Monopsony and Backloaded Compensation: Theory and Evidence from Public Accountants

Michael Rubens* Bernardo Silveira †

February 18, 2026

Preliminary – please do not circulate

Abstract

In monopsony models, wage markdowns induce deadweight loss and are, therefore, inefficient. However, wage markdowns also arise in models with backloaded efficiency pay, in which they are designed to induce effort and are, thus, efficient. These models are non-nested, as they assume either exogenous effort or perfectly competitive labor markets. Therefore, we build a model that features both imperfect labor market competition and endogenous effort and use it to derive empirical tests for whether wage markdowns are due to monopsony power and/or backloaded wage contracts. We implement these tests using novel and uniquely-rich data on 569 U.S. public accounting firms. Our results indicate that these firms carry substantial monopsony power, with a lifetime wage being marked down by 27% below lifetime MRPL. However, 'short-run' wage markdowns are even larger for junior employees, whereas senior employees are paid *above* their marginal revenue product of labor, which indicates the use of backloaded efficiency wage contracts.

Keywords: Monopsony, Wage Markdowns, Team Production, Implicit Contracts, Dynamics, Effort Incentives
JEL codes: J42, L84, M52

*UCLA Economics and NBER and CEPR, rubens@econ.ucla.edu
†UCLA Economics and NBER, silveira@econ.ucla.edu
We thank Daniel Haanwinckel and Moritz Meyer-ter-Vehn for useful feedback.

1 Introduction

There is increasing evidence of monopsony power on labor markets (Card, 2022; Azar and Marinescu, 2024; Caldwell et al., 2025; Kline, 2025). Prior monopsony models mostly treat labor markets as spot markets in which worker effort is exogenous. This implies that wage 'markdowns' below the marginal revenue product of labor is a function of a residual labor supply elasticity (Card et al., 2018; Berger et al., 2022).[1] However, wage markdowns are not necessarily due to the exertion of monopsony power. A large literature on backloaded incentive payments features wage markdowns as a means to incentivize effort of forward-looking employees (Lazear, 1981; Lazear and Rosen, 1981; Malcomson, 1984). These two sources of wage markdowns have markedly different welfare implications. In the classical monopsony model, eliminating markdowns reduces deadweight loss, and eliminating cross-firm variation in markdowns mitigates misallocation. By contrast, markdowns arising from backloaded efficiency wages play a key incentive role, and removing them can weaken worker effort and potentially reduce welfare.

In this paper, we investigate the sources of wage markdowns by empirically testing whether wage markdowns are due to monopsony power and/or backloaded incentive pay. We start by constructing a theoretical model that allows both for monopsony power and backloaded incentive pay. We tailor the model to the context of the professional services industries—such as law firms and consultancies—which are characterized by 'up-or-out' promotion policies and low lateral mobility for senior positions. Our model incorporates a tournament structure with backloaded incentive payments, inspired by Malcomson (1984), and monopsonistic competition with differentiated firms, à la Card et al. (2018). In an overlapping-generations setup, inexperienced workers initially choose a firm and an effort level, and subsequently decide whether to remain with the firm. Workers have idiosyncratic preferences about non-wage amenities of firms and draw heterogeneous outside employment opportunities in each period, which leads to an increasing residual labor supply curve. Firms set both junior and senior wage levels and choose promotion rates based on noisy signals of worker effort, according to an 'up-or-out' policy.

Solving the model in steady state reveals two novel results on wage markdowns under dynamic incentive contracts, which can be tested empirically. First, we find that if effort is endogenous, firms pay junior employees below their marginal revenue product of labor, but senior employees above their marginal revenue product, mirroring the results in prior

[1] A separate class of 'dynamic monopsony' search models allow for labor supply and labor demand, but still implies wage markdowns rather than wage markups (Burdett and Mortensen, 1998; Manning, 2003; Jarosch et al., 2024). In the same vein, Balke and Lamadon (2022) allows for wage backloading under implicit contracts, but not for efficiency wages.

implicit contract models with backloaded pay (Lazear, 1981; Malcomson, 1984). In contrast, when effort is exogenous (or irrelevant), the monopsonist still marks down junior wages, but pays senior employees exactly the marginal revenue product of labor. This implies that the monopsonist only exerts monopsony at the career entry stage, and is in contrast to prior tournament models, which imposed perfect labor market competition.

