, the promotion prospects of a worker in the bottom job depend both on the number of other workers in the bottom job as well as on the number of positions in the top job available next period, which depends on the likelihood that workers at the top will leave the firm. This likelihood in turn depends on the rate of voluntary turnover of workers at the top as well as on the firm's endogenous choice of a retention policy for these workers.

If voluntary turnover at the top is low, then a retention policy that retains all workers at the top yields little separation rents that can be used to motivate workers at the bottom. The firm could, of course, supplement these separation rents with additional efficiency wages for workers at the bottom in order to maintain motivation. However, our main result that workers at the

bottom are motivated solely by separation rents implies that such a policy is suboptimal. Instead, the firm optimally adopts a harsher retention policy and forces a fraction of the workers at the top to leave the firm. Viewed in isolation, adopting a harsher retention policy is a bad idea, since doing so would reduce the employment rents of workers at the top, and their incentive-compatibility constraint would be violated. However, the harsher retention policy is optimally complemented with more generous compensation for workers at the top as well as a more generous promotion policy for workers at the bottom. Policies that clear out space for advancement, such as the aforementioned mandatory-retirement, create separation rents and can indeed be part of an optimal bundle of personnel policies. Whether they are optimal depends on the level of separation rents that results solely from voluntary turnover.

Production Decisions   The optimal personnel policy depends on the firm's personnel needs, which are critically intertwined with the firm's production technology. Coordination-enhancing technological changes such as the IT revolution, can therefore have ramifications for the firm's structure and for the careers of its workers. Yet, in contrast to Neoclassical models of labor demand, technology and product-market factors are not the sole determinants of a firm's organization. Organization optimally balances technological and product-market factors with incentive provision.

Firms may therefore be reluctant to hire workers whose marginal productivity exceeds their wages at the bottom, giving rise to the often-lamented "headcount restriction" policies put in place by human-resource departments. Creating additional positions at the bottom increases output, but it also increases the size of the pool of promotion candidates, reducing the promotion prospects of the existing workers. Further, organizations may optimally become top-heavy: since additional positions at the top improves promotion opportunities at the bottom, positions for which the marginal productivity is smaller than the wage of the worker that occupies it may be nevertheless be created.

Literature Review   This paper is related to several distinct strands of literature. First, it contributes to the literature on efficiency wages–see Akerlof and Yellen (1986) for a collection of important earlier efficiency-wage models and Axelson and Bond (2014) and Board and Meyer-ter-Vehn (2013) for some recent progress. Our paper builds on Shapiro and Stiglitz (1984)'s model which shows that efficiency wages are the optimal stationary scheme for motivating workers. Carmichael (1985) points out that non-stationary compensation schemes such as bonding and backloading can outperform stationary schemes, and a number of papers (Akerlof and Katz (1989), Lazear (1979)) explore how a firm can optimally backload pay within a single job. Our paper studies how firms

can optimally backload pay using job assignments, and in contrast to existing papers, we show that distortions arise even if workers receive no rents.

Since promotions are used to motivate workers, our model shares a number of features prevalent in the literature on tournaments (Lazear and Rosen (1981), Nalebuff and Stiglitz (1983)). For example, as in tournament models, larger wage increases are required in order to motivate effort when promotion prospects are poorer. We contribute to this literature by focusing on how firms can minimize the costs of using promotions as motivation. In particular, we focus on the idea that turnover of workers at the top of an organization creates separation rents that can be used to motivate workers. The firm's objective can therefore be thought of as optimally allocating these separation rents through promotions in order to minimize its overall wage bill. The solution to this problem provides a rationale for why firms might establish internal labor markets (Doeringer and Piore (1971)) and consequently provides one answer to Baker, Jensen, and Murphy (1988)'s puzzle: why are promotions used to motivate when they appear to be dominated by bonus schemes?

Third, our paper contributes to the internal labor market literature by constructing an integrated model that helps explain a number of well-known facts on internal labor markets discussed in the seminal work by Doeringer and Piore (1971), and see Gibbons (1997), Gibbons and Waldman (1999), Lazear (1999), Lazear and Oyer (2013), and Waldman (2013) for reviews of models of internal labor markets. Within this literature, our paper is closely related to Camara and Bernhardt (2009), Krakel and Schottner (2012), MacLeod and Malcomson (1998), Zabojnik and Bernhardt (2001), where hierarchies arise to provide incentives. It is also closely related to Demougin and Siow (1994), DeVaro and Morita (2013), and Zabojnik and Bernhardt (2001), where internal labor markets have implications on the size and span of the firm. One unique feature of our model is it allows the number of top positions to be flexible, and thus the size and span of the firm can be separated.

Finally, there is a sizeable literature looking at how incentive affects organization design (Calvo and Wellisz (1978), Qian (1994), Williamson (1967), Mookherjee (2013)). In these models, the workers stay at a fixed position, and the monitoring technology to detect shirking is the key driver of organization structure. In our model, workers can move up in the hierarchy. As a result, organizational structure is affected by factors that change the availability of promotion opportunities.

2   The Model

A firm and a large mass of identical workers interact repeatedly. Time is discrete and denoted by $t = 1, 2, \ldots$, and all players share a common discount factor $\delta \in (0, 1)$. Throughout, we focus on

the steady state and suppress time subscripts. Production in each period requires two tasks to be performed. Each worker can perform either task if necessary, but he may perform only one task in a given period. A worker performing task i in period t chooses an effort level $e_i \in \{0, 1\}$ at cost $c_i e_i$. A worker who chooses $e_i = 0$ is said to **shirk**, and a worker who chooses $e_i = 1$ is said to **exert effort**. We refer to such a worker as **productive**. A worker's effort choice is his private information, but shirking in task i is contemporaneously detected with probability q_i. If in period t the firm employs masses N_1 and N_2 of productive workers in the two tasks, output is $f(N_1, N_2)$. We assume that f is differentiable, increasing in N_1 and N_2 and is weakly concave.
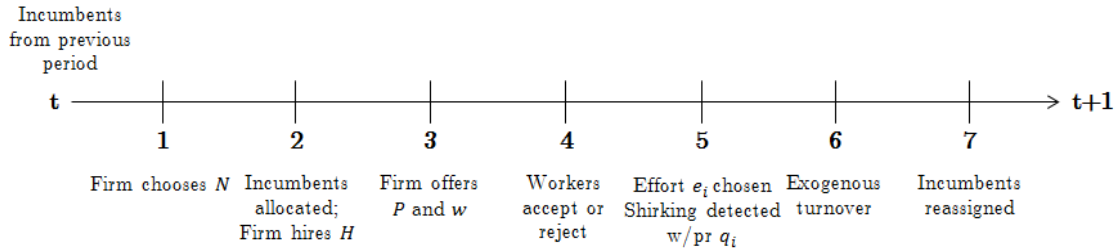

Figure 1: Timing of the stage game.

Figure 1 illustrates the timing of each period. At the beginning of the period, the firm chooses the number of positions N_1 and N_2 for each task. The firm then fills these positions with the incumbent workers and the new hires, where we denote the mass of new hires into task i as H_i, $i = 1, 2$. The firm offers to each worker a contract (w_i, p_{ij}), $i, j = 1, 2$, that includes a **wage policy** and an **assignment policy** that consists of expected promotion, demotion, and retention patterns. We assume that wages are tied to tasks, and denote the wage for task i by w_i. The assignment policy is described by p_{ij}, which denotes the the probability that a worker in task i will take on task j next period if he is not caught shirking. To simplify notation, we assume without loss of generality that a worker who is found shirking is fired with probability 1.

If the worker rejects the contract, he takes his outside option. If he accepts the offer, the wage is paid and the worker chooses his effort level $e_i \in \{0, 1\}$ at cost $c_i e_i$. If he chooses $e_i = 0$, he is caught shirking with probability q_i and fired. For workers not caught shirking, a fraction d_i of workers in task i exogenously leave the firm. We refer to d_i as the **exogenous turnover rate** of workers in task i. The remaining workers (incumbents) are reassigned according to the probability matrix p_{ij}. Notice that if $p_{i1} + p_{i2} < 1$, some workers are asked to leave the firm, receiving their outside utility 0.

3 Efficiency-Wage Benchmark

To provide a benchmark against which to compare our results, we begin by describing what we will refer to as the **efficiency-wage benchmark**. In this benchmark, the firm treats the two tasks independently and offers a wage above the workers' outside options combined with the threat of termination following observed shirking in order to motivate effort. There is no cross-task mobility.

Given a mass $\hat{N}_j$ of workers in task j, the firm chooses N_i and w_i to solve the following program:

$$\max_{N_i, w_i} f\left(N_i, \hat{N}_j\right) - w_i N_i$$

subject to an individual rationality constraint ensuring that the worker receives a greater payoff within the job than outside the job and an incentive-compatibility constraint ensuring that the worker prefers to choose $e_i = 1$ rather than $e_i = 0$. If the worker exerts effort in each period, he receives a total payoff of v_i in the job, where

$$v_i = w_i - c_i + (1 - d_i)\,\delta v_i.$$

That is, in each period, he receives the wage w_i and incurs the effort costs c_i. With probability d_i, he exogenously leaves the firm, but with the remaining probability, he remains in the job and receives v_i again the following period. The worker will exert effort as long as

$$v_i \geq w_i + (1 - q_i)(1 - d_i)\,\delta v_i.$$

A worker who shirks avoids incurring the cost c_i but is caught and fired with probability q_i. To maximize its profit, the firm chooses wages w_i, or equivalently, payoffs v_i, to ensure the incentive-compatibility constraint holds with equality. Given the resulting efficiency wage, the firm hires workers until the marginal benefit of an additional worker is equal to this wage. Finally, the firm hires a mass of new workers into each task to exactly offset the mass of workers who are exogenously separated from that task. The resulting solution, which we refer to as the Shapiro-Stiglitz solution and denote with the superscript ss, is described in the following lemma.

