Managing Careers in Organizations*

Rongzhu Ke Jin Li Michael Powell

September 27, 2014

Abstract

This paper develops a framework for examining how a firm chooses its personnel policies in order to manage promotion opportunities for its employees. Managing employees' careers involves a trade-off between incentive provision at the individual-employee level with productive efficiency at the firm level, so careers are optimally managed at the firm-level. An employee's career therefore depends on his employer's characteristics. Further, employees' careers are optimally interdependent, so the firm may adopt forced-turnover policies for higher-level employees in order to keep the lines of advancement open. Our model is consistent with a host of stylized facts and suggests a number of further empirical implications.

.

*Rongzhu Ke: Department of Economics, Chinese University of Hong Kong. E-mail: rzke@cuhk.edu.hk. Jin Li: Department of Management and Strategy, Kellogg School of Management, Northwestern University. E-mail: jin-li@kellogg.northwestern.edu. Michael Powell: Department of Management and Strategy, Kellogg School of Management, Northwestern University. E-mail: mike-powell@kellogg.northwestern.edu. We thank Chen Cheng and Can Urgun for excellent research assistance. We are grateful to Daron Acemoglu, Ricardo Alonso, Pierre Azoulay, Dan Barron, Tim Bond, Yuk-fai Fong, Bob Gibbons, Bengt Holmstrom, Ben Jones, Kevin Lang, Danielle Li, Niko Matouschek, David Matsa, Giuseppe Moscarini, Arijit Mukherjee, Nicola Persico, Jim Rebitzer, Xianwen Shi, Aloysius Siow,Wing Suen, Mike Waldman,Yanhui Wu, Jano Zabojnik, and seminar participants at CEPR, Fudan, HKU, HKUST, MIT, NBER, Purdue, Queens, Regensburg, SOLE, UCLA, USC Marshall, and UToronto (CEPA Brown Bag) for helpful conversations and suggestions.

1 Introduction

Firms attract, retain, and motivate workers by promising them careers. But delivering on promises to promote workers requires that there are positions to promote them to. Indeed, as Peter Cappelli (2008) concludes, "Frustration with advancement opportunities is among the most important factors pushing individuals to leave for jobs elsewhere." On the other hand, promoting too many workers can lead to firms becoming "top-heavy" in the sense of having unnecessarily many high-level employees relative to the firm's objectives. Managing promotion opportunities is therefore a delicate problem that is important for real firms. And it is one that has been beyond the scope of classic models of internal labor markets, which either treat workers' careers independently (Waldman 1984; Gibbons and Waldman 1999) or take promotion opportunities as given (Lazear and Rosen, 1982).

In this paper, we develop a parsimonious framework in which managing promotion opportunities in firms involves trading off incentive provision at the individual-worker level with productive efficiency at the firm level. At the individual-worker level, contracting imperfections limit transfers and therefore require the firm to provide the worker with **incentive rents**, which are optimally backloaded as in dynamic moral-hazard models (e.g., Lazear 1979, Board 2011). If production requires multiple activities, then activities that require more incentive rents are optimally performed by workers farther along in their careers: promotions arise naturally as a way of reusing incentive rents. But the firm's returns to having a worker perform a given activity is determined by its contribution to production. A tension thus arises between using a promotion to provide incentives for a given worker and using activity assignment for productive efficiency. In turn, conflict arises between providing incentives for one worker and providing incentives for the firm's other workers. Career paths are therefore optimally managed at the firm-level rather than at the individual-worker level.

Formally, we propose a model that builds upon Shapiro and Stiglitz (1984)'s efficiency-wage model by allowing for multiple activities within a single firm. Homogeneous workers privately choose whether to work or shirk, and the firm can motivate workers by committing to a wage that is tied to the activity, coupled with the threat of firing workers who are caught shirking. Each firm has two types of activities that have to be performed, and each worker can perform a single type of activity in each period. The two activities differ in the level of incentive rents that are required to provide motivation, because one activity (the **high-rent activity**) is either more onerous or more difficult to monitor than the other activity (the **low-rent activity**). The firm's output and therefore its revenues depends on the number of workers performing each type of activity, 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 activity. In addition, the firm chooses a bundle of **personnel policies**. How many workers should the firm hire into each activity each period? Should the firm retain its incumbent workers? If so, what activity should they perform next period? What wage should be associated with each activity? The firm's personnel policies are limited by two key constraints. Workers have to be motivated to exert effort in each activity. 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.

Optimal personnel policies resemble an internal labor market. The low-rent activity is performed in the **bottom job**, which serves as a **port of entry** (Doeringer and Piore, 1971). Workers remain in the bottom job until they are promoted to the **top job**, which requires performing the high-rent activity. Once in the top job, **workers are never demoted** (Baker, Gibbs, and Holmstrom QJE 1994; hereafter BGH). As a result, a well-defined career path emerges, and it plays the role of workers' trust funds (Akerlof and Katz, QJE 1989). **Workers in the bottom job receive zero rents** and therefore effectively post a bond by beginning employment in the bottom job. 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 activity. A worker's wages therefore increase upon promotion (BGH).

When a worker departs from the firm, his position can be reallocated to another worker. Worker turnover therefore can expand promotion opportunities, providing a reason for why the firm might want to put in place **forced-turnover policies** such as mandatory-retirement programs. If the promotion prospects created solely from voluntary turnover at the top are insufficient for motivating workers at the bottom, the firm optimally forces a fraction of the workers at the top to leave the firm in every period. Viewed in isolation, adopting forced-turnover policies is a bad idea, since doing so reduces the expected rents of workers at the top, which would violate their incentive-compatibility constraint. However, forced-turnover policies are optimally complemented with more generous compensation for workers at the top as well as a more generous promotion policy for workers at the bottom. A recurring theme in the analysis is that personnel policies are interdependent.

The firm may also expand promotion opportunities by altering its hierarchical structure away from what would be productively efficient. Creating an additional position at the top of the firm expands the opportunities available for those at the bottom of the firm and therefore confers a benefit to the firm in addition to marginal revenue. In contrast, creating an additional position

at the bottom of the firm reduces the career prospects of those at the bottom and therefore the benefit of doing so is less than the marginal revenue the position creates. For both of these reasons, the firm's **hierarchical span**–the ratio of the number of positions at the bottom to the number of positions at the top–is optimally lower than would be productively efficient for the wages it pays.

Since the firm's personnel policies and its production decisions are determined optimally, our model serves as a framework for examining how firms of different size differ in how they manage workers' careers. Larger firms empirically tend to have larger spans. Larger spans call for different personnel policies, which optimally includes higher wages, lower promotion rates, higher forced-turnover rates, and stronger insider bias in hiring at the top. These predictions go beyond the classic evidence on the static firm size-wage effect, and they describe how workers' entire careers differ based on the size of their employers in ways that are consistent with several recent empirical findings. The upshot is that two identical workers may have different career experiences depending on the firm-level characteristics of their employers.

Workers' careers are optimally interlinked, and therefore a firm's demographics affects the careers of all its workers. As we mentioned above, firms may optimally create promotion opportunities for younger workers by putting in place forced-turnover policies that are targeted at older workers. At the aggregate level, many countries have expanded the generosity of government retirement programs in order to encourage turnover of older workers and create opportunities for younger workers, but such policies have often had exactly the opposite effects (see Gruber and Wise, 2011, for many studies documenting these results). Our model predicts that such government policies may indeed negatively affect younger workers' employment prospects, precisely because firms optimally alter their personnel policies in response. We therefore provide an organizational explanation for the findings documented by the studies contained in Gruber and Wise (2011).

Literature Review This paper contributes to the literature on internal labor markets (see Gibbons (1997), Gibbons and Waldman (1999), Lazear (1999), Lazear and Oyer (2013), and Waldman (2013) for reviews of the theory on and evidence for internal labor markets). Relative to this literature, our model focuses on how the organization of internal labor markets provides incentives through task assignment; Lazear and Rosen (1981), MacLeod and Malcomson (1998), Zabojnik and Bernhardt (2001), Camara and Bernhardt (2009), and Krakel and Schottner (2012). In contrast to the existing literature, factors of production in our model are flexible but are subject to diminishing returns. This flexibility allows us to independently vary the firm's span and its size, in order to separately identify their effects. We show that a firm's span is related both to the wage dynamics of workers and to the adoption of various human-resource practices such as insider-bias in hiring

and forced-turnover policies.