Our second main result is that the extent to which firms exert monopsony power is characterized by a markdown of the net-present value of wages below the net-present value of the marginal revenue product of labor, evaluated at the start of the career. We show that this 'lifetime' wage markdown is a function of the residual *long-run* labor supply elasticity to the firm when hiring junior workers. This provides a dynamic analogue to the static labor supply model of Card et al. (2018), in which wage markdowns are a function of short-run labor supply in every period.

We test these two empirical hypotheses— whether firms pay backloaded efficiency wages and whether firms exercise monopsony power —in the context of the U.S. public accounting industry. We obtain novel and uniquely rich internal data on 569 U.S. public accounting firms. Our empirical strategy focuses on estimating wage markdowns for different seniority levels of workers while imposing a minimal set of assumptions on the labor supply and demand side. We specify and estimate a multi-output production model that allows for output differentiation by worker seniority, and for team production in the spirit of Alchian and Demsetz (1972). The team production feature is important because it allows to incorporate, for instance, the effects of partner's effort in attracting new clients on billing revenue generated by more junior employees, or to allow for mentoring externalities within the firm. We estimate our production model using detailed data on worker type-specific billing hours and billing rates, worked hours, and hourly compensation.

Using our estimated team production model, we compute marginal revenue products of labor for each worker type, and use these to infer 'instantaneous' wage markdowns/markups for workers at different stages of their careers. Our first main finding is clear evidence of backloaded efficiency wages. Junior employees are paid only slightly more than half of their marginal revenue product, implying substantial wage markdowns, while workers with more than ten years of experience earn wage mark*ups*. This indicates the presence of backloaded incentive pay contracts.[2][3]

Our second main finding is that the net present value (NPV) of wages at career entry

[2] Although some of the prior production function-based markdown estimates have also found wage markups (Dobbelaere and Mairesse, 2013; Treuren, 2022), these have been typically attributed to collective bargaining, and have been documented mainly for European countries with collective bargaining institutions. In contrast, we focus on a U.S. industry without unionized workers.

[3] Our estimates also reject static monopsony models, which cannot rationalize wage markups.

is, on average, 27% below the NPV of the marginal revenue product of labor, implying a statistically significant lifetime wage markdown. This result provides direct evidence of monopsony power, in contrast to classical models of backloaded pay under perfect labor market competition, in which the NPV of wages equals the NPV of the MRPL.[4] Using the model, we convert these lifetime markdowns into short- and long-run residual labor supply elasticities. We find that residual labor supply is highly inelastic in the short run, with an average elasticity of 0.22, reflecting that workers choose firms primarily based on the prospect of very high future wages rather than the much lower entry-level pay. In the long run, labor supply is substantially more elastic, with an average elasticity of 2.7. Yet, because long-run labor supply is not perfectly elastic, firms continue to exert monopsony power in the junior labor market.

Our empirical findings carry multiple important implications. First, estimated wage markdowns for employees can substantially overestimate the true extent of monopsony power in the presence of implicit contracts, as there is no longer a relationship between the static labor supply elasticity and wage markdowns. Second, not all markdown dispersion is necessarily inefficient, as wage markdowns in part are designed to induce effort provision by junior employees, in contrast to spot market monopsony models, in which wage markdown variation necessarily leads to misallocation (Berger et al., 2022). Third, there is an important distinction between ex-ante and ex-post markdown inequality and, therefore, wage inequality. Ex-post, there is substantial income inequality between workers who left the firm prematurely before earning wage markups and the employees that win the tournament and become senior employees or partners. However, our model posits that workers are homogeneous ex-ante, and face the same wage markdowns in expectation at the start of their careers.