LEMMA 0. *A firm maximizing its profits separately over the two tasks chooses wages* $w_i^{ss} = (1 + (1 - (1 - d_i)\,\delta)\,/\,(q_i\,(1 - d_i)\,\delta))\,c_i$ *to provide rents* $v_i^{ss} = c_i/(\delta q_i\,(1 - d_i))$ *to each worker performing task* $i = 1, 2$. *The firm hires* $H_i^{ss} = (1 - d_i)\,N_i^{ss}$ *workers, where* $\partial f\left(N_i^{ss}, N_j^{ss}\right)/\partial N_i = w_i^{ss} > c_i$.

Lemma 0 reproduces the following observations from Shapiro and Stiglitz. First, the firm has to give workers rents to provide incentives to exert effort. Second, the level of rents required to

provide incentives, and hence the wage level, increases in the turnover rate d_i and decreases in the firm's monitoring ability, q_i. Third, the firm optimally chooses an employment level for each task that is smaller than the socially optimal level, which would satisfy $\partial f / \partial N_i = c_i$. Moreover, the gap between the firm's employment-level choice and the socially optimal level is greater for jobs that require higher rents to provide incentives. To facilitate our discussion, define

$$R_i = \frac{c_i}{(1 - d_i)\, \delta q_i} \qquad\qquad \text{(Shapiro-Stiglitz Rents)}$$

as the **Shapiro-Stiglitz rents** associated with task i. We assume throughout that $R_2 > R_1$, so that in the efficiency-wage benchmark, more rents are provided to workers in task 2 than in task 1.

4   Managing Careers

In the efficiency-wage benchmark, the firm chooses only a mass of workers to perform each task and a wage paid to each of these workers. In this section, we study more general personnel policies that allow for reassignment across tasks. We demonstrate that the firm always performs better by linking the tasks together than by treating them independently. Moreover, we characterize the firm's optimal choices and show that they lead to features characteristic of internal labor markets.

4.1   Preliminaries

The firm chooses wage, hiring, and assignment policies jointly to maximize its steady-state profits

$$f\left(N_1, N_2\right) - w_1 N_1 - w_2 N_2.$$

As in the benchmark, denote v_i as the expected discounted payoff of a worker performing task i. The firm maximizes its profits subject to the following constraints.

Promise-Keeping Constraints. Assuming that workers always exert effort, their payoffs must satisfy the following equations

$$\begin{aligned} v_1 &= w_1 - c_1 + (1 - d_1)\, \delta\left(p_{11} v_1 + p_{12} v_2\right); & \text{(PK-1)} \\ v_2 &= w_2 - c_2 + (1 - d_2)\, \delta\left(p_{21} v_1 + p_{22} v_2\right). & \text{(PK-2)} \end{aligned}$$

Individual-Rationality Constraints. To ensure that workers prefer working for the firm rather than taking their outside options, they must receive a greater payoff from doing so. That is,

$$\begin{aligned} v_1 &\geq 0; & \text{(IR-1)} \\ v_2 &\geq 0. & \text{(IR-2)} \end{aligned}$$

Incentive-Compatibility Constraints. Workers prefer to exert effort if the following constraints are satisfied:

$$w_1 - c_1 + (1 - d_1)\,\delta\,(p_{11}v_1 + p_{12}v_2) \;\geq\; w_1 + (1 - q_1)\,(1 - d_1)\,\delta\,(p_{11}v_1 + p_{12}v_2)\,;$$
$$w_2 - c_2 + (1 - d_2)\,\delta\,(p_{21}v_1 + p_{22}v_2) \;\geq\; w_2 + (1 - q_2)\,(1 - d_2)\,\delta\,(p_{21}v_1 + p_{22}v_2)\,,$$

where we use the fact that if the worker leaves the firm, he receives a payoff of 0. For notational convenience, we rewrite these constraints as follows:

$$p_{11}v_1 + p_{12}v_2 \;\geq\; c_1/\,(1 - d_1)\,\delta q_1 = R_1; \qquad\qquad\text{(IC-1)}$$
$$p_{21}v_1 + p_{22}v_2 \;\geq\; c_2/\,(1 - d_2)\,\delta q_2 = R_2, \qquad\qquad\text{(IC-2)}$$

where recall R_i is the Shapiro-Stiglitz efficiency rent associated with task $i = 1, 2$. It will sometimes be useful to denote the **excess rents** offered in task i by $\Delta_i \equiv p_{i1}v_1 + p_{i2}v_2 - R_i$.

Flow Constraints. In the steady state, the number of workers in a particular task must remain constant. Given the hiring and assignment policies, the following constraints ensure that the mass of workers flowing into each task equals the mass of workers flowing out of that task:

$$(1 - d_1)\,p_{11}N_1 + (1 - d_2)\,p_{21}N_2 + H_1 \;=\; N_1; \qquad\qquad\text{(FL-1)}$$
$$(1 - d_1)\,p_{12}N_1 + (1 - d_2)\,p_{22}N_2 + H_2 \;=\; N_2, \qquad\qquad\text{(FL-2)}$$

where $H_i \geq 0$ is the mass of new workers hired into task i. In addition, since the p_{ij} are probabilities, they must be non-negative, and

$$p_{i1} + p_{i2} \leq 1, \text{ for } i = 1, 2.$$

A fraction of workers who are neither caught shirking nor exogenously separated from the firm are fired if $p_{i1} + p_{i2} < 1$.

We solve the firm's problem in two steps. First, we fix the number of positions for each task, and we solve for the firm's cost-minimizing levels of p_{ij}, H_i, and v_i. In the second step, we allow the firm to optimize over N_1 and N_2. Throughout, we refer to the ratio N_2/N_1 as the firm's **span** and $N_1 + N_2$ as the firm's **size**. The vector $H = [H_i]_i$ is the firm's **hiring policy**, and the rent vector $v = [v_i]_i$ determines the firm's **wage policy** for a given **assignment policy** $P = [p_{ij}]_{ij}$. The values $1 - p_{i1} - p_{i2}$ represent the probability that the firm asks a productive worker in task i to leave the firm, so the assignment policy P represents the firm's **promotion, demotion, and**

retention policies. If $1 - p_{i1} - p_{i2} = 0$, we say that task i has **full job security**; that is, a worker performing task i departs the firm only for exogenous reasons unless he is caught shirking. We refer to a collection (H, w, P) as a **personnel policy**.

4.2   Optimal Personnel Policy

Given the span and size of the firm, the firm chooses an optimal personnel policy. This involves choosing hiring, wage, and assignment policies to minimize the steady-state wage bill $N_1 w_1 + N_2 w_2$. In this section, we describe the optimal personnel policy and provide intuition for the results. Formal derivations of the results are included in the appendix.

We have assumed that the Shapiro-Stiglitz rents associated with task 2 exceed the Shapiro-Stiglitz rents of task 1 (i.e., $R_2 > R_1$). Throughout this section, we will assume (and formally verify in the appendix) that under the optimal personnel policy, the rents provided in task 2 exceed those provided in task 1 (i.e., $v_2^* > v_1^*$). For reasons that will soon become clear, we refer to task 1 as the **bottom job** and task 2 as the **top job**. We also refer to workers who perform task 1 as **bottom workers** and those who perform task 2 as **top workers**. If $N_2 d_2 > N_1 (1 - d_1)$, so that there are not enough incumbent bottom workers to fill all the top-job vacancies generated by voluntary turnover, we say that **top jobs are abundant**. Otherwise, **top jobs are scarce**. Whenever top jobs are scarce, the firm will never hire directly into the top job.

LEMMA 1. *If top jobs are scarce, all new workers are hired into the bottom job (i.e., $H_2^* = 0$).*

To see why firms prefer to hire workers into the bottom job, notice that a vacancy in the top job can be filled either by directly hiring into the top job or by hiring into the bottom job and promoting an incumbent bottom worker. We refer to the former policy as **replacement hiring** and the latter as **push hiring**. Replacement hiring requires the firm to provide a rent of v_2^* to the new worker. In contrast, push hiring only requires the firm to provide a rent of v_1^* to the new worker. Both policies preserve the flow constraint, since the vacancy in the top job is filled and the mass of bottom workers remains constant. Push hiring also makes the incentive-compatibility and participation constraints for bottom workers easier to satisfy, because it involves a higher promotion probability. Promoting from within helps motivate bottom workers using the rents associated with the top job, which in turn allows the firm to lower the wages associated with the bottom job.

Next, we describe workers' careers within the firm. There will be two important cases to consider, which are related to the rents that are freed up by voluntary turnover at the top. Consider the efficiency-wage benchmark in which there are no promotions, and each task is associated with full job security and is paid a wage that corresponds to its Shapiro-Stiglitz rents. At the end of

any period, a mass $d_2 N_2$ workers depart from the top, which frees up an amount $d_2 N_2 R_2$ of rents that may be reallocated. Additionally, at the end of the period, there are a mass $(1 - d_1) N_1$ of incumbent bottom workers who must be promised rents R_1 to exert effort. We say that there are **sufficient separation rents** if $d_2 N_2 R_2 \geq (1 - d_1) N_1 R_1$. In this case, the prospect of receiving rents from exogenous turnover of the top job is sufficient to motivate the workers at the bottom job. If this condition is not satisfied, we say that there are **insufficient separation rents**. The next lemma describes workers' careers when there are sufficient separation rents.

LEMMA 2. *When there are sufficient separation rents, in an optimal personnel policy, bottom workers receive zero rents and top workers receive Shapiro-Stiglitz rents. There are no demotions, and workers receive full job security.*

Lemma 2 illustrates the benefits of using promotions to reduce rents given to new workers. In the efficiency-wage benchmark, efficiency wages motivate workers and also determine their equilibrium payoffs. By using promotions, the firm can separate incentive provision from equilibrium payoffs for bottom workers. Since top workers are never promoted, they must receive at least Shapiro-Stiglitz rents in order to exert effort. When there are sufficient separation rents, promotion prospects alone provide enough motivation for bottom workers. The firm then sets the bottom wage just high enough to induce participation, leaving bottom workers with no rents. Bottom workers' per-period payoffs are lower than their outside options, but they are willing to work for the firm, because of the prospect of being promoted to the top job.