We also contribute to the vast literature on dynamic moral hazard; see Bolton and Dewatripont (2005, chapter 10) for a textbook treatment. As in efficiency-wage models (Shapiro and Stiglitz, 1984), our model assumes that wages are tied to jobs, giving rise to incentive rents. The efficiency-wage literature has studied the question of how firms can extract these incentive rents from workers by backloading pay within a given job (Lazear, 1979; Carmichael, 1985; Akerlof and Katz, 1989; Board 2011; Fong and Li, 2013). In our model, backloading pay occurs across activity assignments, and our setting is indeed one in which the firm is able to extract all the surplus from workers. More importantly, we show that how a worker's pay is optimally backloaded (i.e., how his career progresses) is not determined in isolation. Rather, how a workers' pay is optimally backloaded depends on the firm's production technology and the careers of his coworkers. Our model therefore highlights how firm-level factors such as its production technology and its organizational demographics affect the dynamic moral-hazard problem at the individual-worker level.

There is a sizeable literature looking at how the need to provide incentives interacts with organizational design (Williamson, 1967; Calvo and Wellisz, 1978; Qian, 1994; Mookherjee, 2013). In these models, workers remain in a fixed position within the firm, and the firm's monitoring technology is the key driver of its organizational structure. In our model, workers' positions within the hierarchy are not fixed, and their promotion opportunities determine their incentives. The need to provide incentives, therefore, affects the firm's optimal organizational structure.

Finally, there is a literature outside economics that examines how the careers of individual workers progress within organizations; Simon (1951), White (1970), Bartholomew (1973), Keyfitz (1973), Stewman and Konda (1983), Rosenbaum (1984), and Stewman (1986). Similar to this paper, the span of the organization plays a key role in determining the speed of career advancement. Unlike this paper, this literature takes the span as fixed and does not consider how organizations can adjust their hierarchy to facilitate incentive provision. Our model shows that endogenizing the hierarchical structure can reverse the predictions from this literature. In particular, we provide a rationale for why programs that encourage older workers to retire do not always facilitate employment for younger workers.

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 have a common discount factor $\delta \in (0, 1)$. We focus on the steady state and suppress time subscripts. Production requires two types of activities to be performed, and each

worker can perform a single activity each period. A worker performing activity 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 is his private information, but shirking in activity i is contemporaneously detected with probability q_i. If the firm employs masses N_1 and N_2 of productive workers in the two activities, revenues are $F(N_1, N_2)$. F is differentiable, increasing, concave, and satisfies $F_{12} \geq 0$.

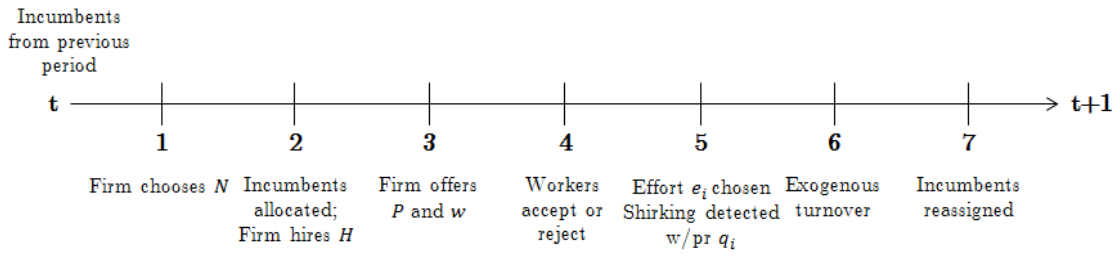

Figure 1: Timing of the stage game.

Figure 1 illustrates the timing of each period. The firm chooses the masses of positions N_1 and N_2 for each activity. The firm then fills these positions with incumbent workers and new hires, where we denote the mass of new hires into activity i as H_i, $i = 1, 2$. The firm offers each worker a contract (w_i, p_{ij}), $i, j = 1, 2$, that includes a **wage policy** and an **assignment policy** consisting of expected promotion, demotion, and retention patterns. We assume that wages are tied to activities, and denote the wage for activity i by w_i. The assignment policy is described by p_{ij}, which denotes the the probability that a worker in activity i will take on activity j next period if he is not caught shirking. We assume that a worker who is caught shirking is fired with probability 1, which constitutes an optimal penal code since it occurs only off the equilibrium path.

If a worker rejects the contract, he receives his outside option, yielding 0 utility. If he accepts the offer, the wage is paid and he 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 activity i exogenously leave the firm. We refer to d_i as the **voluntary departure rate** of workers in activity i. Incumbent workers are reassigned according to the probability matrix p_{ij}. If $p_{i1} + p_{i2} < 1$, some workers are asked to leave the firm and receive their outside utility. We refer to $1 - p_{i1} - p_{i2}$ as the **forced-turnover rate** for activity i.

3 Parallel-Careers Benchmark

To provide a benchmark against which to compare our results and to develop some useful notation and terminology, we begin by describing what we will refer to as the **parallel-careers benchmark**.

In this benchmark, the firm treats the two activities 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 worker mobility across activities.

Given a mass $\hat{N}_j$ of workers in activity j, the firm chooses N_i and w_i to solve the 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. A worker's motivation to work therefore derives from his expected future payoffs within the firm. Define the **incentive rents for activity** i as the minimum future payoffs necessary to satisfy the worker's incentive-compatibility constraint in activity i, and denote this value by R_i. The incentive-compatibility constraint can be rearranged to verify that

$$R_i = \frac{c_i}{(1 - d_i)\,\delta q_i} \qquad \text{(Incentive Rent)}$$

To maximize its profits, the firm chooses wages w_i, or equivalently, payoffs v_i, to ensure the incentive-compatibility constraint holds with equality. Given the resulting wage, the firm hires workers until the marginal revenue product of an additional worker is equal to this wage. Finally, the firm hires a mass of new workers into each activity to exactly offset the mass of workers who are exogenously separated from that activity. The resulting solution, which we refer to as the parallel-careers solution and denote with the superscript pc, is described in the following lemma.

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

Lemma 0 is consistent with several observations of Shapiro and Stiglitz (1984). First, the firm has to pay wages that exceed workers' outside options in order to provide incentives. The resulting "efficiency wage" is increasing in the departure rate d_i and decreasing in the firm's monitoring ability, q_i. Second, the firm optimally chooses an employment level for each activity that is lower than the socially optimal level, which would satisfy $F_i = c_i$. Moreover, the gap between the firm's employment-level choice and the socially optimal level is greater for activities that require higher incentive rents. We assume throughout that $R_2 > R_1$, so that in the parallel-careers benchmark, more incentive rents are required in activity 2 than in activity 1.

4 Managing Careers

In the parallel-careers benchmark, the firm chooses only a mass of workers to perform each activity and a wage paid to each of these workers. In this section, we study more general personnel policies that allow for reassignment across activities. We show that the firm always performs better by linking the activities together in the form of a career. We then 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

$$F(N_1, N_2) - w_1 N_1 - w_2 N_2.$$

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

Promise-Keeping Constraints. Productive workers' payoffs have to satisfy:

$$v_1 = w_1 - c_1 + (1 - d_1)\,\delta\,(p_{11} v_1 + p_{12} v_2)\,; \tag{PK-1}$$

$$v_2 = w_2 - c_2 + (1 - d_2)\,\delta\,(p_{21} v_1 + p_{22} v_2)\,. \tag{PK-2}$$

Individual-Rationality Constraints. Workers prefer working for the firm rather than taking their outside options if:

$$v_1 \geq 0; \tag{IR-1}$$

$$v_2 \geq 0. \tag{IR-2}$$

Incentive-Compatibility Constraints. Workers prefer to exert effort if the following conditions hold:

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

where we use the fact that if the worker leaves the firm, he receives a payoff of 0. Equivalently, future payoffs have to exceed activity i's incentive rents:

$$p_{11}v_1 + p_{12}v_2 \geq c_1/(1-d_1)\,\delta q_1 = R_1; \tag{IC-1}$$

$$p_{21}v_1 + p_{22}v_2 \geq c_2/(1-d_2)\,\delta q_2 = R_2, \tag{IC-2}$$

where R_i is the incentive rent for activity $i = 1, 2$.