This paper relates to four main strands of literature. First, we build on monopsony models of horizontally differentiated employers in the tradition of Card et al. (2018), in which markdowns are the result of preference heterogeneity of workers (Berger et al., 2022; Azar et al., 2022; Lamadon et al., 2022; Roussille and Scuderi, 2021; Volpe, 2024). In these static models of labor markets, wage markdowns are proportional to short-run labor supply elasticities. Extensions of this framework that incorporate dynamics have maintained the assumption of exogenous worker effort, and introduced dynamics through firm-specific human capital accumulation, rather than efficiency wages (Jungerman, 2023; Berger et al., 2024; Agostinelli et al., 2025). [5] We contribute to this literature by incorporating endoge-

[4]This equivalence holds under additional assumptions, such as risk neutrality and a common discount factor for firms and workers.

[5]Different strands of 'dynamic' monopsony models exist without with search frictions also features wage markdowns in equilibrium (Burdett and Mortensen, 1998; Manning, 1987; Gottfries and Jarosch, 2023). Also

nous effort of forward-looking workers into the workhorse monopsony model. We derive a relationship between long-run labor supply elasticities and the markdown of the net present value of wages below MRPL which is a dynamic analog of the static markdown rule of Card et al. (2018). Another difference with the monopsony literature is that it has focused predominantly on lower-wage occupations and industries (Card and Krueger, 1994), whereas we study the exertion of monopsony power in a high-wage industry, public accounting, which shares many features with other professional services industries, such as law firms and consulting firms.

Second, we contribute to the literature on implicit labor contracts with endogenous effort, much of which features backloaded pay as incentive wages (Lazear, 1981; Malcomson, 1984; Medoff and Abraham, 1980). We contribute to this literature by incorporating monopsony power in the tournament model, thereby allowing for the NPV of wages to be marked down below the NPV of the marginal revenue product of labor. We also allow for nonlinear production with various types of workers, which is crucial when bringing the model to the data, and incorporate multi-output production with productivity externalities across the various labor inputs into the tournament model.

Third, we contribute to the prior literature on estimating wage markdowns using production models (Brooks et al., 2021; Yeh et al., 2022; Morlacco, 2019; Mertens, 2022; Morlacco, 2019; Delabastita and Rubens, 2025; Treuren, 2022), which has mainly focused on manufacturing industries, whereas we estimate a multi-output team production model of professional services firms. Our work builds on prior approaches that have used production function estimation to test the backloaded wage hypothesis (Kotlikoff and Gokhale, 1992; Hellerstein et al., 1999), with the crucial difference that we specify and estimate a multi-output team production model in which we allow for team production and in which output is differentiated between employees of different seniority levels. Our approach to estimating the production model consists of estimating a matrix of own- and cross-output elasticities for all types of workers, rather than just the own-output elasticities as is usual in single-output production models. Our model relates to the literature on multi-product production models (De Loecker et al., 2016; Dhyne et al., 2017; Orr, 2022; Valmari, 2023), in which we follow an approach similar to (Dhyne et al., 2017) by allowing inputs to be nonrival within the firm, thereby avoiding having to assign input usage to individual product lines, an approach which fits well with the nature of production in professional services firms.

Finally, this paper relates to prior work on competition and market power in the market for auditing and accounting services (Gerakos and Syverson, 2015). Whereas this literature

related are models of implicit labor contracts such as Lamadon et al. (2022).

has mainly focused on market power of auditing firms on the product market, we highlight and quantify their market power on labor markets.

The remainder of this paper is structured as follows. We start by describing the public accounting industry, our dataset, and key facts on billing-rate-to-salary spreads by worker seniority and promotion rates in Section 2. In Section 3, we set up a tournament model with monopsonistic competition and nonlinear multi-output production. We use this model to derive two empirical hypotheses for the exertion of monopsony power and for efficiency-wage contracts. In Section 4, we test these hypotheses for the U.S. public accounting industry by estimating wage markdowns using our production model, and turn these markdown estimates into short- and long-run labor supply elasticities. Section 5 concludes.

2 Industry Background and Data

2.1 U.S. Public Accounting Firms

We study public accounting firms, which provide auditing, tax preparation, and consulting services to clients. Although the industry is dominated by the "Big Four" (EY, Deloitte, PwC, KPMG), it contains many smaller firms that together represent non-negligible market shares. For instance, SEC filings show that the Big Four together accounted for 48.4% of public firm audits in 2024 (Ideagen Audit Analytics, 2024). As explained in Section 2.2, our analysis focuses on non–Big Four firms, for data availability reasons. Accordingly, unless otherwise noted, our description of the public accounting industry below refers to this group.[6] Revenue of public accounting firms is generated by auditing (29%), tax compliance services (40%), and non-compliance advisory services (31%), which include other tax services, business advisory, valuation, and IT consulting (Inside Public Accounting, 2023b).