If top workers were demoted or asked to leave the firm with positive probability, Shapiro-Stiglitz rents would not be sufficient to motivate them. Since they receive Shapiro-Stiglitz rents under the optimal personnel policy, it must therefore be the case that they are never demoted, and they receive full job security. For bottom workers, full job security is optimal, but not uniquely so. The firm could, in each period, fire bottom workers with positive probability and increase the promotion prospects of the remaining bottom workers correspondingly. As long as the promotion probability of bottom workers at the *beginning* of each period remains unchanged, workers are motivated, and the firm's wage bill is the same. If hiring or firing were exogenously costly, full job security for bottom workers would be uniquely optimal. This is because full job security for bottom workers minimizes the mass of workers who are hired and fired.

Workers' career patterns are different in firms in which there are insufficient separation rents. We explore these patterns in the next lemma.

LEMMA 3. *When there are insufficient separation rents, in an optimal personnel policy, bottom workers receive zero rents, and top workers receive rents that exceed Shapiro-Stiglitz rents. There*

are no demotions, bottom workers receive full job security, and top workers do not receive full job security.

When there are insufficient separation rents, the personnel policies described in Lemma 2 no longer provide enough motivation for bottom workers. To increase the incentives for bottom workers, the firm could in principle pay efficiency wages at the bottom. Lemma 3 shows that doing so is never optimal–in the optimal personnel policy, bottom workers receive zero rents. The firm provides additional motivation entirely by increasing bottom workers' promotion prospects. To do this, the firm fires top workers with positive probability in each period and offers them rents that exceed Shapiro-Stiglitz rents. This increase in turnover at the top allows the firm to increase the promotion prospects for bottom workers. Coupled with the associated increase in rents upon promotion, such a policy maintains motivation for both top workers and bottom workers.

To see in another way why the firm prefers to use promotion incentives rather than efficiency wages to motivate bottom workers, notice that if efficiency wages are paid at the bottom, the firm must be giving rents to new workers. Doing so constitutes a pure loss for the firm. In contrast, the firm can recapture increased wages for top workers by lowering wages for bottom workers. Raising wages for top workers backloads a worker's pay and therefore is more effective than using efficiency wages throughout the firm. Moreover, if the firm offers rents that exceed Shapiro-Stiglitz rents for the top job, top workers' incentive constraints would be slack if they were given full job security. The firm can therefore reduce top workers' job security, increase bottom workers' promotion prospects, and decrease bottom workers' wages still further.

The firm weakly prefers to fire top workers than to demote them. Firing and demoting top workers creates promotion opportunities for bottom workers, but they also reduce the value that workers place on the top job. The relative amount by which they do so depends on how top workers' outside options compare to the value of the bottom job, which under the optimal personnel policy is equal to the bottom workers' outside options. Firing workers is therefore preferred whenever top workers' outside options exceed bottom workers' outside options. For demotions to be optimal, it has to be the case that bottom workers' outside options are greater than top workers' outside options. In our model, both are zero.

Proposition 1 summarizes the main features of an optimal personnel policy.

PROPOSITION 1. *An optimal personnel policy has the following features. (i) Hiring occurs only in the bottom job, unless top jobs are abundant. (ii) There is a well-defined career path: bottom workers stay at the bottom job or are promoted. Top workers are never demoted but may be fired. (iii) Bottom-job wages are lower than the Shapiro-Stiglitz wages for the low-rent task. Top-job wages*

exceed the Shapiro-Stiglitz wages for the high-rent task whenever there are insufficient separation rents.

The analysis in this section takes as given the number of workers performing each task in the firm and characterizes the optimal personnel policies that minimize the costs of motivating these workers. We turn next to the firm's choice of the number of workers.

5   Optimal Production

In this section, we examine how the availability of promotion opportunity affects the span and size of the firm.   Notice that Proposition 1 characterizes the optimal personnel policies, taking the firm's size and span as given and therefore results in a **labor-cost function**, $W(N_1, N_2)$. Given the wage bill and the production function, characterizing the firm's optimal production decisions are straightforward. Before proceeding to do so, we first discuss several properties of the wages. The exact expressions are simpler if we assume that $q_1 = q_2 = 1$–that is, monitoring is perfect so that a worker who shirks is always caught–but doing so is otherwise inconsequential.

COROLLARY 1. *The following holds.*

    (i):   When $N_1/N_2 \le d_2c_2/((1-d_2)c_1)$, the wages are given by $w_1 = c_1 - d_2c_2N_2/((1-d_2)N_1)$ and $w_2 = c_2/((1-d_2)\delta)$.   The labor-cost function is given by

$$W(N_1, N_2) = c_1N_1 + \frac{1-\delta d_2}{1-d_2}\frac{1}{\delta}c_2N_2.$$

    (ii):   When $N_1/N_2 > d_2c_2/((1-d_2)c_1)$, the wages are given by $w_1 = 0$ and $w_2 = (c_1N_1 + c_2N_2)/(\delta N_2)$. The labor-cost function is given by

$$W(N_1, N_2) = \frac{1}{\delta}c_1N_1 + \frac{1}{\delta}c_2N_2.$$

As a consequence of Proposition 1, Corollary 1 shows that the labor-cost function depends on whether there are sufficient separation rents.   There are sufficient separation rents if and only if the span of the firm, N_1/N_2, is smaller than a cutoff $(d_2c_2/((1-d_2)c_1))$. It follows that the sufficient-separation-rents case arises when the span is small and/or the turnover rate of the top job is high. To simplify our discussion below, we denote the cutoff span as

$$\kappa \equiv d_2c_2/((1-d_2)c_1).$$

Notice that the effect of span on wages differs in the two cases. When there is sufficient separation rents, the top wage (w_2) is equal to the Shapiro-Stiglitz wage. The span only affects the

bottom wage. In contrast, when there is insufficient separation rents, the bottom wage does not depend on the span. Instead, the top wage is increasing and linear in span.

Regardless of size of the span, a key feature of the labor-cost function is that it displays constant return to scale, just as in the neoclassical labor cost function without incentives concerns. Unlike the neoclassical cost function, the marginal cost of having an extra position is not equal to the wage of the position. Since the wages are backloaded, the marginal cost of having a bottom position is higher than the bottom wage and the opposite is true for the top position. Moreover, turnover rate matters for the labor-cost function whereas it is irrelevant in the neoclassical cost function. This has consequence on the span of the firm which we come to next.

5.1 Span

To determine the optimal span, we now fix the number of the top positions (N_2) and find the optimal bottom positions $N_1^*(N_2)$. This is carried out by comparing the value of marginal product of the bottom position with its marginal cost. Notice that Corollary 1 implies that the marginal cost of the bottom position is given by

$$\frac{dW(N_1, N_2)}{dN_1} = \begin{cases} c_1 & if \quad N_1 < \kappa N_2 \\ \frac{c_1}{\delta} & if \quad N_1 > \kappa N_2, \end{cases}$$

which is illustrated in Figure 2. The marginal cost is equal to c_1 to the left of κN_2 and jumps to c_1/δ to the right of κN_2. Once the number of bottom workers exceeds κ, the firm must increase the top wage to maintain the the incentive constraint of the bottom workers.
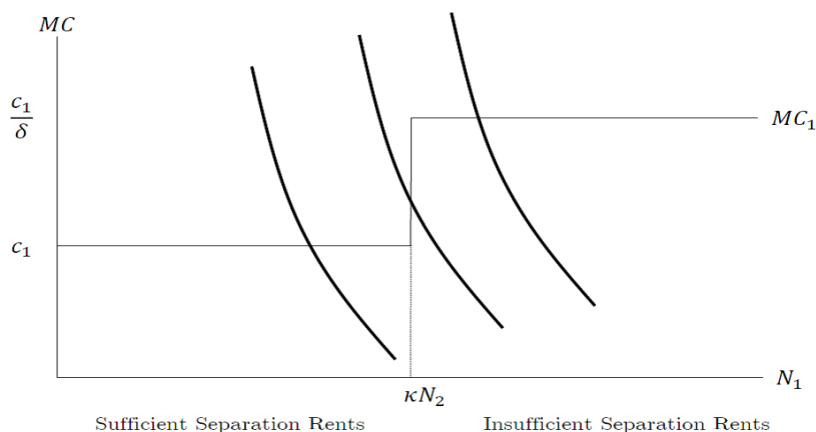
Figure 2: The MC_1 curve represents the marginal cost of creating an additional bottom position when the number of top positions is fixed. The three curved lines represent the marginal benefits of doing so for different levels of technology.

To see why the marginal cost of the bottom position is c_1 when there is sufficient separation rents, recall from Lemma 2 that bottom workers have zero rents. Moreover, hiring an extra bottom worker in this case does not require adjustment at the top. In particular, the firm does not need to increase the wage of the top job above the Shapiro-Stiglitz wage to maintain the incentives of the bottom workers. The firm's cost of having one more bottom position, therefore, is only that of compensating the worker for his effort cost (since the opportunity cost is normalized to 0).

In contrast, when separation rents are insufficient, adding another bottom position has a spillover effect on the top job. In particular, each additional hire now requires the firm to increase the wage at the top (and/or increase the number of top positions) to maintain the incentives of the bottom workers. The adjustment at the top implies that marginal cost of the bottom job is higher than c_1 even if the bottom worker receives no rents.

Given the marginal cost curve, the firm can determine the optimal number of bottom positions $(N_1^*(N_2))$ by finding where the marginal benefit and cost curve intersect. As Figure 2 illustrates, there are three cases, depending on whether at the intersection $N_1^*(N_2)$ is smaller than, bigger than, or equal to κN_2. The next corollary illustrates how the turnover rate affects the span $N_1^*(N_2)$.