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

$$(1-d_1)\,p_{11}N_1 + (1-d_2)\,p_{21}N_2 + H_1 = N_1; \tag{FL-1}$$

$$(1-d_1)\,p_{12}N_1 + (1-d_2)\,p_{22}N_2 + H_2 = N_2, \tag{FL-2}$$

where $H_i \geq 0$ is the mass of new workers hired into activity 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 activity, 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_1/N_2 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** $w = [w_i]_i$ 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 activity 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 activity i has **full job security**; that is, a worker performing activity 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

We now characterize the optimal personnel policy. Given the span and size of the firm, (N_1, N_2), the firm chooses an optimal personnel policy (H, w, p) to solve the following program:

$$W(N_1, N_2) = \min_{(H,w,P)} w_1 N_1 + w_2 N_2$$

subject to $(PK - i), (IR - i), (IC - i)$, and $(FL - i)$. That is, the firm chooses hiring, wage, and assignment policies to minimize the steady-state wage bill. 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 assume that the incentive rents for activity 2 are higher than the incentive rents for activity 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 activity 2 exceed those provided in activity 1 (i.e., $v_2^* > v_1^*$). For reasons that will soon become clear, we refer to activity 1 as the **bottom job** and activity 2 as the **top job**. We also refer to workers who perform activity 1 as **bottom workers** and those who perform activity 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. We will assume this is the case in what follows.

ASSUMPTION 1 (Top jobs are scarce). $N_2 d_2 \leq N_1 (1 - d_1)$.

Assumption 1 reflects what is likely to be the overwhelmingly more common situation, so this is where we focus our attention in the analysis. In the appendix, we solve for the optimal personnel policy for the full model.

LEMMA 1. *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 departures at the top. Consider the parallel-careers 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 incentive 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 the incentive rents associated with activity* 2*. 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 parallel-careers benchmark, high 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 the incentive rents for activity 2 in order to exert effort. When there are sufficient separation rents, promotion prospects alone provide enough motivation for bottom workers, so that their incentive-compatibility constraints are slack. 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, the incentive rents for task 2 would not be sufficient to motivate them. Since they receive the incentive rents for activity 2 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. 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 in excess of the incentive rents for activity 2. There are no demotions, bottom workers receive full job security, and there is forced turnover at the top.*

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 higher 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 the incentive rents for activity 2. 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 higher 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 offering high wages throughout the firm. Moreover, if the firm offers rents that exceed the incentive rents for activity 2 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 forced turnover to demoting workers at the top. Forced turnover and demotions create 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. Forced turnover 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.

Finally, it is worth remarking that optimal wages, promotion prospects, and forced-turnover rates depend on (N_1, N_2) only through the span, N_1/N_2. This is because for any (N_1, N_2) for which top jobs are scarce, hiring only occurs in the bottom, and bottom workers receive zero rents. Wages at the bottom are therefore determined by bottom workers' promotion prospects, which depend on the firm's span. Wages at the top are determined by the incentive rents for task 2 and the forced-turnover rate, which depends on the firm's span.

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. (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 correspond to rents that are lower than the incentive rents for activity* 1*. Top-job wages correspond to rents that exceed the incentive rents for activity* 2 *whenever there are insufficient separation rents. (iv) Wages, promotion rates, and forced-turnover rates depend on* (N_1, N_2) *only through the span,* N_1/N_2*.*

Proposition 1 characterizes the optimal personnel policies, taking the firm's size and span as given and therefore results in a **labor-cost function**, $W(N_1, N_2)$. We now discuss several properties of the optimal wage policy and labor-cost function. We simplify the expressions by assuming that shirking is detected with probability one (i.e., $q_1 = q_2 = 1$). For our purposes, this restriction is inconsequential. Given N_1 and N_2, the expressions for optimal wages and for the labor-cost function depend on whether or not there are sufficient separation rents (i.e., whether $d_2 N_2 R_2 \geq (1 - d_1) N_1 R_1$). There are sufficient separation rents if $N_1/N_2 \leq \kappa$, where we define the cutoff

$$\kappa \equiv (c_2/c_1) \cdot (d_2/(1 - d_2)).$$

That is, there are sufficient separation rents whenever the firm's span is low and/or the turnover rate of the top job is high. These expressions for wages and for the labor-cost function are described in the following corollary to Proposition 1.

COROLLARY 1. *The following are true.*

(i) When there are sufficient separation rents, wages at the bottom are $w_1 = c_1 - c_1 \kappa N_2/N_1$*, and wages at the top are* $w_2 = c_2/((1 - d_2)\delta)$*. The labor-cost function is*

$$W(N_1, N_2) = c_1 N_1 + \frac{1 - \delta d_2}{1 - d_2} \frac{1}{\delta} c_2 N_2.$$

(ii) When there are insufficient separation rents, wages at the bottom are $w_1 = 0$*, and wages at*

the top are $w_2 = (c_1 N_1 + c_2 N_2) / (\delta N_2)$. *The labor-cost function is*

$$W(N_1, N_2) = \frac{1}{\delta} c_1 N_1 + \frac{1}{\delta} c_2 N_2.$$

In each region, the labor-cost function is linear in N_1 and N_2, so the coefficient on N_i has a natural interpretation as the marginal cost to the firm of adding a position in activity i. Under the Neoclassical model of labor supply in which there are no incentive problems, this coefficient would be equal to the wage for workers in task i, which would be equal to the associated effort cost c_i, since workers' outside options are 0 and they are on the long side of the market.

In contrast, when effort is not contractible, the marginal cost accounts for the effect that adding an additional position in task i affects the firm's optimal personnel policy problem, which in turn depends on whether or not there are sufficient separation rents. When there are sufficient separation rents, the marginal cost of adding a position in activity 1 coincides with the Neoclassical costs of adding the position, c_1, which in turn exceeds the bottom wage, because compensation is backloaded in workers' careers. When there are insufficient separation rents, adding another position at the bottom reduces the promotion prospects for bottom workers and therefore requires that the firm adjust its personnel policies in order to keep bottom workers motivated. The resulting effective marginal cost of adding such a position is then $c_1/\delta > c_1$. Relatedly, there are benefits of adding positions at the top that exceed the marginal revenue product of such positions, since additional positions at the top create promotion opportunities for workers at the bottom, in turn relaxing the firm's optimal personnel-policy problem.

5 Optimal Production

Given the labor-cost function and the production function, we use standard tools from Neoclassical production theory to characterize the optimal span, N_1^*/N_2^*, for a given level of output, y. The firm's span entirely determines its optimal personnel policies, so the firm's production-expansion path describes not only the number of positions $(N_1^*(y), N_2^*(y))$ but also its optimal personnel policies as a function of its scale. We conclude with a description of how personnel policies vary with a firm's scale when production functions are non-homothetic.

There is broad evidence that larger firms tend to have larger spans (see Rushing, 1966; Blau and Scheonherr, 1971; Kasarda, 1971, Colombo and Delmastro 1999 for cross-sectional evidence; see Caliendo, Monte, and Rossi-Hansberg, Forthcoming, for more recent evidence that conditional on organizational structure, firms expand by increasing their span.) We will therefore focus on

non-homothetic production technologies for which **expansion favors activity 1**. We write the firm's revenues as the product of its output price and its output: $F(N_1, N_2) = P \cdot f(N_1, N_2)$. Assumption 2 provides sufficient conditions for a production function f to have an increasing and convex production-expansion path. Assumption 2 includes as an extreme case production functions in which there is a fixed number of top positions in the firm, such as in Zabojnik and Bernhardt (2001) and DeVaro and Morita (2013).

ASSUMPTION 2 (Production Expansion Favors Activity 1). For all $k \geq 0$,

$$kN_2 \left.\frac{\partial (f_1/f_2)}{\partial N_1}\right|_{N_1 = kN_2} + N_2 \left.\frac{\partial (f_1/f_2)}{\partial N_2}\right|_{N_1 = kN_2} > 0.$$

Assumption 2 ensures that the marginal rate of technical substitution between N_1 and N_2 falls along any ray from the origin. In other words, as production expands, N_2 becomes a worse substitute for N_1 in production.

Given the firm's labor-cost function $W(N_1, N_2)$, the firm's optimal production problem is to solve:

$$\max_{N_1, N_2} P \cdot f(N_1, N_2) - W(N_1, N_2).$$

We solve this problem in two steps. We first solve for the cost-minimizing numbers of positions necessary for producing output y. For each y, the cost-minimizing numbers of positions in turn determine the optimal personnel policy. We then solve for the optimal output y^*, which determines the firm's optimal scale.

Given an output level y, the firm wants to choose $(N_1^*(y), N_2^*(y))$ to solve the following cost-minimization problem

$$C(y) = \min_{N_1, N_2} W(N_1, N_2)$$

subject to $f(N_1, N_2) \geq y$. This cost-minimization problem will trace out the set of conditionally efficient input pairs. From Corollary 1, we know that $W(N_1, N_2)$ is piecewise linear in (N_1, N_2), and the coefficients on N_1 and N_2 depend on whether the firm operates in the sufficient separation rents region or the insufficient separation rents region. Figure 2 below depicts the producer-theory approach to the firm's cost-minimization problem. The isocost is piecewise linear with different coefficients on either side of the $N_1 = \kappa N_2$ boundary.