Public accounting firms are typically organized as partnerships, with 'equity partners' being the residual claimants of profits. Firms employ 'professional staff' members, who mostly hold Certified Public Accountant (CPA) licenses, and which are promoted in an 'up-or-out' career path that can end in a promotion to partnership. Given the up-or-out system, turnover of professional staff is high, at an annual 15% rate. The select few who make it to the partnership usually have around 15 years of experience before reaching that stage. Lateral mobility is limited, with merely 12% of new equity partners coming in from another firm (Inside Public Accounting, 2023b). In addition to the professional staff, public accounting firms employ 'paraprofessionals'—who are not in the partnership track—as well as interns and administrative staff.

[6] For ease of exposition, we omit the "non–Big Four" qualifier in the remainder of the paper.

Two thirds of firms set prices using hourly billing rates, charging clients staff-specific hourly billing rates multiplied by each employee's billable hours. One quarter rely instead on fixed or upfront fees, while the remaining 10% use other fee arrangements. Employees are paid salaries based on their actual working hours, which are on average about one-third above their billable hours (Inside Public Accounting, 2023b).

Partners are divided into non-equity partners, who hold the partnership title but do not own a share of the firm, and equity partners (around two thirds of all partners), who earn on average around twice as much as non-equity partners. Most firms impose a mandatory retirement age for partners (Inside Public Accounting, 2023b).

2.2 Data

Our dataset consists of annual surveys of 501-588 U.S. public accounting firms conducted by Inside Public Accounting from 2015 until 2024. The survey does not include the Big Four companies, so our focus lies on small and medium-sized public accounting firms. For confidentiality reasons, we do not observe firm-level data; instead, we observe averages for bins of firms that contain on average 20 firms, divided based on location and firm size, with the smallest bins containing 5 firms. We observe 25 firm bins per year over 10 years, yielding a total of 250 observations.

We observe averages of firm-level net revenues and billing revenues, disaggregated into seven employee categories: professional staff with 0-2, 3-5, 6-8, and 9+ years of experience, non-equity partners, equity partners, and paraprofessionals. For each of these categories, we observe billing rates (USD per billed hour), billed hours, hourly salaries (USD per worked hour), and hours worked. The only exception is that hours worked and billed hours are not reported for non-equity partners, although their compensation and billing rates are available. We deflate all monetary variables using the CPI for urban consumers, averaged over U.S. cities.[7].

In addition to the professional staff and paraprofessionals, we observe the number of interns and administrative staff employed. In terms of non-labor expenses, we observe expenditure on marketing, training and continued professional education, recruiting, and technology. We also observe the value of the firm's working capital.

Table 1 reports selected summary statistics. Annual revenue ranges from USD 1M to USD 4B, with an average of USD 58M. Firms employ on average 239 employees, among which 22 are equity partners and 55% of whom are CPA-licensed. Firms spend 48% of revenue on labor, 4.5% on technology, 2% on marketing, and 1% each on training and recruiting costs. Firm capital is on average around 20% of annual revenue.