COROLLARY 2. *There exists a $\underline{d}_2$ and $\overline{d}_2$ such that the following holds.*

(i): When $d_2 \geq \overline{d}_2$, the number of bottom positions satisfies $\partial f(N_1^*(N_2), N_2)/\partial N_1 = c_1$, and the span $N_1^*(N_2)/N_2 < \kappa$.

(ii): When $d_2 \in [\underline{d}_2, \overline{d}_2)$, the number of bottom positions is given by $N_1^*(N_2) = \kappa N_2$.

(iii): When $d_2 < \underline{d}_2$, the number of bottom workers is given by $\partial f(N_1^*(N_2), N_2)/\partial N_1 = c_1/\delta$, and the span $N_1^*(N_2)/N_2 > \kappa$.

For any N_2, the socially optimal number of bottom positions $(N_1^{FB}(N_2))$ satisfies $\frac{df(N_1^{FB}(N_2), N_2)}{dN_1} = c_1$. Corollary 2 then implies that the number of bottom positions never exceeds the socially optimal level. It is equal to the socially optimal level only when the turnover rate is above $\overline{d}_2$, so the marginal product curve intersects the marginal cost curve to the left of κN_2 (part (i)). This corresponds to the sufficient-separation-rent case.

When the turnover rate d_2 drops, κN_2 moves to the left and at some point the marginal product curve intersects that marginal cost curve at the vertical part (part (ii)). In this case, productive efficiency requires the firm to hire more than κN_2 workers. But to do so, the firm must increase

the wage of the top job to maintain the incentives of bottom workers. This extra cost overweighs the gain in production, and as a result, the firm does not expand beyond κN_2. In other words, the concern for providing career incentives prevents the firm from adjusting its production decision in response to productivity gain.

As the turnover rate drops further, eventually the marginal product curve intersects the marginal cost curve to the right of κN_2 (part (iii)). In this case, the span of the firm is again inefficiently small (for the given N_2) since the value of the marginal product is above c_1. This inefficiency arises because for each new bottom position, the firm must increase the wage at the top and terminate some top workers. In particular, the firm pushes out just enough workers so that the total turnover rate does not change with d_2 and is exactly equal to $\underline{d}_2$.

The reason for a constant total turnover rate follows from part (iii) of Corollary 2, which shows that the total number of bottom positions is always the same in this case. To keep these bottom workers motivated, the minimal total turnover rate associated with the optimal internal labor market is at least $\underline{d}_2$ so as to provide sufficient promotion opportunity. Once d_2 falls short of $\underline{d}_2$, the firm must increase the involuntary turnover rate to insure that the total turnover rate is at least $\underline{d}_2$. Of course, the firm has no incentive to raise the total turnover rate above $\underline{d}_2$ because doing so leads to an unnecessarily high wage at the top and lowers its profit. In other words, $\underline{d}_2$ is the lowest total turnover rate for the firm.

For a fixed number of top positions N_2, Corollary 2 shows that the span increases with the turnover rate at the top. In other words, firms are more top-heavy when the turnover rate is smaller. Notice that the number of top positions can be difficult to adjust in some cases–in small stores and plants there is typically a single manager ($N_2 = 1$). In these cases, the span of the firm moves one-to-one with size of the firm. Corollary 2 then suggests that both the size and span are inefficiently small when the turnover rate is low. In general, the turnover rate also affects the number of top position. Its effect on the span and size of the firm are more therefore nuanced. We come to these issues next.

5.2   Size

Now we study the effect of turnover rate on the size of the firm. When the top and bottom positions are chosen jointly, the isoquant is tangent to the isocost curve at the optimal levels of positions. To determine the isocost curve, recall in Corollary 2 that the marginal cost of adding a new worker depends on whether there is sufficient separation rents, i.e., whether N_1/N_2 is smaller than κ. The slope of the isocost curve, therefore, depends on the relative number of positions for each job.

Figure 3 draws a isocost curve arising from the labor-cost function, and for the same cost level, the corresponding isocost curve for the contractible-effort case (when there is no incentive issues). The downward-sloping dotted line represents the isocost curve if effort were contractible. This socially optimal isocost curve is a straight line with slope $-c_1/c_2$. When incentive matters, however, the resulting isocost curve is kinked around the $N_1/N_2 = \kappa$ ray. To the left of the ray, the slope of the isocost is given by $-(1 - d_2)\delta c_1/c_2$, which is smaller (in absolute value) than contractible-effort case. To the right of the ray, the slope is equal to $-c_1/c_2$, paralleling the socially optimal isocost curve.
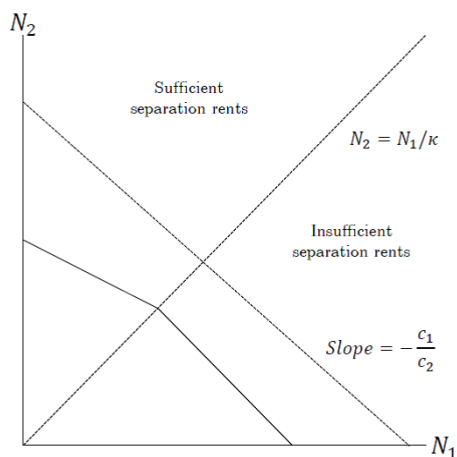

Figure 3: This figure plots two iso-cost curves. The straight, downward-sloping curve is the iso-cost curve that would be obtained if effort were contractible. The kinked curve is the iso-cost curve that is obtained in our model. Note that these two iso-cost curves represent the same level of production.

It is immediate from Figure 3 that the level of production will be distorted since for a given cost level, the maximum quantity that can be produced at that cost is lower. In addition, there will be distortion in the production mix between N_1 and N_2. This follows because the isocost curve with contractible effort has a constant slope of $-c_1/c_2$, and therefore, the socially optimal marginal rate of technical substitution is always equal to c_1/c_2. The slope of the isocost curve in our model, in contrast, depends on the firm's span (N_1/N_2). This implies that the marginal rate of technical substitution in our model will be away from the socially optimal level in some cases. Corollary 3 captures these distortions in levels and production mix by characterizing the conditions for the optimal number of positions.

COROLLARY 3. *Suppose the production function f is concave. The optimal number of positions N_1^* and N_2^* falls into one of three cases below.*

(i): When $N_1^* < \kappa N_2^*$, the optimal numbers of position satisfy

$$\frac{\partial f\left(N_1^*, N_2^*\right)}{\partial N_1} = c_1; \quad \frac{\partial f\left(N_1^*, N_2^*\right)}{\partial N_2} = \frac{c_2}{\delta\left(1 - d_2\right)}.$$

(ii): When $N_1^* = \kappa N_2^*$, the optimal numbers of position satisfy

$$\frac{\partial f\left(N_1^*, N_2^*\right)}{\partial N_1} + \frac{1}{\kappa}\frac{\partial f\left(N_1^*, N_2^*\right)}{\partial N_2} = \frac{c_1}{\delta} + \frac{c_1}{\delta d_2}.$$

(iii): When $N_1^* > \kappa N_2^*$, the optimal number of position satisfy

$$\frac{\partial f\left(N_1^*, N_2^*\right)}{\partial N_1} = \frac{c_1}{\delta}; \quad \frac{\partial f\left(N_1^*, N_2^*\right)}{\partial N_2} = \frac{c_2}{\delta}.$$

Corollary 3 is the counterpart of Corollary 2 where the top position is fixed. As in Corollary 2, the marginal product of the bottom position depends on whether the optimal span N_1^*/N_2^* is smaller than, equal to, or greater than κ. Unlike Corollary 2, we cannot directly order the span N_1^*/N_2^* according to the turnover rate d_2 since N_2^* is also affected by d_2. Nevertheless, the marginal rate of technical substitution $\left(\frac{\partial f\left(N_1^*, N_2^*\right)}{\partial N_1}\middle/\frac{\partial f\left(N_1^*, N_2^*\right)}{\partial N_2}\right)$ is higher in part (iii) than in part (i), suggesting that the firm is more top-heavy when there is there is insufficient separation rents.

To better link the turnover rate with the firm size and span, we next consider a specific example, where the production function is given by $f\left(N_1, N_2\right) = \alpha_1 \log N_1 + \alpha_2 \log N_2$. In this case, we can compute the number of top- and bottom positions explicitly.

COROLLARY 4. *Let $f\left(N_1, N_2\right) = \alpha_1 \log N_1 + \alpha_2 \log N_2$. The following holds.*
(i): When $d_2 \geq \frac{1}{\delta}\frac{\alpha_1}{\alpha_1+\alpha_2}$, we have $N_1^* = \frac{\alpha_1}{c_1}$ and $N_2^* = \frac{(1-d_2)\alpha_2\delta}{(c_2(1-\delta d_2))}$.
(ii): When $d_2 \in \left(\frac{\alpha_1}{\alpha_1+\alpha_2}, \frac{1}{\delta}\frac{\alpha_1}{\alpha_1+\alpha_2}\right)$, we have $N_1^* = \frac{d_2(\alpha_1+\alpha_2)\delta}{c_1}$ and $N_2^* = \frac{(1-d_2)(\alpha_1+\alpha_2)\delta}{c_2}$.
(iii): When $d_2 \leq \frac{\alpha_1}{\alpha_1+\alpha_2}$ we have $N_1^* = \frac{\delta\alpha_1}{c_1}$ and $N_2^* = \frac{\delta\alpha_2}{c_2}$.