Each of the three isoquants represents a different production technology producing the same level of output y. Isoquant 1 is an activity-1-heavy production technology and will favor production at

point A at which the firm operates with a large span and has insufficient separation rents. Isoquant 3 is an activity-2-heavy production technology and will favor production at point C at which the firm operates with a small span and has sufficient separation rents. At each of these points, the marginal rate of technical substitution, f_1/f_2, is equal to the cost ratio, W_1/W_2. By Corollary 1, the cost ratio is c_1/c_2 at point A, and it is $(c_1/c_2)\cdot(1-(1-\delta)/(1-\delta d_2))$ at point C. For intermediate production technologies such as the one that generates isoquant 2, if the marginal rate of technical substitution, f_1/f_2, at $N_1 = \kappa N_2$ is between c_1/c_2 and $(c_1/c_2)\cdot(1-(1-\delta)/(1-\delta d_2))$, then the firm optimally produces at point B, which lies on the boundary between the insufficient separation rents region and the sufficient separation rents region.

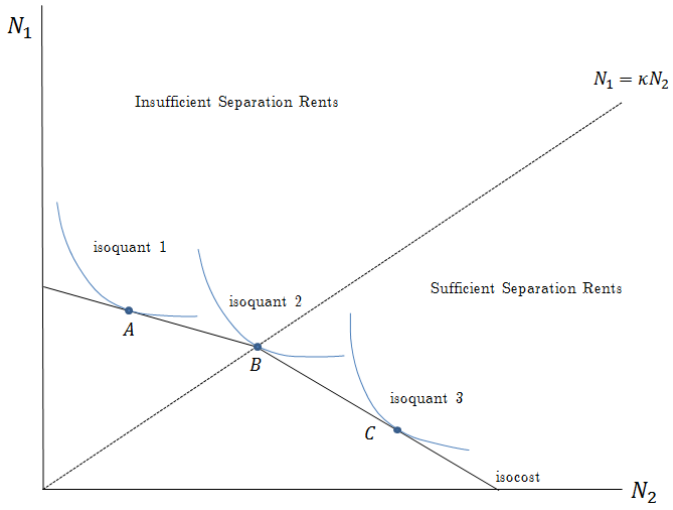

Figure 2: Producer-theory approach to firm's cost-minimization problem. The isocost curve is piecewise linear with different coefficients on either side of the dotted boundary. The isoquants represent different production technologies, with higher-numbered isoquants representing production technologies increasingly favoring activity 2 relative to activity 1.

The conditionally efficient number of positions $(N_1^*(y), N_2^*(y))$ trace out a production-expansion path and determine a minimized production cost $C(y)$. Because the labor-cost function is kinked, the appropriate generalization of the Kuhn-Tucker conditions (Wolkowicz, 1983) for the optimal masses of each position are

$$\begin{aligned} C_+'(y)\cdot f_1(N_1^*(y), N_2^*(y)) &\geq W_{1+}(N_1^*(y), N_2^*(y)) \\ C_+'(y)\cdot f_2(N_1^*(y), N_2^*(y)) &\leq W_{2+}(N_1^*(y), N_2^*(y)), \end{aligned}$$

where W_{j+} is the right derivative of W with respect to N_j and C_+' is the right derivative of the cost function. These hold with equality everywhere except on the boundary. When production

expansion favors activity 1, the associated production-expansion path will be convex within each of the two regions, and it will be linear on the boundary. Figure 3 below depicts a production-expansion path for such a production function.

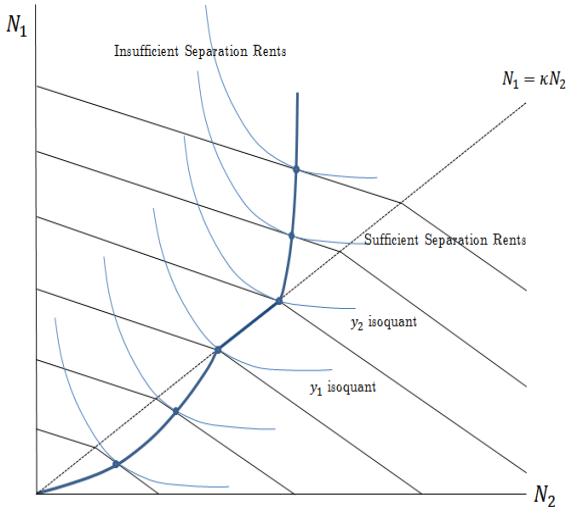

Figure 3: This figure plots a production-expansion path for a non-homothetic production technology for which expansion favors activity 1. Firms producing at high levels of output optimally operate in the insufficient separation rents region. Firms producing at low levels of output optimally operate in the sufficient separation rents region. Firms producing intermediate levels of output optimally operate on the boundary between the two regions.

Figure 3 highlights the result, summarized in Lemma 3, that the firm's optimal output level determines whether it produces in the sufficient separation rents region or the insufficient separation rents region. As a result, the firm's optimal output level determines the firm's span and therefore its optimal personnel policies.

LEMMA 3. *Suppose production expansion favors activity 1. Then there exists two cutoffs,* y_1 *and* y_2 *such that the following are true. (i) if* $y^* < y_1$*, the firm's optimal span is* $N_1^*/N_2^* < \kappa$*, (ii) if* $y^* \in [y_1, y_2]$*, the firm's optimal span is* $N_1^*/N_2^* = \kappa$*, and (iii) if* $y^* > y_2$*, the firm's optimal span is* $N_1^*/N_2^* > \kappa$*.*

We can compare the optimal personnel policies of firms that operate at large scales and those that operate at small scales. Suppose there is a small firm (denoted by superscript S) that operates at $y^{*S} < y^1$ and a large firm (denoted by superscript L) that operates at $y^{*L} > y^2$. Lemma 4 summarizes key differences in the optimal personnel policies for these two firms.

LEMMA 4. *Suppose production expansion favors activity 1.* $w_1^{*L} > w_1^{*S}$, $w_2^{*L} > w_2^{*S}$, $p_{12}^{*L} < p_{12}^{*S}$, $p_{22}^{*L} < p_{22}^{*S}$, and $v_2^{*L} > v_2^{*S}$.

Given the conditionally efficient numbers of positions $(N_1^*(y), N_2^*(y))$, the firm then wants to choose a level of output to solve the unconstrained program

$$\max_y Py - C(y).$$

By Corollary 1, $W_1(N_1, N_2) > w_1^*$ and $W_2(N_1, N_2) < w_2^*$. Optimal production occurs at the point where $C'(y) = P$, so by the optimality conditions above, bottom workers are paid less than their marginal revenue product, and top workers are paid more than their marginal revenue product. Creating additional positions at the top relaxes the firm's incentive problem and therefore the firm will create more positions than would equalize the marginal revenue product with the wage paid to top workers. Similarly, creating additional positions at the bottom tightens the firm's incentive problem, so the firm will create fewer positions than would equalize the marginal revenue product with the wage paid to bottom workers. These results are summarized in Lemma 5, where we denote by $MRP_i^* = P \cdot f_i(N_1^*(y^*), N_2^*(y^*))$.

LEMMA 5. *For any production function f, at the optimum, $w_1^* < MRP_1^*$ and $w_2^* > MRP_2^*$.*

We now examine how an increase in the voluntary turnover rate at the top, d_2, affects optimal production and optimal personnel policies. Figure 4 depicts the effects of an increase in d_2 for a given labor-cost level. Increasing d_2 increases the threshold span κ below which the firm operates in the sufficient separation rents region. Further, in the sufficient separation rents region, an increase in d_2 increases the wage necessary to motivate top workers, with no offsetting effect on the motivation of bottom workers, leading to a counterclockwise rotation of the isocost line.

Firms operating at points like A are unaffected by an increase in the voluntary-turnover rate at the top, since they optimally offset this increase by decreasing the forced-turnover rate. Firms operating at points like C optimally reduce the number of positions at the top and the bottom and instead produce at a point like C'. Firms operating at boundary span points like B will reduce production and will reduce N_2, but depending on the production technology, they might increase or decrease N_1.