[7]https://fred.stlouisfed.org/series/CPIAUCSL

Table 1: Summary Statistics

	Avg.	Med.	S.D.
Revenue (100 millions USD)	0,695	0,188	1,271
No. Employees (1000s)	0,239	0,080	0,397
Interns (Share of Employment)	0,017	0,016	0,007
Paraprofessionals (Share of Employment)	0,053	0,050	0,024
Admin staff (Share of Employment)	0,162	0,162	0,019
Professionals 0-2 Years (Share of Employment)	0,152	0,150	0,025
Professionals 3-5 Years (Share of Employment)	0,150	0,147	0,026
Professionals 6-8 Years (Share of Employment)	0,104	0,101	0,021
Professionals 9+ Years (Share of Employment)	0,171	0,174	0,031
Non-Equity Partners (Share of Employment)	0,046	0,047	0,018
Equity Partners (Share of Employment)	0,114	0,111	0,026
Labor Expenditure (Share of Revenue)	0,482	0,482	0,028
Technology Expenditure (Share of Revenue)	0,045	0,045	0,008
Marketing Expenditure (Share of Revenue)	0,019	0,019	0,005
Training Expenditure (Share of Revenue)	0,010	0,010	0,002
Recruiting Expenditure (Share of Revenue)	0,009	0,007	0,044
Working Capital (Share of Revenue)	0,193	0,190	0,045
Share of Staff holding CPA	0,548	0,543	0,067
No. Firms per Bin	20,446	17,000	11,638
No. of Bin-Years		250	

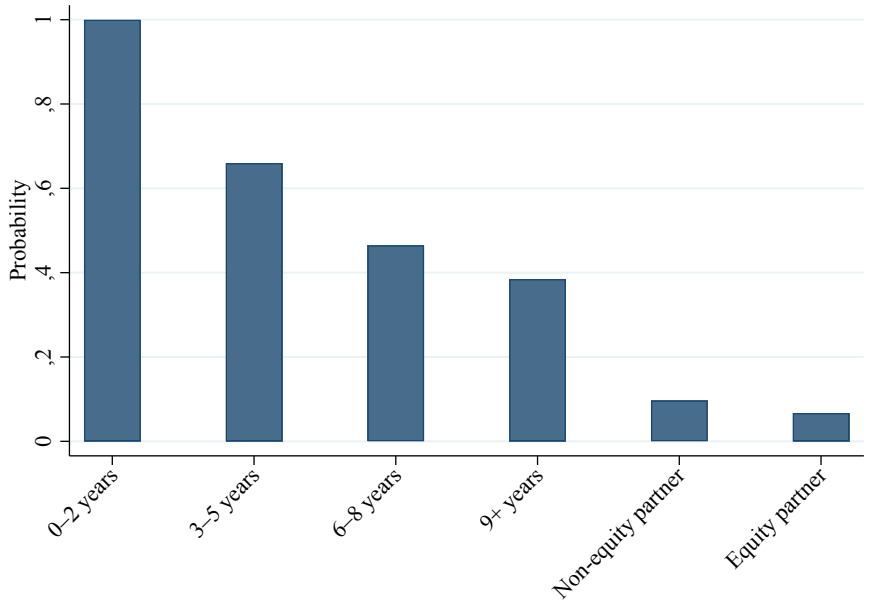
2.3 Stylized Facts

Fact 1: Firms use an 'up-or-out' promotion policy

Public accounting firms use an 'up-or-out' promotion policy, similarly to many other professional services industries. We compute promotion probabilities by assuming a steady-state labor force at the industry level, while allowing for different promotion probabilities for small, medium-sized, and large firms.[8] The resulting promotion probabilities from the viewpoint of a career starter without experience are shown in Figure 1. 35% of starters

[8]Details on the computation of promotion probabilities are in Appendix B.2.

Figure 1: Cumulative Promotion Probabilities



Notes: Bars indicate the cumulative probabilities of reaching every grade i. Details on computation are in Appendix B.2.

exit after 2 years, another 20% exits after 5 years, etc. The ex-ante probability of becoming a non-equity partner is 10%, whereas merely 7% of starters eventually become equity partners.

Labor markets are mostly internal, with very low lateral mobility between public accounting firms (Inside Public Accounting, 2023a). Workers that are not promoted to partnerships leave the firm, either voluntarily or involuntarily,[9] and usually transition to in-house public accounting positions in firms, rather than partnerships at competing public accounting firms. IPA reports that merely 14% of new equity partners are, on average, lateral hires rather than in-house promotions (Inside Public Accounting, 2023b). We will assume away any lateral mobility in the model, to keep it tractable.