The three parts of Corollary 4 again correspond to those in Corollary 3 where the span of the firm N_1^*/N_2^* is smaller than, equal to, or larger than κ. Unlike Corollary 3, we now have explicit formula for the number of positions in terms of the exogenous turnover rate at the top. These formula highlight the following properties of the size and span of the firm. First, the size of the firm is small than the socially optimal level. Notice that the first-best number of positions are given by $N_1^{FB} = \alpha_1/c_1$ and $N_2^{FB} = \alpha_2/c_2$. Therefore, only the number of bottom position when the turnover rate is high is equal to the first-best. In all other cases, both the number of top- and bottom-positions are too small.

Second, higher exogenous turnover rates can increase the size of the firm. Notice that a higher turnover rate actually reduces the number of positions in the Shapiro-Stiglitz model, and similarly, the number of top positions are weakly decreasing in d_2 here. However, the higher turnover rate at the top makes it easier to motivate the bottom workers, and therefore the number of bottom positions increases. When the turnover rates are in the intermediate range (part (ii)), in particular, the gain in the bottom positions overweighs the loss in the top. The total size of the firm therefore increases in d_2.

Finally, even if the total output is additive in the two jobs, the number of bottom positions depend on the productivity of the top job and vice versa. The provision for career incentives therefore creates connections in the two jobs even if they are productively independent. In particular, lower productivity at the top reduces the number of top positions. As a result, the number of bottom positions must be reduced to give the bottom workers sufficient promotion opportunities.

6   Applications

In this section, we relate the predictions of the model to observable patterns on hiring, job transition, compensation, span, and size of the firm. Our focus is on how the availability of promotion opportunities affects the use of different personnel policies, and relatedly, its effect on how firms adjust their wages, spans, and sizes in response to changes in exogenous environments.

6.1   Hiring Policies

In our model, all hiring occurs at the bottom job, which serves as a port of entry into the firm. Port of entry is an important feature of the internal labor market (Doeringer and Piore (1971)). Empirically, however, the evidence for port of entry appears to be mixed (Baker, Gibbs, and Holmstrom (1994)). Nevertheless, port of entry appears to be a good description of the functioning of the labor market in many situations; see Milgrom and Roberts (1992) for a discussion of the pilot labor markets and Ken'ichi and Hiroyuki (1988) for the Japanese internal labor markets.

Notice that port of entry can be viewed as an extreme form of insider-bias: for top positions, only the insiders are considered. When workers are heterogeneous, the model would predict that the firms favor insiders for the top job. The insider-bias has been documented in a number of studies; see for example, Huson, Malatesta and Parrino (1994) and Oyer (1994). Bond (2013), however, reports that only about one quarter of firms favor insiders in hiring.

Our model implies that the insider bias is likely to be more prevalent when the promotion opportunities for the bottom workers are more limited. In particular, a lower turnover rate at the

top is like to exacerbate the insider-bias. Relatedly, to the extent that the promotion opportunity is more limited when the span of the firm is large, our model supports the finding that insider-bias is stronger when the span of the firm is bigger; see DeVaro and Morita (2013).

6.2 Career Path

Our model also predicts that workers follow a well-defined career trajectory within the firm. We discuss below the way the firm manages its workforce in terms of its demotion, promotion, and retention policy.

6.2.1 Demotion

Our model predicts that there is no demotion. This extreme prediction results because the workers are homogeneous. If workers differ in their abilities, our model then suggests that demotion is rare. The empirical literature generally finds that demotion is rare but there are also notable exceptions. Baker, Gibbs, and Holmstrom (1994) show that demotion is rare by using data from a large U.S. firm. Seltzer and Merrett (2000) report similar findings using data from an Australian bank, and Treble et al. (2001) also find that demotion is rare in the British operation of a service-sector firm. Dohmen et al (2004) and Lin (2005), however, find that demotions are more common than previous studies suggested.

6.2.2 Promotion

Part of firm's promotion policy has been discussed earlier, our model predicts that the firms will promote from within. In addition, our model implies that the effect of turnover on promotion depends on the availability of exiting promotion opportunities. When there are enough promotion opportunities, a higher turnover rate increase the promotion prospect because more top positions become available (part (i) of Corollary 4). When promotion opportunities are limited, however, a higher turnover rate can actually reduce the promotion probability (part (ii) of Corollary 4). This follows because a higher turnover rate can lead to a larger span. The negative effect from the larger span dominates, leading to a lower promotion probability.

6.2.3 Retention

Our model shows that it is optimal for the firm to push workers out even if they are not caught shirking. A number of personnel policies, such as buyout and mandatory retirement age can all be viewed as ways to push workers out so as to create more promotion opportunities for bottom workers. In the U.S., the Age Discrimination in Employment Act was modified in 1979 and increased

the mandatory retirement age from 65 to 70 for tenured professors. Lazear (1998, p65) mention that "In order to induce older professors to leave voluntarily, a number of universities offered buyout plans where professors beyond 55 years of age were offered a sweetened pension if they would retire immediately." The buyout policy helps free up positions, and thus, can facilitate hiring. Relatedly, U.S. Department of Labor (1981) reports that

"When firms were asked their reasons for using mandatory retirement all firms, but particularly large firms, put greatest emphasis on assuring promotional opportunities for younger workers. The promotion rationale was stronger than the oft-cited rationale for retiring unproductive workers."

Our model suggests that these policies are more likely to be used when the promotion opportunities are limited.

6.3 Compensation Policies

When the promotion prospect of younger workers becomes more limited, the firm can adjust its compensation policy to motivate the workers in a number of ways. First, a direct implication of our model is that the pay increase needs to be higher to motivate them. Our model, therefore, suggests that there is a negative relationship between promotion probability and wage increase upon promotion. This result, of course, is of the same spirit as standard tournament models, and there is some related evidence broadly supporting it. Grund (2005) examined the yearly promotion rates and the wage increase upon promotion for four firms and find a negative correlation. Leonard (1990) shows that there is a negative correlation between promotion rates and wage increase upon promotion for executives in large U.S. companies.

A related implication of our model is that a larger span can lead to higher wages upon promotion since a larger span reduces the promotion prospect. Empirically, Garicano and Hubbard (2009) show that both the within-firm wage inequality and the associate/partner ratio have increased from 1977 to 1992. Garicano and Hubbard (2009, footnote 9) also report that entry into the law profession dramatically increased in the 1970s, suggesting a large crop of recently partnered lawyers may have limited the promotion aspect of young associates. This makes it more important for the firm to raise the wage increase upon in providing career incentive.

In addition, our model is consistent with the observation that both the bottom and top wages are higher in larger and more productive firms. This sheds light on a puzzle raised by Rebitzer and Taylor (1995). They find that the wages of associates and partners are higher at larger law firms and interpreted the finding as evidence against efficiency wages. Their reasoning is that larger firms allow for better backloading, so associates in these firms should have lower pay. In our model,

however, firms with more productive bottom workers can result in bigger spans and more limited promotion prospect for their workers. Corollary 3 then implies that the wages of both the bottom- and top workers are higher in these firms.

6.4 Span

Our model implies that limited promotion opportunities tend to make firms more top heavy. The empirical evidence on this front is somewhat scant in part because data on the hierarchical structure of firms are costly to obtain. At the very top of the hierarchy, Rajan and Wulf (2006) and Guadalupe and Wulf (2010) find that the spans of control of the CEOs have recently increased. It is not clear to what extent this is related to changes in promotion prospects of insiders, although Kaplan and Minton (2008) find that CEO turnover rate has also increased recently.

Notice that in our model, firms become top heavy even if they can increase the wage level to motivate the workers. When there is limit to the degree money can be used to motivate the workforce, our model suggests that firms and organizations can become even more top heavy. One possible example is the U.S. military. The ratio of generals and admirals to the overall size of the U.S. military increased by 5 times since the World War II. Even if in 2010 the then-Defense Secretary Robert Gates noted the trend and called for cutting the flag officer positions, in 2013 there were 10 more three-star and 14 more two-star generals.

6.5 Size

Limited promotion opportunities affect how firms choose their sizes in a number of ways. First, Corollary 4 has shown that the size of the firm is inefficiently small in this model, but how firm adjusts its size in response to changes to turnover depend on the availability of opportunities. In particular, when there are enough promotion opportunities, an increase in the turnover rate lowers the firm size (part (i) of Corollary 4), and therefore, one more exiting worker means less-than-one new worker is hired to replace him. This is the familiar effect from Shapiro-Stiglitz, i.e., turnover reduces the size of the firm.

More importantly, when promotion opportunities are more limited (part (ii) of Corollary 4), firm size increases with the turnover rate. As a result, more than one new worker is hired to replace an exiting one. Conversely, one fewer exit at the top means that more than one potential hires no longer receive job offers. Notice that much of the hiring in our model reflects churning: new workers hired to replace exiting ones. Various studies (Abowd, Corbel, and Kramarz 1996, Lazear and Spletzer 2012, 2013) have shown that churning is an important part of total hiring and that

factors that affect churning can affect the aggregate employment level. These studies have focused on the importance of match-quality in affecting hiring. Our model shows that factors that affect churning can influence the employment level beyond the match-quality channel.

Second, when promotion opportunities are more limited, firm size can become less responsive to productivity shocks. Part (i) of Corollary 4 shows that there are ample promotion opportunities, the size of the firm increases at a rate of $1/c_1$ for increase in productivity $(d(N_1 + N_2)/d\alpha_1 = 1/c_1)$. When the promotion opportunities are more limited, however, the firm size becomes less responsive to changes in productivity in the bottom $(d(N_1 + N_2)/d\alpha_1 < 1/c_1)$. The reason is that even if the bottom workers become more productive, their promotion prospect remains limited. The concern for career opportunities therefore dampens the sensitivity of hiring to productivity.

This result sheds light on the use of headcount restrictions. Firms often set hard limits in each fiscal year on the number of permanent employees, known as the headcounts. Typically, the headcount is smaller than the number of workers necessary to carry out the work (Barley and Kunda 2011). Our model shows that part of the reason for the headcount is to make sure the new hires will have a career in the firm. Moreover, for firms with more limited promotion opportunities, their headcounts are less sensitive to changes in firm productivity.