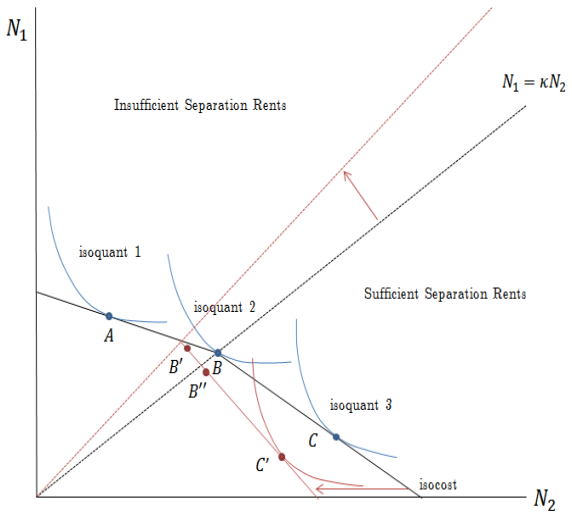

Figure 4: This figure examines the effects of an increase in the voluntary departure rate at the top. Holding labor costs constant, this increase rotates the boundary between the ISR and SSR regions counterclockwise, and it rotates the isocost curve in the SSR region clockwise. If point A was optimal, then it remains optimal. If point B was optimal, then the new optimum will be either B' or B". If point C was optimal, then the new optimum will be C'.

When there are sufficient separation rents, an increase in the voluntary departure rate requires that the firm increase its wages at the top in order to satisfy top-workers' incentive-compatibility constraints. In turn, they will reduce the number of positions at the top (i.e., $dN_2^*/dd_2 < 0$), and since there are complementarities between activity 1 and activity 2, they will also reduce the number of positions at the bottom (i.e., $dN_1^*/dd_2 < 0$). These effects of an increase in the voluntary-turnover rate at the top are summarized in Lemma 6.

LEMMA 6. The following are true. (i) If $N_1^*/N_2^* > \kappa$, then $dN_1^*/dd_2 = dN_2^*/dd_2 = 0$, (ii) if $N_1^*/N_2^* = \kappa$, then $dN_2^*/dd_2 < 0$, and (iii) if $N_1^*/N_2^* < \kappa$, then $dN_1^*/dd_2 < 0$ and $dN_2^*/dd_2 < 0$.

Many other comparative-statics results are possible, but we have narrowed our focus to results that we will discuss in more detail in the next section.

6 Discussion of Empirical Implications

Our model delivers a rich set of predictions that are consistent with a broad pattern of evidence, and it provides a framework to think about how the need to manage workers' promotion prospects interact with firm size and the firm's demographics. In this section, we first highlight a number of predictions that accord with a host of stylized facts that prompted the development of early

models of internal labor markets. These predictions are core to the individual dynamic moral hazard problems between the firm and each worker.

We then highlight two sets of predictions that go beyond the standard facts and arise precisely because workers' careers are managed at the firm-level and therefore involve trade-offs among workers. The first set of implications examine how firm size affects its personnel policies, and relatedly, the career dynamics of its workers. The second set of implications relate the demographics of the workers to the organizations of the firms. Each of the facts we discuss is likely due to and consistent with many factors beyond the scope of our model, but taken as a whole, they are supportive of our model's main mechanisms and trade-offs.

6.1 Features of Internal Labor Markets

Proposition 1 shows that optimal personnel policies give rise to hiring and mobility patterns that are consistent with the functioning of internal labor markets.

OBSERVATION 1. *Employees perform different activities at different stages of their career with the firm.*

In other words, there is an internal labor market. A core force in our model is the optimal backloading of incentive rents through activity assignments. This force offers a clear rationale for why organizations offer careers internally, an idea that dates back at least to Weber (1947) and was advanced in Doeringer and Piore (1971)'s seminal work on internal labor markets. Because workers in our model are homogeneous, our main mechanism operates independently of the alternative forces of learning about the qualities of employees and firm-specific human-capital acquisition by employees that have also been proposed as reasons why firms link sequences of activities together into careers. We would therefore expect careers within organizations to be important even in settings in which firm-specific human capital and employer learning is less important.

OBSERVATION 2. *Employees begin their careers in the bottom job.*

According to Doeringer and Piore (1971), "Entry into such markets is limited to particular jobs or ports of entry." Indeed, ports of entry appear to be the rule in some industries and countries. For example, Milgrom and Roberts (1992) report that "airlines used a strict system of hiring only at the bottom of the job ladder." As a result, even experienced pilots who lost their jobs due to the industry shake-out in the late 1980s and early 1990s started over at the bottom when they changed airlines. Ken'ichi and Hiroyuki (1988) observe that entry into Japanese firms is limited in the sense that most hiring is done at the time of graduation and mid-career ports of entry are almost

non-existent. A port of entry is an extreme feature of a personnel policy that is, of course, not present in every firm. For example, Baker, Gibbs, and Holmstrom (1994) study detailed personnel records from a large U.S. firm and find that 25% of workers filling higher positions in the firm are hired externally. We discuss below how our model's stark result that there is a port of entry can be cast in terms of insider bias in hiring at the top.

OBSERVATION 3. *Workers are never demoted.*

With a couple recent exceptions (Dohmen et al., 2004; Lin, 2005), studies of detailed personnel data suggest that demotions are rare. For example, in Baker, Gibbs, and Holmstrom (1994)'s study, demotions almost never occur. Seltzer and Merrett (2000) report similar findings using data from an Australian bank, and Treble et al. (2001) find that demotions are rare in a British service-sector firm. In our model, a firm might in principle want to demote top workers in order to create opportunities for bottom workers. However, because the outside option of top workers is weakly higher than the payoffs bottom workers receive, top workers are better off being forced out of the firm than being demoted. In turn, the firm is better off adopting forced-turnover policies rather than demoting top workers.

OBSERVATION 4. *There are wage jumps at promotion.*

Many studies have found that promotions are associated with large wage increases (Murphy, 1985; Lazear, 1992; Baker, Gibbs, Holmstrom 1994a,b; McCue 1996). The wage increases may result from a number of factors such as human capital accumulation (Gibbons and Waldman 1998) and signaling of ability to the outside market (Waldman 1984). Lazear and Rosen (1981)'s labor-market tournament model provides one of the first incentive-based explanations for large wage increases upon promotion–these wage increases are used to provide incentives for effort for workers at the bottom of the job ladder. But in principle, the wage increases could have taken the form of a large one-time bonus. In our model, these wage increases optimally serve not only to provide incentives for bottom workers, but they also provide incentives for top workers. In other words, if a worker is willing to work hard to get a promotion, he will also be willing to work hard to keep that job to which he has been promoted.

OBSERVATION 5. *Wages at the bottom are below workers' marginal revenue product. Wages at the top exceed workers' marginal revenue product.*

Lemma 5 shows that at the conditionally optimal scale, bottom workers' wages are below their marginal revenue product. This reluctance to hire workers whose marginal revenue product exceeds their wages arises, because hiring an additional worker at the bottom reduces the promotion

prospects of *other* workers at the bottom, giving rise to a shadow cost of additional bottom positions. This wedge between marginal revenue product and wages can manifest itself as often-lamented "headcount restriction" policies that human-resource departments put in place.

On the flip side, top workers' wages exceed their marginal revenue product. Taken together with the prediction that workers start with the bottom job and receive a wage below their marginal product of labor, this implies that the wage growth increases faster than the productivity growth. This is a well known prediction from the incentive-based theory of labor market, and has received considerable empirical support; Lazear (1979), Medoff and Abraham (1981), Lazear and Moore (1984), Hutchens (1987) and Kotlikoff (1992). The existing theories typically focus on wage gains on the job. Here, we show that another source of wage gain is that older workers are assigned to better paying jobs.

6.2 Organizational Demographics

Our results suggest that workers' careers are optimally interlinked: firms may adopt policies that affect one set of workers in order to improve incentives for a different set of workers. Further, changes in labor-market conditions that affect turnover at one level of the organization will, through endogenous changes in personnel policies, affect workers' entire career paths.

OBSERVATION 6. *Firms may adopt forced-turnover policies to create promotion opportunities.*

In the United States, prior to 1986 when it was outlawed, many firms put in place mandatory-retirement programs often with the stated objective of creating promotion opportunities for the young. For example, Cappelli (2008) reports that executives at Sears put in place mandatory retirement policies "entirely to keep the lines of advancement open." The U.S. Department of Labor (1981) surveyed employers regarding this practice and summarized their results as follows, "When firms were asked for reasons for using mandatory retirement, all firms, but particularly large firms, put greatest emphasis on assuring promotional opportunities for younger workers." Recently, Hong Kong Civil Service Bureau (2014) in determining whether to increase the mandatory retirement age for civil servants wrote, "Any proposal for extending the service of staff beyond retirement age must be carefully balanced against its adverse impacts on the promotion prospects of serving officers and the need for healthy injection of new blood into the civil service."

Forced-turnover policies are not limited to mandatory-retirement programs. Many firms, including GE, Motorola, Dow Chemical, IBM, and in the past, Microsoft, put in place "stack ranking" or "vitality curve" policies in which a fraction of workers at each level of the hierarchy is regularly dismissed. Descriptions of these policies often emphasize both the motivational effects of dismiss-

ing low-performing workers and that dismissing workers in higher positions creates opportunities throughout the firm.