Fact 2: Spreads between billing rates and compensation decrease with seniority

We compute the spread between the hourly salary (compensation to the worker per billable hour) and the billing rate (price charged to the client per billable hour) for employees of

[9]The 2024 IPA Practice Management Report mentions that around 4/5 of turnover between grades is voluntary, with the remainder being involuntary (firings) (Inside Public Accounting, 2023b).

seven different seniority levels i:[10]

$$\text{spread}_i = \frac{\text{compensation}_i}{\text{billing revenue}_i}$$

This spread amounts to measuring the labor share of revenue by employee seniority level. If one would assume independent and linear production by each worker i and no product market power, this margin would be equal to the wage markdown (or wage markup) $\delta_i \equiv W_i/MRPL_i$.[11]

Figure 2 shows the salary-billing rate spreads for the various levels of employees recorded in the dataset. Upon joining the firm, professional staff in the 'partner-track' are paid on average 30% of the amount they are billed to the clients. As seniority grows, this compensation rate slightly increases, up to 37% for staff with 9+ years of experience. Non-equity partners are paid 62% of the revenue that they generate. In contrast, equity partners receive 38% more than the billing revenue they generate. If equity partners are the only residual claimants of profits, this is mechanical: summing the various margins across labor types (weighted by the numbers of employees in each category) and overhead costs per revenue would have to add up to unity in order to ensure steady-state firm profits. Finally, paraprofessionals, which are outside the 'partner promotion track' are paid on average 38% of their billing revenue.

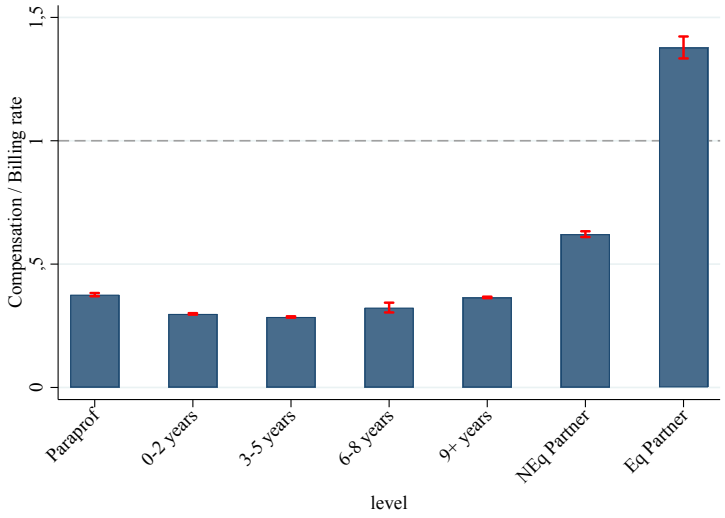
Taken together, the promotion rates, lack of lateral turnover, and compensation-billing spreads suggest that a tournament takes place in public accounting firms, which implies that dynamics should be important for understanding labor supply decisions: forward-looking junior staff members should internalize that despite earning well below their billable revenues initially, there is a probability of earning substantially above billing revenues later upon becoming an equity partner.

However, these billing-earning spreads cannot be readily interpreted as markdowns because in reality, production by different accountants does not happen independently, the production function might be nonlinear, and firms likely have some product market power. For these reasons, we need a structural model to estimate wage markdowns and wage markups for the various employees and partners within the firm, which we construct and estimate in Section 4.

[10] Given that we do not observe worked and billed hours for non-equity partners, we assume that their hours are identical to those performed by equity partners.

[11] Under this assumption, production for type-i services would be $Q_i = a_i L_i$, implying $(W_i/MRPL_i) = (W_i L_i)/(P_i Q_i)$, where W_i denotes hourly wages, P_i billing rates, L_i hours worked, and Q_i billable hours.

Figure 2: Compensation-Billing Spreads and Seniority



Notes: Average ratios of hourly compensation over hourly billing rate, by seniority. Confidence intervals based on standard errors clustered at the firm-group level.

Roadmap

The wage markdown patterns and promotion schedule documented above suggest a tournament structure occurs in the public accounting industry. However, wage markdowns for employees could also be due to the exertion of monopsony power, so unlike the prior tournament literature, we do not want to impose perfect labor market competition by assumption. In this section, we set up a model of monopsonistic competition with backloaded compensation that nests both monopsony models with exogenous worker effort and tournament-style efficiency wage models. We start with a general model of production and labor supply, and impose more functional form assumptions when bringing the model to the data in the next section.