7   Conclusion

This paper shows that career ladder arises naturally in response to imperfect contracting within organizations. Jobs with lower rents serve as ports-of-entry, and workers are motivated by the promotion opportunities to advance to jobs with higher rents. When promotion opportunities become limited, organizations optimally push out workers at the top to keep line of advancement open. Organizations also become more top heavy. Finally, organizations become less able to respond to outside changes.

The simplicity of the model allows one to extend it in a market setting, and therefore, allows one to study the effects of labor market policies on the organization of the internal labor markets. In a separate paper, we show that a progressive-tax policy that directly affects the top workers have indirect effects on bottom workers. Fewer workers are hired at the bottom although the existing workers are more likely to be promoted. We also show that by banning the firms from firing its workers, not only the wages of the top job becomes lower and the employment level goes down, but also the span of the hierarchy becomes less flat. Finally, we show that a minimum wage policy can both reduce and increase the employment level, depending on the parameter range.

In addition, there are a number of other future directions for this model. Currently, all workers

are homogenous. By incorporating ability heterogeneity into the model and allowing the ability to be learnt over time, we can use the model to address additional types human resources policies. For example, a direct application of the model is that once internal labor market is important, the firm may prefer promoting from within to hiring from outside, i.e. there can be an insider-bias. In addition, the model can potentially address the wage-seniority puzzle by Medoff and Abraham (1982): workers on a job longer receive higher wages (without performing better) because the chances for promotion for these workers are smaller, so to motivate them, a higher efficiency-wage is called for. Finally, since our model focuses on managing turnover for promotion purposes, allowing for heterogeneous workers can help shed light on human resources policies (such as up-or-out) in professional industries in which both the incentive and selection of workers are important.

Another direction for future research is to consider firms with multiple levels of hierarchy. This allows us to study not only the turnover policies at the top jobs but also the turnovers in the middle-rank. In addition, multiple levels of hierarchy enable us to make prediction about how wages, promotion probabilities, and spans change at different levels of hierarchy. To mechanism in our model suggests that wage can be convex in hierarchy levels: a larger wage increase at higher level provides incentives for more workers below, c.f. Rosen (1986). Studying the promotion probabilities and spans for multi-layer hierarchies may allow one to uncover further patterns on careers within organization and therefore better understand the effects of labor market policies on internal labor markets.

Third, the outside options of the workers are currently left as exogenous. In human-capital-intensive industries, firms often affect the outside options of their workers by engaging in training in general human capital. Such training is often considered a puzzle because they raise the outside options of the workers; see Becker (1975). In general, the conventional wisdom within economics suggests that a higher outside options of the workers hurt the firms, but in practice firms often take effort to boost the outside options of their workers. In particular, the consulting firms worry about whether their workers can find good employment elsewhere, and one reason appears to be related to creating promotion opportunities. A McKinsey insider comments that "if international companies stopped recruiting former McKinsey staff, it could clog the "up or out" refining process." In our context, if training increases the outside option of the top job more than that for the bottom job, it makes promotion more valuable and can help reduce distortions in the span of hierarchy and wages. This provides a justification for why firms would like to provide general training, which is one way to increase the worker's outside options. More importantly, extending the model to allow for training could shed light on how training policies interacts with the organizational structure of

the firm.

Fourth, our model has ruled out formal contracts as a way to motivate the workers. More generally, firms use a mix of formal contracts and career-based incentives to motivate their workers. For example, one way to increase the turnover rate at the top is to use of buy-out policies. By committing to a severance pay to the top workers, the firms increase the value of promotion, and this makes it easier to motivate the bottom workers. Of course, the effectiveness of these formal contracts depends on the qualities of legal institutions, and when the formal contracts are harder to enforce, the firms reply more on career-based incentives. This model therefore allows us to explore how the availability of formal contracts affects the organization of the internal labor markets and the careers of workers.

Finally, this model considers firms in steady states: the size of the firm and its hierarchy does not change over time. One future research direction is to study how the human resources policies vary for high-growth and more stable more firms. It appears natural that firms with higher growth rate can rely more on promotion incentives to motivate their workers. At some point, however, high-growth firms become more mature and their growth slows down. Understanding how firms will change their human resources policies when the promotion opportunities shrink is a fascinating theoretical question with important practical implications.

8 Appendix

8.1 Proof of Lemma 0.

Proof. For task i, the firm will choose the lowest wage such that the worker feels indifferent between working or shirking, i.e.,

$$w_i - c_i + (1 - d_i)\delta v_i = w_i + (1 - q_i)(1 - d_i)\delta v_i,$$

which gives

$$v_i^{ss} = \frac{c_i}{(1 - d_i)\delta q_i}.$$

Therefore, by the relationship between wage and continuition value, we have

$$w_i^{ss} = (1 - (1 - q_i)(1 - d_i)\delta)v_i^{ss} = \frac{(1 - (1 - q_i)(1 - d_i)\delta)}{(1 - d_i)\delta q_i}c_i.$$

Given w_i^{ss}, the optimal hiring N_i^{ss} is determined by the first order condition

$$\frac{\partial f(N_i, N_j)}{\partial N_i} = w_i^{ss},$$

and $H_i^{ss} = (1 - d_i)N_i^{ss}$ is chosen to satisfy the requirement of number of productive workers. ∎

8.2 Proof of Lemma 1.

Proof. For convenience, we introduce a notation

$$M_i \equiv (1 - d_i)N_i \ (i = 1, 2).$$

Using the promise-keeping constraint $(PK - 1)$ and $(PK - 2)$, the firm's labor cost can be rewritten as

$$
\begin{aligned}
W &= w_1 N_1 + w_2 N_2 \\
&= N_1(v_1 + c_1 - \delta(1 - d_1)(p_{11}v_1 + p_{12}v_2)) + N_2(v_2 + c_2 - \delta(1 - d_2)(p_{21}v_1 + p_{22}v_2)) \\
&= N_1 c_1 + N_2 c_2 + v_1(N_1 - \delta((1 - d_1)p_{11}N_1 + (1 - d_2)p_{21}N_2)) \\
&\quad + v_2(N_2 - \delta((1 - d_1)p_{12}N_1 + (1 - d_2)p_{22}N_2)) \\
&= N_1 c_1 + N_2 c_2 + v_1(N_1 - \delta(N_1 - H_1)) + v_2(N_2 - \delta(N_2 - H_2)) \\
&= N_1 c_1 + N_2 c_2 + v_1((1 - \delta)N_1 + \delta H_1) + v_2((1 - \delta)N_2 + \delta H_2),
\end{aligned}
$$

where the third step uses flow constraints (FL-1) and (FL-2).

Therefore, to minimize $w_1 N_1 + w_2 N_2$, is equivalent to minimize

$$v_1 \left((1-\delta)N_1 + \delta H_1 \right) + v_2 \left((1-\delta)N_2 + \delta H_2 \right).$$

Now we show $H_2^* = 0$. We first assume $v_2^* \geq v_1^*$, which we will later verify. Further, we will first consider the case where $M_1 + M_2 > N_2$ so the top jobs are scarce.

Consider an optimal W^* with $H_2^* > 0$ and $v_2^* \geq v_1^*$. Since $M_1 + M_2 > N_2$, either $p_{12}^* < 1$ or $p_{22}^* < 1$. In the first case, let $\tilde{H}_1 = H_1^* + M_1\varepsilon$, $\tilde{H}_2 = H_2^* - M_1\varepsilon$, $\tilde{p}_{11} = p_{11}^* - \varepsilon$, and $\tilde{p}_{12} = p_{12}^* + \varepsilon$. In the second case, let $\tilde{H}_1 = H_1^* + M_2\varepsilon$, $\tilde{H}_2 = H_2^* - M_2\varepsilon$, $\tilde{p}_{21} = p_{21}^* - \varepsilon$, and $\tilde{p}_{22} = p_{22}^* + \varepsilon$. Let $\tilde{W}_j$ denote the wage bill under perturbation j. Then

$$\tilde{W}_j = W - \delta M_j \varepsilon \left(v_2^* - v_1^* \right) \leq W^*.$$

If $v_2^* > v_1^*$ is strict, the above inequality show a contradictions of the optimality of original W^*. If $v_2^* = v_1^*$, then the above perturbations do not increase the cost, we can do so until $\tilde{H}_2 = 0$. Therefore, $H_2^* = 0$. ∎

8.3 Proof of Lemma 2.

Proof. We first show $v_2^* = R_2$ and $v_1^* = 0$. By (IC-2), it is easy to see

$$v_2 \geq R_2.$$

Note $v_1 \geq 0$ (by IR-1). Therefore, if $(v_1, v_2) = (0, R_2)$ is attainable, it will minimize the labor cost. Since p_{ij} does not enter the cost function directly, it suffices to show the existence of some assignment probability P such that $(v_1, v_2) = (0, R_2)$ satisfies all constraints. That is indeed the case, by sending $p_{22}^* = 1$ and

$$p_{12}^* = \frac{N_2 - M_2}{M_1} \leq 1,$$

so that (IC-1)

$$p_{12}^* R_2 \geq R_1$$

is satisfied given $d_2 N_2 R_2 \geq (1 - d_1) N_1 R_1$. Therefore, we conclude $(v_1^*, v_2^*) = (0, R_2)$ and $v_1^* < v_2^*$ is confirmed. Clearly, $p_{22}^* = 1$ implies no demotion ($p_{21}^* = 0$) and full job security for the top job.