Lazear (1979) provides a justification for mandatory-retirement policies as being part of an optimal long-term employment contract in which wage payments are backloaded in order to motivate workers, and the value of a workers' entire wage stream equals the entire stream of his contribution to profits. At termination, a worker's spot wages optimally exceed his marginal product, and therefore retirement would not be ex-post voluntary and so has to be mandated. In our model, as in Lazear's, backloaded compensation implies that workers are paid less than their marginal product when young and more than their marginal product when old. However, older workers are not fired *because* their wages exceed their marginal products–their replacements, old or young, will also have to be paid wages exceeding their marginal products. Rather, old workers may be fired precisely to increase the vertical flow of workers through the organization. This result holds *even though* young workers know they may be forced out after being promoted.

As we will see in the next section, frictions in wage payments are key to our results on forced-turnover policies. In firms that make extensive use of pay-for-performance bonuses, forced-turnover policies are unnecessary. Our model therefore suggests that forced-turnover policies are more likely to be prevalent in organizations in which pay is less flexible.

OBSERVATION 7. *Increased turnover at the top may lead to less employment at the bottom and less hiring at the bottom.*

Even though forced-turnover policies may be optimal for some firms, the desirability of pushing workers out in order to create opportunities for others does not hold universally. Yet this motive has been cited extensively as a justification for increasing the generosity of government retirement programs. For example, in the UK, the Job Release Scheme "was introduced in 1977 and was described as 'a measure which allows older workers to retire early in order to release jobs for the registered unemployed.'" (Banks, Blundell, Bozio, Emmerson, 2010 p. 7). Changes in the skill mix notwithstanding, the argument has some intuitive appeal. After all, in the steady state, hiring at the bottom of the organization is carried out exactly to offset departures from the firm. That is, $H_1^* = d_1 N_1^* + D_2^* N_2^*$, where D_2^* represents the sum of the voluntary and involuntary departure rates at the top. All else equal, an increase in the rate of voluntary departures at the top increases hiring at the bottom, since $\partial H_1^*/\partial d_2 = N_2^* > 0$. However, in response to an increase in the voluntary departure rate, firms optimally adjust their personnel policies and the number of workers they employ. That is,

$$\frac{\partial H_1^*}{\partial d_2} = d_1 \frac{\partial N_1^*}{\partial d_2} + d_2^* \frac{\partial N_2^*}{\partial d_2} + \frac{\partial D_2^*}{\partial d_2} N_2^*.$$

We show in Lemma 6 that an increase in the voluntary departure rate at the top of the firm can lead to a decrease in the steady-state employment level at the bottom of the firm. The reasons for this are two-fold. First, when there are insufficient separation rents, firms adopt forced-turnover policies. An increase in the rate of voluntary departures causes the firm to scale back on these forced-turnover policies (i.e., $\partial D_2^*/\partial d_2 = 0$), but they otherwise make no other changes. In other words, firms are already able to expand opportunities for entry-level workers by increasing turnover at the top and will do so themselves when they find it profitable. Second, when there are sufficient separation rents, an increase in the voluntary departure rate causes the firm to reduce the number of positions at the top and at the bottom (i.e., $dN_i^*/dd_2 < 0$).

Empirically, the effects of increased retirement rates on the employment of younger workers is exactly the opposite of the objectives stated for the Job Release Scheme. Using changes in the generosity of government retirement programs in twelve countries in the 20th century, several authors have shown that when government retirement programs became more (less) generous, older workers retired earlier (later), and youth and prime-age unemployment went up (down). (Gruber and Wise, 2011). The flow constraint in our model captures the intuition behind proposals like the Job Release Scheme, but it also highlights organizational reasons for why its intended outcomes might fail to materialize.

6.3 Firm Size, Span, and Workers' Careers

In this section, we contrast the careers of workers working in large firms relative to those of workers working in small firms. We group our model's predictions into static, cross-sectional observations and dynamic predictions relating to a worker's entire career. The first set of predictions speaks to the firm size-wage effect that has been widely documented in labor economics (see Oi and Idson, 1999 for a survey). The second set of predictions is consistent with the findings of many disparate single-firm studies.

To think about these issues in the context of our model, we make use of the empirical pattern that larger firms tend to have larger spans (see Rushing, 1966; Blau and Schoenherr, 1971; Kasarda, 1971, Colombo and Delmastro 1999 for cross-sectional evidence of this pattern; see Caliendo, Monte, and Rossi-Hansberg, Forthcoming, for recent evidence related to within-firm growth). That is, we make the assumption that expansion favors activity 1, and we explore the implications of this assumption for the effects of firm size on workers' careers. As in Lemmas 3 and 4, we will consider a small firm to be one that operates in the sufficient separation rents region, and we consider a large firm to be one that operates in the insufficient separation rents region.

OBSERVATION 8. *Larger firms pay higher wages for all workers.*

A positive relationship between firm size and wages has been documented in many studies going back to at least Moore (1911) (Brown and Medoff, 1989; see Oi and Idson, 1999 for a review of the literature). Our explanation is closest to the efficiency-wage story that was originally posited as an explanation for the size-wage effect. Under the efficiency-wage explanation, larger firms have relatively worse monitoring technologies, which corresponds to a lower q_i in our model, and therefore have to offer higher wages for all workers.

In contrast to the efficiency-wage explanation, in our model, firms need not possess different technologies in order for a size-wage effect to exist. In our model, a small firm has a small span, which means that bottom workers' promotion prospects are relatively strong. As a result, the firm can offer bottom workers a lower wage, while still keeping these workers motivated. In such firms, top workers have full job security and therefore can be motivated with a relatively lower wage. Our model generates a firm size-wage effect, but the primary mechanism through which it operates is through a firm span-wage effect. We would therefore predict that controlling for a firm's span in a regression of wages on firm size should reduce the magnitude of the size-wage effect. To the best of our knowledge, there are no studies showing a size-wage effect controlling for firm span or promotion prospects.

OBSERVATION 9. *Larger firms have more of an insider bias in hiring at the top.*

The stark result that in an optimal personnel policy, hiring only occurs at the bottom is due in part to worker homogeneity. The result is more continuous, however, in a way we shall make precise. Suppose the firm has a one-time opportunity to hire into the top job an external candidate whose incremental productivity over existing workers is Δ (i.e., the incremental increase in the NPV of future profits from hiring this worker are Δ). Since workers in the top job receive rents of v_2^*, if the firm hires this external worker, the firm has to offer him total rents of v_2^*. In contrast, if the firm instead hires into the bottom job and promotes a bottom worker instead, the firm will pay total rents of 0, since the expected future rents from eventually being promoted are extracted from the bottom worker. As a result, the firm will hire the external candidate into the top job only if $\Delta > v_2^*$.

The firm therefore exhibits an insider bias in hiring into higher-level positions, for which there is empirical support. For example, Huson, Malatesta and Parrino (1994) find that outside CEOs bring about better firm performance; Agrawal, Knoeber, and Tsoulouhas (2006) find that external candidates are superior to internal candidates in observable qualities; and Oyer (2007) finds that there is an insider advantage for tenure decisions for academic economists.

Since larger firms offer greater rents to top workers than smaller firms do, the insider bias in hiring at the top is greater in larger firms. There is extensive support for the idea that larger firms have more of an insider bias for hiring CEOs (Dalton and Kesner, 1983; Lauterbach and Weisberg, 1994; Parrino, 1997; Lauterbach, Vu, and Weisberg, 1999; Agrawal, Knoeber, and Tsoulouhas, 2006). More broadly, in a nationally representative sample of UK firms, recent papers find support for a positive size-insider bias relationship (DeVaro and Morita, 2013; Bond, 2014).

DeVaro and Morita (2013) explain this positive relationship between firm size and insider bias in hiring at the top by arguing that firms differ in the "returns to managerial capability." Firms with greater returns to managerial capability will hire more workers at the bottom and therefore operate at a larger scale. Additionally, the returns to training subordinates to become managers and the returns to promoting from within are higher in such firms. In equilibrium, there is therefore a positive correlation between firm size and insider bias in hiring at the top, driven by unobserved returns to managerial capability. In contrast, our model suggests that this positive relationship is likely to hold even among firms within narrowly defined industries, for which one might expect the returns to managerial capability to be similar.