3 Model: Monopsony and Backloaded Compensation

We develop a model in the spirit of the rank-order tournament framework of Lazear and Rosen (1981) and Malcomson (1984), extending it in two important dimensions. First, we allow for nonlinear production and multiple differentiated outputs, in contrast to the linear technologies typically assumed in the tournament literature. Second, we incorporate monopsonistic competition.

The model is set in an overlapping-generations environment. For expositional purposes, we focus on two worker levels—juniors and seniors—while Appendix C presents a more general version with N levels. Workers live and work for two periods: they enter

the labor market as juniors and, if promoted, become seniors before retiring.

To further simplify the exposition, we assume that both junior and senior workers are employees, and that wages and promotion policies are chosen by a residual claimant who does not participate in the tournament. Section 3.5.1 discusses how to interpret the model when senior workers are themselves the residual claimants of profits, as is the case in most (though not all) public accounting firms.

3.1 Model Primitives

3.1.1 Environment and Timing

Time is discrete and indexed by t. All agents share the same discount rate, $\rho < 1$. Each period, a mass $\mathcal{L}$ of new junior workers enters the labor market. A junior hired in period t becomes a senior in $t+1$ and retires at the beginning of $t+2$.

Each worker i of seniority type $s \in \{1, 2\}$ (junior or senior) employed in period t exerts effort $e_{i,t} \in [\underline{e}, \infty)$, incurring cost $C(e_{i,t})$ with $C'(\cdot) > 0$ and $C''(\cdot) > 0$. Worker utility increases linearly in wages, parametrized by the coefficient α, and decreases in effort.

$$U_{i,t} = \alpha W_{s(i),t} - C(e_{i,t}) \tag{1}$$

The firm observes a noisy, non-contractible signal of effort,

$$\theta_{i,t} = e_{i,t} + \kappa_{i,t}, \qquad \kappa_{i,t} \sim \Phi(\cdot),$$

with density $\phi(\cdot)$ and $\kappa_{i,t}$ independent across i and t.

3.1.2 Production and Compensation Policy

The firm employs both juniors and seniors. Let $L_{1,t}$ and $L_{2,t}$ denote their respective employment levels, and let $\bar{e}_{s,t}$ be the average effort of type-s workers. The firm produces two outputs,

$$Q_{s,t} = F_s(L_{s,t}, L_{-s,t}, \bar{e}_{s,t}, \bar{e}_{-s,t}), \quad s \in \{1, 2\}$$

where $L_{-s,t}$ and $\bar{e}_{-s,t}$ denote the employment and effort of the other type. In our empirical application, $Q_{s,t}$ corresponds to the billing hours produced by workers of seniority level $s \in \{1, 2\}$. Effort weakly increases production; specifically, $\partial F_s / \partial \bar{e}_{k,t} \geq 0$ for all $s, k \in \{1, 2\}$.

At the time of hiring junior workers, the firm commits to a compensation–promotion policy, $\{W_{1,t}, W_{2,t}, \tau_t\}$, consisting of:

- $W_{1,t}$: wage for juniors in period t;

- $W_{2,t}$: wage for seniors (promoted juniors) in period $t+1$
- τ_t: promotion cutoff, defined as the share of juniors dismissed; i.e., $\theta_{i,t} \geq \theta_t^P(\tau_t)$ implies promotion, where $\theta_t^P(\tau_t)$ denotes the τ_t-quantile of effort signals among junior workers hired by the firm in period t.

We assume that compensation of employees is set by the firm, rather than negotiated.[12] We further assume away a lateral labor market within the industry. In particular, firms cannot hire senior workers directly; instead, workers must first be hired as juniors and then attain seniority through promotion in the following period.