Now we show the full job security for the bottom job. Given $H_2^* = 0$ and $p_{22}^* = 1$ as we have shown, we add two flow constraints (FL-1) and (FL-2) to obtain

$$(p_{11}^* + p_{12}^*)M_1 + M_2 + H_1^* = N_1 + N_2,$$

which implies

$$H_1^* \geq (1 - p_{11}^* - p_{12}^*)M_1 + N_1 - M_1.$$

Suppose that by contradiction $p_{11}^* + p_{12}^* < 1$. We let $\tilde{H}_1 = H_1^* - M_1\varepsilon$ and $\tilde{p}_{11} = p_{11}^* + \varepsilon$ for some $\varepsilon > 0$. All other choice variables are kept the same. So we still have the flow constraint

$$\tilde{p}_{11}M_1 + p_{21}^* M_2 + \tilde{H}_1 = N_1,$$

and the increasing of p_{11}^* will not destroy (IC-1). Then all other constraints are satisfied. Under the above perturbation, the labor cost is weakly decreased. So we can continue to do perturbation until $1 = p_{11}^* + p_{12}^*$, where $\tilde{H}_1 > 0$ is still true. Therefore, at the optimum, $1 = p_{11}^* + p_{12}^*$, i.e., the full job security. ■

8.4   Proof of Lemma 3

Proof.   Using nations Δ_i and multiplying it by M_i, we have

$$
\begin{aligned}
M_1\Delta_1 &= M_1p_{11}v_1 + M_1p_{12}v_2 - M_1R_1 \\
M_2\Delta_2 &= M_2p_{21}v_1 + M_2p_{22}v_2 - M_2R_2.
\end{aligned}
$$

Add the above two equalities up, we obtain

$$
\begin{aligned}
M_1\Delta_1 + M_2\Delta_2 &= (M_1p_{11} + M_2p_{21})v_1 + (M_1p_{12} + M_2p_{22})v_2 - M_1R_1 - M_2R_2 \\
&= (N_1 - H_1)v_1 + (N_2 - H_2)v_2 - M_1R_1 - M_2R_2
\end{aligned}
$$

where the last step uses flow constraints (FL-1) and (FL-2). Therefore, we can rewrite the objective function as

$$
\begin{aligned}
W &= N_1c_1 + N_2c_2 + v_1\left((1-\delta)N_1 + \delta H_1\right) + v_2\left((1-\delta)N_2 + \delta H_2\right) \\
&= N_1c_1 + N_2c_2 + H_1v_1 + H_2v_2 + (1-\delta)\left[(N_1 - H_1)v_1 + (N_2 - H_2)v_2\right] \\
&= N_1c_1 + N_2c_2 + H_1v_1 + H_2v_2 + (1-\delta)(M_1\Delta_1 + M_2\Delta_2) + (1-\delta)(M_1R_1 + M_2R_2)
\end{aligned}
$$

As we have shown that at the optimum $H_2^* = 0$, and $H_1^* = d_1N_1 + d_2N_2$ is independent of v_i and p_{ij}. Therefore, if $v_1 = 0$ and $\Delta_i = 0$ $(i = 1, 2)$ is attainable, the labor cost will be minimized. When $d_2N_2R_2 < (1 - d_1)N_1R_1$, we can confirm that $v_1 = 0$ and $\Delta_i = 0$ satisfy all the constraint as follows. From $\Delta_i = 0$, and use flow constraints (FL-1) and (FL-2), we can solve that

$$v_2^* = \frac{R_1M_1 + R_2M_2}{N_2} > R_2 > 0.$$

The corresponding assignment probabilities are feasible by noting that

$$p_{12}^* = \frac{R_1 N_2}{R_1 M_1 + R_2 M_2} \in (0, 1),$$

$$p_{22}^* = \frac{R_2 N_2}{R_1 M_1 + R_2 M_2} \in (0, 1).$$

Therefore, the optimal solution is

$$v_1^* = 0, v_2^* = \frac{R_1 M_1 + R_2 M_2}{N_2}.$$

Now we show no demotion, i.e., $p_{21}^* = 0$. Since $v_1^* = 0$, for any $p_{21}^* > 0$, we can decrease p_{21}^* by ε. Let $\tilde{p}_{21} = p_{21}^* - \varepsilon$ and $\tilde{H}_1 = H_1^* + M_2 \varepsilon$ for some $\varepsilon > 0$. All other choice variables are kept the same. So we still have the flow constraint

$$p_{11}^* M_1 + \tilde{p}_{21} M_2 + \tilde{H}_1 = N_1,$$

and the decreasing of p_{21}^* will not destroy (IC-1) given $v_1^* = 0$. We can do this pertubation until $\tilde{p}_{21} = 0$. It becomes clear that $p_{21}^* + p_{22}^* = \frac{R_2 N_2}{R_1 M_1 + R_2 M_2} < 1$, which implies that the top job is not fully secure.

Finally, the bottom job is still fully secure by the same logic that we argue in the proof of Lemma 2. ∎

8.5 Proof of Corollary 1

Proof. (i) Recall $(v_1^*, v_2^*) = (0, R_2)$ (by Lemma 2). Then, based on promise-keeping constraints (PK-1) and (PK-2),

$$w_1 = c_1 - \delta(1 - d_1)p_{12}^* R_2 = c_1 - \frac{d_2 c_2 N_2}{N_1(1 - d_2)}$$

and

$$w_2 = R_2 + c_2 - \delta(1 - d_2)R_2 = \frac{c_2}{(1 - d_2)\delta}.$$

Therefore, plugging the above two formulas into the labor cost $w_1 N_1 + w_2 N_2$, we obtain the desired labor cost function.

(ii) Recall $(v_1^*, v_2^*) = (0, \frac{R_1 M_1 + R_2 M_2}{N_2})$ (by Lemma 3). Then, based on promise-keeping constraints (PK-1) and (PK-2),

$$w_1 = c_1 - \delta(1 - d_1)R_1 = 0$$

and

$$w_2 = v_2 + c_2 - \delta(1 - d_2)R_2 = \frac{R_1 M_1 + R_2 M_2}{N_2} = \frac{c_1 N_1 + c_2 N_2}{\delta N_2}.$$

Therefore, plugging the above two formula into the labor cost $w_1 N_1 + w_2 N_2$, we obtain the desired labor cost function. ∎

8.6 Proof of Corollary 2

Proof. We can solve the optimal N_1 by the first order condition

$$\frac{\partial f(N_1, N_2)}{\partial N_1} = \frac{dW(N_1, N_2)}{dN_1}.$$

Since $\kappa = \frac{d_2 c_2}{(1-d_2)c_1}$ is increasing in d_2, the cut-off $\bar{d}_2$ is

$$\bar{d}_2 = \frac{N_1^*(N_2)c_1}{N_1^*(N_2)c_1 + N_2 c_2},$$

where $N_1^*(N_2)$ satisfies $\frac{\partial f(N_1^*(N_2), N_2)}{\partial N_1} = c_1$. For $d_2 \geq \bar{d}_2$, we have $\frac{N_1^*(N_2)}{N_2} < \kappa$, then the optimal $N_1^*(N_2)$ is determined by $\frac{\partial f(N_1^*(N_2), N_2)}{\partial N_1} = c_1$. Similarly, cut-off $\underline{d}_2$ is determined by the similar manner, replacing $N_1^*(N_2)$ with the one that satisfies the first order condition $\frac{\partial f(N_1^*(N_2), N_2)}{\partial N_1} = \frac{c_1}{\delta}$. Since $f(., N_2)$ is concave, so $\underline{d}_2 < \bar{d}_2$. When $d_2 \in [\underline{d}_2, \bar{d}_2)$,

$$c_1 < \frac{\partial f(N_1, N_2)}{\partial N_1} \leq \frac{c_1}{\delta}$$

for $N_1 = \kappa N_2$. Therefore, the optimal solution is $N_1^*(N_2) = \kappa N_2$. We complete these three cases. ■

8.7 Proof of Corollary 3

Proof. (i) If at the optimum, $N_1^* < \kappa N_2^*$, then according to the labor cost function defined in Part (i) of of Corollary 1, we obtain the first order condition for the optimality as desired. (ii) If $N_1^* = \kappa N_2^*$, N_1 is the decision variable, and the first order condition w.r.t. N_1 gives the desired equation. (iii) If at the optimum, $N_1^* > \kappa N_2^*$, then according the labor cost function defined in Part (ii) of Corollary 1, we obtain the desired first order condition. ■

8.8 Proof of Corollary 4

Proof. (i) According to Corollary 3, we can calculate $N_1^* = \alpha_1/c_1$ and $N_2^* = (1 - d_2)\alpha_2\delta/(c_2(1 - \delta d_2))$. And the cut-off d_2 is solved by

$$\frac{(1 - \delta d_2)\alpha_1}{(1 - d_2)\alpha_2\delta} = \frac{N_1^* c_1}{N_2^* c_2} = \frac{d_2}{1 - d_2},$$

which is $\frac{1}{\delta}\frac{\alpha_1}{\alpha_1 + \alpha_2}$. (ii) The first order condition in Part (ii) of Corollary 3 implies $N_1^* = d_2(\alpha_1 + \alpha_2)\delta/c_1$ and $N_2^* = (1 - d_2)(\alpha_1 + \alpha_2)\delta/c_2$. The cut-off d_2 is determined by (i) and (ii). (iii) We can calculate $N_1^* = \delta\alpha_1/c_1$ and $N_2^* = \alpha_2\delta/c_2$. And the cut-off d_2 is solved by

$$\frac{\delta\alpha_1}{\alpha_2\delta} = \frac{N_1^* c_1}{N_2^* c_2} = \frac{d_2}{1 - d_2},$$

which is $\frac{\alpha_1}{\alpha_1 + \alpha_2}$. ■

References

[1] Abraham, Katharine G., and James L. Medoff. "Length of service and the operation of internal labor markets." MIT, Sloan School of Management, *Working Paper*, 1394-83, (1983).

[2] Abowd, J. M. , Patrick Corbel, and Francis Kramarz, "The Entry and Exit of Workers and the Growth of Employment: an Analysis of French Establishments", *Review of Economics and Statistics*, May 1999, 81(2): 170–187.