OBSERVATION 10. *Larger firms have higher starting wages and higher wages for the promoted workers, but the promotion prospects for bottom workers are worse.*

We conclude this section with a discussion of how firm characteristics affect workers' wage and career dynamics. Workers in larger firms begin their employment with higher wages but worse promotion prospects than workers in smaller firms. Their promotion prospects are worse, because larger firms optimally choose to have larger spans, which in essence creates more competition among bottom workers for promotions. To maintain incentives through promotions, larger firms also choose higher wages at the top and put in place forced-turnover policies to create more vacancies at the top. Nevertheless, expected future rents for bottom workers in larger firms are smaller than they are for bottom workers in smaller firms, and therefore to keep bottom workers motivated, the firm offers higher wages at the bottom.

While we are unaware of any studies that directly examine how firm size is related to career dynamics, a number of papers have shown relationships between firm size and various aspects of wage and promotion dynamics. Taken together, their results are consistent with our predictions. For example, Barron, Black, and Loewenstein (1987) and Brown and Medoff (1989) find that larger firms offer higher starting wages. Others find that the wage differential is larger at higher levels of the hierarchy (Brown, Hamilton, and Medoff, 1990; Meagher and Wilson, 2004). In terms of our predictions regarding promotion prospects, Belzil and Bognanno (2008) study the careers of over

30,000 American executives across many firms and find that the rate at which they are promoted to higher positions is negatively related to the size of their employers.

Rebitzer and Taylor (1995) study the labor market for lawyers and, in particular, focus on the relationship between firm size and various aspects of career dynamics. They find that larger law firms offer higher wages to both their associates and partners, and they interpret these findings as evidence against an efficiency-wage model–higher pay for partners in larger firms should be viewed as backloaded pay for associates, implying that associates should have lower wages in larger firms. In larger firms in our model, bigger wage increases upon promotion do not imply lower wages at the bottom, because bottom workers' promotion prospects are lower. Indeed, this is consistent with Galanter and Palay (1991)'s broad study of law firms in which they claim that, "the chances of promotion to partner are accordingly lower in big firms than small firms."

The differences in career dynamics between small firms and large firms, of course, are likely to result from many other factors, notably differences in the quality of labor. As a result, many of the empirical findings above have alternative explanations, and moreover, there are aspects of wage dynamics that cannot be explained by our model. For example, Barron, Black and Loewenstein (1987)'s finding that larger firms offer higher within-job wage growth is beyond the scope of our model, because we assume wages are tied to jobs. Nevertheless, there is some evidence that a firm's hierarchical span is positively related to wage growth upon promotion (Smeets and Warzynski, 2008; Garicano and Hubbard, 2009).

7 Pay-for-Performance Contracts

In our main model, we made the assumption that wages are tied to activities. This assumption ruled out both performance pay and other more flexible compensation arrangements such as seniority-based pay increases within an activity. We now expand the firm's contracting possibilities by allowing firms to write history-contingent pay-for-performance contracts, but we assume that workers are subject to a limited-liability constraint. We therefore solve for optimal history-contingent pay-for-performance contracts, and we show that they share most of the features of optimal personnel policies in the main model. One notable difference, however, is that optimal pay-for-performance contracts no longer require forced turnover, even if the firm's span is large. Contractual flexibility therefore interacts with the organization of internal labor markets.

Specifically, we assume that the firm pays a minimum wage $\underline{w} \geq 0$ at the beginning of each period and a performance-contingent bonus $b_t \geq 0$ at the end of each period. As in the main model, we assume that any worker who is caught shirking is terminated with probability 1, and

he also receives no bonus. Indeed, this constitutes an optimal penal code. If a worker is not caught shirking, the firm pays a bonus $b_t \geq 0$, which can depend on the worker's entire past employment history within the firm. For a worker in his t-th period in the firm, his employment history can be described as $h^t = (h_1, \ldots, h_t)$, where $h_s \in \{1, 2\}$, $s = 1, 2, \ldots, t$ denotes the activity the worker was assigned in period s. We also assume that the firm's assignment policy depends on h^t. Denote $p_i\left(h^t\right)$, $i = 1, 2$ as the probability the worker will be assigned to activity i next period. The complementary probability $1 - p_1\left(h^t\right) - p_2\left(h^t\right)$ is the forced-turnover rate for a worker with employment history h^t. We assume that the firm offers the same contract to all workers, so the firm's optimal personnel policy in this setting can be described by $\left\{b\left(h^t\right), p_1\left(h^t\right), p_2\left(h^t\right)\right\}_{t=1}^{\infty}$.

Since bonuses and activity assignments can depend on the worker's entire employment history, this extension allows for a variety of personnel policies. For example, we are allowing firms to adopt seniority-based promotion policies in which each worker performs, say, activity 1 for a number of periods before being promoted to activity 2. Firms can rotate workers among jobs as well. We are also allowing the firm to backload pay by increasing the size of performance bonuses as the worker accumulates more time on the job. The set of feasible contracts and personnel policies is therefore large, but the optimal personnel policy takes a simple form.

To describe the optimal personnel policy, define the incentive rents under pay-for-performance for activity i as $r_i = (1 - q_i)\, c_i/q_i$. As in the main model, we assume that $r_2 > r_1$, so activity 2 requires greater incentive rents, either because its associated effort costs are higher or because performance is more difficult to monitor. As in the main model, there is a top job and a bottom job. Activity 1 is performed by workers in the bottom job, and activity 2 is performed by workers in the top job.

PROPOSITION 2. *An optimal personnel policy has the following features: (i) Hiring occurs only in the bottom job. (ii) There is a well-defined career path: bottom workers stay in the bottom job or are promoted. The promotion rate is constant and given by* $d_2 N_2/\left((1 - d_1)\, N_1\right)$. *Top workers are never demoted. Workers are not fired unless they are caught shirking. (iii) The performance bonus in the top job is constant and independent of the firm's span. The performance bonus in the bottom job is also constant, and it is equal to zero if the span* N_1/N_2 *is below a threshold and positive and increasing in the span above this threshold.*

Proposition 2 shows that performance bonuses within each job are stationary–under the optimal personnel policy, pay is optimally backloaded across jobs rather than within jobs. An internal labor market therefore emerges. New hires enter the firm through a port of entry in which they perform the low-rent activity; incumbent workers climb a job ladder, and there are no demotions on the

equilibrium path. As in the main model, the firm's span affects the form of the optimal personnel policy.

As in the main model, if the firm's span is below a threshold, separation rents created from voluntary turnover at the top, along with minimum-wage payments are enough to motivate bottom workers. In this sufficient-separation rents case, workers in the top job receive the minimal incentive bonus necessary to motivate them. Workers in the bottom job are not given performance bonuses, and therefore their pay in each period is equal to $\underline{w}$. No workers are terminated unless they are caught shirking. As in the main model, top workers receive rents equal to the incentive rents for activity 2, and workers in the bottom are motivated by the rents they get from receiving minimum-wage payments and from the prospects of being promoted to the top job. When there are sufficient separation rents, promotion prospects are therefore sufficiently strong to ensure bottom workers remain motivated, even though they receive no immediate bonuses for their work.

If the firm's span exceeds the threshold, we say that there are insufficient separation rents. Top workers again receive the minimal incentive bonus necessary to motivate effort. Workers in the bottom job now receive positive performance bonuses, and the bonus amount increases with the firm's span. Again, no workers are terminated unless they are caught shirking. As in the main model, the firm adjusts its personnel policy to provide additional incentives for the bottom workers when the prospects of being promoted to the higher-paying job alone is not enough to motivate bottom workers. In contrast to the main model, however, the firm does not do so by increasing rents and putting in place forced-turnover policies at the top. Rather, the firm optimally increases performance bonuses at the bottom.

Before focusing on the differences, we note that when pay-for-performance contracts are possible, optimal personnel policies share many features with those of the main model. These similarities arise because the same set of economic forces are at work both when wages are tied to jobs and when bonuses payments are allowed, but the workers are subject to limited-liability constraints. In particular, optimal incentive provision at the individual-worker level implies that firms should first assign workers to the low-rent activity before promoting them to the high-rent activity, since doing so allows the firm to reuse the incentive rents for the top job to motivate both top and bottom workers. This mechanism is responsible for the emergence of internal labor markets. At the firm-level, the flow constraint implies that the firms' span affects workers' promotion prospects, and therefore the firm adjusts its personnel policies to be consistent with the firm's hierarchical structure.

Important differences in the optimal personnel policy do arise when pay-for-performance is

allowed. In particular, forced-turnover policies are unnecessary. When promotion prospects are weak, the firm now uses bonuses to motivate bottom workers. Using bonuses involves only a monetary transfer from the firm to the workers. In contrast, using promotions typically involves both a transfer and other types of distortions. Forced-turnover policies, for example, increases top workers' effective discount rates, limiting the firm's ability to extract rents from top workers. As a result, forced-turnover policies are more likely to be adopted in settings in which a firm's ability to put in place pay-for-performance contracts is limited. This leads to the following observation.

OBSERVATION 11. *Mandatory retirement is more prevalent in large organizations in settings in which pay-for-performance contracts are limited.*

We are unaware of any systematic investigation into how the prevalence of pay-for-performance contracts relates to the adoption of mandatory-retirement policies. Casual empiricism suggests that occupations with mandatory-retirement policies are often those in which pay-for-performance contracts are rare–judges, police and military officers, government officials, and clerks. To the extent that effective pay-for-performance contracts are differentially limited by firm-level heterogeneity in monitoring technology, our model suggests that mandatory-retirement policies are more likely to be adopted in firms in which worker performance is more difficult to measure. More generally, Observation 11 reinforces the idea that there are complementarities among personnel policies. Difficulties in implementing pay-for-performance contracts may render necessary the use of mandatory-retirement policies. Conversely, where mandatory-retirement policies are impossible to put in place, the firm's returns to putting in place pay-for-performance contracts are higher.

Finally, we note that even when pay-for-performance contracts are feasible, they are not necessarily used to motivate bottom workers.

OBSERVATION 12. *Both bonus and promotion are used to motivate bottom workers, but bonus is less likely to be used when the promotion prospects are high.*

This observation contributes to the discussion of "promotion-based incentives versus bonus-based incentives." Baker, Jensen, and Murphy (1988) highlight that the prevalence of promotions as incentive instruments within firms is puzzling, because "promotion-based incentive schemes appear to have many disadvantages and few advantages relative to bonus-based incentive schemes." This puzzle arises, because if it costs the firm one dollar to provide one dollar in pay to a worker, then the firm should reward performance with money rather than with distorted decisions that move the firm away from productive efficiency. In our model, promotions serve as an optimal way to re-use rents–the firm's costs of using promotions to motivate workers is therefore zero when there

are sufficient separation rents. The marginal cost of using promotions to motivate workers becomes positive when there are insufficient separation rents, and in this case, Baker, Jensen, and Murphy's intuition prevails–the firm will indeed choose to put in place pay-for-performance contracts to provide incentives for bottom workers.

8 Conclusion

This paper shows that career ladders arise naturally within organizations in response to contractual imperfections. Jobs requiring lower levels of incentive rents serve as ports of entry, and workers are motivated in part by the opportunity to advance to jobs requiring, and therefore delivering, higher levels of incentive rents. When promotion opportunities are naturally limited by the firm's hierarchical structure and the voluntary departure rate of its employees, firms optimally push out higher-level employees in order to keep the lines of advancement open. Firms may also optimally alter their hierarchical structures, becoming more top heavy, in order to expand promotion opportunities.

The model is sufficiently tractable to embed in a market setting, therefore allowing us to study the effects of labor-market policies on the careers of workers, which we do in a separate paper. We show that progressive taxation, which disproportionately affect tops workers has indirect effects on bottom workers–fewer workers are hired at the bottom, but the workers that are hired have greater promotion opportunities. If firms are subject to employment-protection legislation that introduces costs to adopting forced-turnover policies, optimal personnel policies involve lower wages at the top and fewer positions at the top, which in turn reduces bottom workers' promotion prospects. Finally, we demonstrate that minimum-wage policies can either increase or decreasing employment in the firm.

We have deliberately abstracted from many of the conventional forces that have been identified in the literature, including employer learning, human capital acquisition, and signaling, in order to emphasize the richness of empirically relevant patterns that are generated by this single force. These forces can be incorporated into the model, however, and the resulting interactions can generate relevant patterns. For example, by allowing worker-level heterogeneity and employer learning about worker quality, the associated optimal personnel policy is consistent with Medoff and Abraham (1982)'s seniority-wage puzzle that worker tenure in a job is associated with higher wages but not higher performance. In our model, since wages on the current job and promotions serve as substitute mechanisms for motivation, a worker who has been revealed to be a particularly bad fit for promotion will have to be compensated with higher wages in the current job in order to

maintain motivation. Selection would therefore account for the seniority-wage puzzle.

The model currently considers firms with production functions requiring two activities to be performed. Allowing for more activities would generate a career ladder with more than two levels and would allow us to study how forced-turnover rates, wages, and promotion policies differ throughout the hierarchy. Our model's main mechanism suggests that wages ought to be increasing at an increasing rate as workers climb the career ladder–larger wage increases at higher levels provide stronger incentives than corresponding wage increases at lower levels (Rosen, 1986).

We also assume that workers' outside options are exogenous. In human-capital-intensive industries, many firms adopt practices that intentionally or unintentionally increase workers' outside options. For example, firms often provide training that increases a worker's general human capital (Becker, 1975), and many firms offer outplacement services for workers whose jobs are eliminated. Some firms, especially in the management consulting industry, actively invest in placing workers who are forced to leave because of up-or-out policies. A McKinsey insider commented that, "if international companies stopped recruiting former McKinsey staff, it could clog the 'up or out' refining process." In the context of our model, if training increases the outside option of top workers by more than it increases the outside option for bottom workers, it increases the value that workers place on being promoted and can therefore help reduce distortions in the firm's hierarchical structure and wages. The model therefore suggests important interactions between firms' training and outplacement policies, and the rest of their personnel policies.

Finally, we have focused on a steady-state analysis. The firm's size and hierarchical structure therefore do not change over time. Allowing for a non-stationary environment would allow us to examine how firm growth interacts with a firm's optimal personnel policies. It seems natural to think that a firm experiencing a higher growth rate can better rely on promotion incentives to motivate their workers. At some point, however, high-growth firms mature, and their growth slows. Understanding how firms optimally change their personnel policies in response to a slowdown in growth is an intriguing theoretical question with important practical implications.

9 Appendix

9.1 Proof of Lemma 0.

For task i, the firm will choose a wage that ensures the incentive-compatibility constraint holds with equality:

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

which gives us

$$v_i^{pc} = \frac{c_i}{q_i\left(1-d_i\right)\delta}$$

9.2 Proof of Lemma 1.

Proof. For convenience, we introduct 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

$$\begin{aligned}
W &= w_1N_1 + w_2N_2 \\
&= N_1(v_1 + c_1 - \delta\left(1-d_1\right)\left(p_{11}v_1 + p_{12}v_2\right)) + N_2(v_2 + c_2 - \delta\left(1-d_2\right)\left(p_{21}v_1 + p_{22}v_2\right)) \\
&= N_1c_1 + N_2c_2 + v_1\left(N_1 - \delta((1-d_1)\,p_{11}N_1 + (1-d_2)\,p_{21}N_2)\right) \\
&\quad + v_2\left(N_2 - \delta((1-d_1)\,p_{12}N_1 + (1-d_2)\,p_{22}N_2)\right) \\
&= N_1c_1 + N_2c_2 + v_1\left(N_1 - \delta(N_1 - H_1)\right) + v_2\left(N_2 - \delta(N_2 - H_2)\right) \\
&= N_1c_1 + N_2c_2 + v_1\left((1-\delta)N_1 + \delta H_1)\right) + v_2\left((1-\delta)N_2 + \delta H_2\right),
\end{aligned}$$

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

Therefore, to minimize $w_1N_1 + w_2N_2$, 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, $H_2^* = 0$. ■

9.3 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. ■

9.4 Proof of Lemma 3

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

$$\begin{aligned} M_1 \Delta_1 &= M_1 p_{11} v_1 + M_1 p_{12} v_2 - M_1 R_1 \\ M_2 \Delta_2 &= M_2 p_{21} v_1 + M_2 p_{22} v_2 - M_2 R_2. \end{aligned}$$

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

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

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

$$\begin{aligned} p_{12}^* &= \frac{R_1N_2}{R_1M_1 + R_2M_2} \in (0, 1), \\ p_{22}^* &= \frac{R_2N_2}{R_1M_1 + R_2M_2} \in (0, 1). \end{aligned}$$

Therefore, the optimal solution is

$$v_1^* = 0, v_2^* = \frac{R_1M_1 + R_2M_2}{N_2}.$$

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 $p_{21}^* + p_{22}^* = \frac{R_2N_2}{R_1M_1+R_2M_2} < 1$, 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. ■

9.5 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. ■

9.6 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. ■

9.7 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. ■

9.8 Proof of Corollary 4

Proof. (i) According to Corollary 3, we can calculate $N_1^* = \alpha_1/c_1$ and $N_2^* = (1-d_2)\,\alpha_2\delta/\left(c_2\,(1-\delta d_2)\right)$. 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.