[3] Akerlof, George A., and Janet L. Yellen, eds. *Efficiency wage models of the labor market.* Cambridge University Press, 1986.

[4] Akerlof, George and Lawrence Katz (1989) "Workers' Trust Funds and the Logic of Wage Profiles," *Quarterly Journal of Economics*, 104 (3) pp. 525-536.

[5] Axelson, Ulf and Phillip Bond. "Wall Street occupations." (2014).

[6] Baker, George P., Michael C. Jensen, and Kevin J. Murphy. "Compensation and incentives: Practice vs. theory." *The journal of Finance* 43.3 (1988): 593-616.

[7] Baker, George, Michael Gibbs, and Bengt Holmstrom. "The internal economics of the firm: evidence from personnel data." *The Quarterly Journal of Economics* 109.4 (1994): 881-919.

[8] Barley, S. R. and Gideon Kunda, *Gurus, Hired Guns, and Warm Bodies: Itinerant Experts in a Knowledge Economy,* Princeton University Press, (2011).

[9] Beaudry, P., and R. DiNardo, "The Effect of Implicit Contracts on the Movement of Wages over the Business Cycle: Evidence from Micro Data," *Journal of Political Economy*, 99 (1991), 665-688.

[10] Becker, G., *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, 2nd Edition.* Chicago and London: University of Chicago Press, (1975).

[11] Bernhardt, D. "Strategic Promotion and Compensation," *Review of Economic Studies* 62 (1995), 315-339.

[12] Board, S. "Relational Contracts and the Value of Loyalty," *American Economic Review* 101(7) (2011), pp. 3349–67.

[13] Board, Simon, and Moritz Meyer-ter-Vehn. "Relational Contracts in Competitive Labor Markets." (2013).

[14] Calvo, Guillermo A., and Stanislaw Wellisz. "Supervision, loss of control, and the optimum size of the firm." *Journal of Political Economy* 86.5 (1978): 943-52.

[15] Câmara, Odilon, and Dan Bernhardt. "The dynamics of promotions, quits and layoffs." (2009).

[16] Cappelli, P., Talent on Demand. Harvard Business Press, (2008).

[17] Carmichael, L. "Can Unemployment Be Involuntary?" *The American Economic Review* (1985): 75, pp. 1213-1214.

[18] Chiappori, A, Salanie B, and J. Valentin "Early Starters Versus Late Beginners," *Journal of Political Economy*, 107 (1999), 731-760.

[19] Demougin, Dominique, and Aloysius Siow. "Careers in ongoing hierarchies." *The American Economic Review* (1994): 1261-1277.

[20] DeVaro, Jed, and Hodaka Morita. "Internal promotion and external recruitment: a theoretical and empirical analysis." *Journal of Labor Economics* 31.2 (2013): 227-269.

[21] Dickens William, Kevin Lang, Larry Katz, and Larry Summers (1990), "Why Do Firms Monitor Workers?" in Yoram Weiss and Gideon Fishelson, eds., Advances in the Theory and Measurement of Unemployment (London: MacMillan), pp. 159-71.

[22] Doeringer, Peter and Michael Piore (1971). *Internal Labor Markets and Manpower Analysis*. Lexington, MA: Heath Lexington Books.

[23] Dohmen, T.J., B. Kriechel, and G.A. Pfann, "Monkey Bars and Ladders: The Importance of Lateral and Vertical Job Mobility in Internal Labor Market Careers," *Journal of Population Economics*, 17 (2004), pp. 193-228.

[24] Fuchs, William (2007), "Contracting with Repeated Moral Hazard and Private Evaluations", *American Economic Review*, 97(4); pp. 1432-1448.

[25] Garicano, L. and T. Hubbard, "Specialization, Firms, and Markets: The Division of Labor within and between Law Firms", *Journal of Law, Economics, and Organizations*, 25 (2008), pp. 339-71.

[26] Gibbons, Robert. Incentives and careers in organizations. No. w5705. National bureau of economic research, 1997.

[27] Gibbons, R., and M. Waldman, "A Theory of Wage and Promotion Dynamics inside Firms," *Quarterly Journal of Economics*, 114 (1999), 1321-58.

[28] Gibbons, R., and M. Waldman, "Careers in Organizations: Theory and Evidence," Chapter 36 in Volume 3B of O. Ashenfelter and D. Card (eds.), Handbook of Labor Economics, North Holland, (1999).

[29] Grund, C., The Wage Policy of Firms—Comparative Evidence for the U.S. and Germany from Personnel Data, *Discussion Paper* No. 605, University of Bonn and IZA Bonn (2005).

[30] Guadalupe, Maria, and Julie Wulf, "The Flattening Firm and Product Market Competition: The Effect of Trade Liberalization on Corporate Hierarchies." *American Economic Journal: Applied Economics (*2010), 2(4): 105–127.

[31] Hall, Robert (1982), "The Importance of Lifetime Jobs in the U.S. Economy," *American Economic Review*, 72 (4), pp. 716-724.

[32] Harris, M., and B. Holmstrom, "A Theory of Wage Dynamics," *Review of Economic Studies* 72 (1982) 315-333.

[33] Imai, Ken'ichi, and Itami Hiroyuki, Allocation of Labor and Capital in Japan and the United States, in *Inside the Japanese System*, Okimoto, Daniel and T. Rohlen, eds., pp 112-7, Stanford University Press, 1988.

[34] Kaplan, Steven N., and Bernadette A. Minton, "How has CEO Turnover Changed? Increasingly Performance Sensitive Boards and Increasingly Uneasy CEOs," *NBER Working Paper* No. W12465, (2008).

[35] Krakel, Matthias, and Anja Schottner. "Internal Labor Markets and Worker Rents." *Journal of Economic Behavior and Organization* 84 (2012), pp. 491-509.

[36] Lazear, Edward P. "Why is there mandatory retirement?." *The Journal of Political Economy* (1979): 1261-1284.

[37] Lazear, Edward P. *Personnel economics: past lessons and future directions*. No. w6957. National bureau of economic research, 1999.

[38] Lazear, Edward P., and Paul Oyer. "Personnel Economics." *The Handbook of Organizational Economics* (2012): 479.

[39] Lazear, Edward P., and Sherwin Rosen. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89 (1981), 841—64.

[40] Leonard, J, "Executive Pay and Firm Performance", *Industrial and Labor Relations Review* (1990), 43(3), 13–29

[41] Lin, M.J., "Opening the Black Box: The Internal Labor Markets of Company X," *Industrial Relations*, 44 (2005), pp. 659-706.

[42] MacDonald, G., "A Market Equilibrium theory of Job Assignment and Sequential Accumulation of Information," *American Economic Review*, 72 (1982) 1038-1055.

[43] MacLeod, B., and J. Malcomson, "Reputation and Hierarchy in Dynamic Models of Employment," *Journal of Political Economy*, 96 (1988) 832-854.

[44] Merton, R. *A Life of Learning*, New York: American Council of Learned Societies (1994).

[45] Mincer, J., Schooling, Experiences, and Earnings, New York: Columbia University for National Bureau of Economic Research, (1974).

[46] Mookherjee, D., 'Incentives in Hierarchies', in *Handbook of Organizational Economics*, edited by Robert Gibbons and John Roberts, eds. Princeton University Press, 2013.

[47] Morris, S., "Stalled Professionalism: The Recruitment of Railway Officials in the United States, 1885-1940," *The Business History Review*, 47 (1973) 317-334

[48] Nalebuff, Barry J., and Joseph E. Stiglitz. "Prizes and incentives: towards a general theory of compensation and competition." *The Bell Journal of Economics* (1983): 21-43.

[49] Neal, D., and S. Rosen., "Theories of the Distribution of Earnings," in Handbook of Income Distributions, Vol. 1, North Holland (2000).

[50] Qian, Yingyi. "Incentives and loss of control in an optimal hierarchy." *The Review of Economic Studies* 61.3 (1994): 527-544.

[51] Sattinger, M., "Assignment Models of the Distribution of Earnings," *Journal of Economic Literature*, 31 (1993), 831-80.

[52] Rajan, R. and J. Wulf, "The Flattening Firm: Evidence on the Changing Nature of Corporate Hierarchies, Review of Economics and Statistics, 88. 4, (2006), pp. 759-73.

[53] Rebitzer, J. B. and Lowell J. Taylor, Efficiency Wages and Employment Rents: The Employer-Size Wage Effect in the Job Market for Lawyers, *Journal of Labor Economics*, 13 (1995), pp. 678-708.

[54] Seltzer, Andre, and David Merrett, "Personnel Policies at the Union Bank of Australia: Evidence form the 1888-1900 Entry Cohorts", *Journal of Labor Economics,* 18 (2000), pp. 573-613.

[55] Shapiro, Carl, and Joseph E. Stiglitz. "Equilibrium unemployment as a worker discipline device." *The American Economic Review* (1984): 433-444.

[56] Treble, J., E.V. Cameren, S. Bridges, and T. Barmby, "The Internal Economics of the Firm: Further Evidence from Personnel Data," *Journal of Labor Economics*, 8 (2001), pp. 531-   552.

[57] Waldman, Michael. "Job Assignments, Signalling, and Efficiency " *RAND Journal of Economics,* Vol. 15, No. 2 (Summer, 1984), pp. 255-267.

[58] Waldman, Michael. "Classic promotion tournaments versus market-based tournaments." *International Journal of Industrial Organization* 31.3 (2013): 198-210.

[59] Williamson, Oliver E. *The economics of discretionary behavior: Managerial objectives in a theory of the firm.* Chicago, Illinois: Markham Publishing Company, 1967.

[60] Vogel, Ezra F. *Deng Xiaoping and the transformation of China.* Harvard University Press, 2011.

[61] Zabojnik, Jan, and Dan Bernhardt. "Corporate Tournaments, Human Capital Acquisition, and the Firm Size—Wage Relation." *The Review of Economic Studies* 68.3 (2001): 693-716.