3.1.3 *Workers' Outside Options*

Each worker draws an outside option $U_{i,s}^{alt}$ from G_s^{alt} with density g_s^{alt}. These draws are i.i.d. across workers and independent over time. Juniors observe $U_{i,1}^{alt}$ before deciding whether to accept employment and how much effort to exert, but they do not know $U_{i,2}^{alt}$ in advance. A worker hired in period t and subsequently promoted observes $U_{i,2}^{alt}$ at the beginning of period $t+1$, before deciding whether to remain with the firm.[13]

3.1.4 *Product Market*

The firm is a monopolistic competitor in the product market. The demand for accounting services of type s is loglinear with elasticity $\eta < -1$. That is, denoting by $P_{s,t}$ the price of service s and by $\overline{Q_t}$ and $\overline{P_t}$ the market-level quantities and prices, we have

$$Q_{s,t} = \overline{Q}_t \left(\frac{P_{s,t}}{\overline{P}_t} \right)^{\eta}. \tag{2}$$

Then, denoting by $P_s(Q_{s,t})$ the price of service s as a function of the firm's output, we can write the firm's revenue from s in period t as

$$\begin{aligned} R_s(L_{s,t}, L_{-s,t}, \bar{e}_{s,t}, \bar{e}_{-s,t}) &\equiv P_s\left(F(L_{s,t}, L_{-s,t}, \bar{e}_{s,t}, \bar{e}_{-s,t})\right) F(L_{s,t}, L_{-s,t}, \bar{e}_{s,t}, \bar{e}_{-s,t}) \\ &= d_t \left[F(L_{s,t}, L_{-s,t}, \bar{e}_{s,t}, \bar{e}_{-s,t})\right]^{\frac{1+\eta}{\eta}} \qquad s \in \{1,2\}, \end{aligned}$$

[12] Wage bargaining could provide an additional explanation for why wage markdowns evolve with experience, and why wage markups arise, as discussed in Dobbelaere and Mairesse (2013); Treuren (2022). However, unionization rates are near zero for U.S. public accounting firms, as in much of private-sector employment in the U.S. Although individual wage bargaining may occur, especially at the senior level, such effects can to some extent be captured by modeling senior workers as residual profit claimants, as we discuss in Section 3.5.1.

[13] Consistent with our assumption of no lateral market, the outside option for senior workers reflects job opportunities outside of public accounting firms.

where $d_t \equiv \overline{P}_t / \left[\overline{Q}_t^{1/\eta}\right]$. This expression makes it explicit that prices adjust to input choices via the implied change in output. For notation simplicity—and consistent with our focus on the steady-state equilibrium of the model, as explained in Section 3.3—we suppress the time index t from d_t in the remainder of the analysis.

3.2 Worker and Firm Behavior

3.2.1 Workers' Problem

Senior workers are not incentivized by the contract, since no further promotion is available. Accordingly, they exert an exogenous level of effort, denoted by $e^*_{2,t}$, and their value is[14]

$$U_{2,t} = \alpha W_{2,t} - C(e^*_{2,t}) \tag{3}$$

Juniors choose effort to maximize expected discounted utility:

$$U_{1,t} = \max_e \ \alpha W_{1,t} - C(e) + \rho \left\{ \left[1 - \Phi(\theta_t^P(\tau_t) - e)\right] \bar{\bar{U}}_{2,t} + \Phi(\theta_t^P(\tau_t) - e) \bar{U}_2^{alt} \right\} \tag{4}$$

where

$$\bar{\bar{U}}_{2,t} \equiv \mathbb{E}\left[\max\left\{U_{2,t}, U_{i,2}^{alt}\right\}\right] = G_2^{alt}(U_{2,t}) U_{2,t} + \int_{U_{2,t}}^{\infty} u \, dG_2^{alt}(u)$$

and $\bar{U}_2^{alt}$ is the expected value of outside options for seniors. The term $\bar{\bar{U}}2, t$ is the expected continuation value of a worker who is promoted in period t, prior to the realization of the period-$t+1$ outside offer. A worker exerting effort e is promoted with probability $1-\Phi(\theta_t^P(\tau_t)-e)$ and dismissed with the complementary probability. Accordingly, the term in braces in ((4)) represents the worker's *ex-ante* continuation value—before the realization of effort signals and the determination of promotion outcomes—as a probability-weighted average of $\bar{\bar{U}}_{2,t}$ and $\bar{U}_2^{alt}$.

The effort choice of junior workers, denoted by $\tilde{e}_t$, must satisfy

$$C'(\tilde{e}_t) = \rho \phi(\kappa^P(\tau_t))[\bar{\bar{U}}_{2,t} - \bar{U}_2^{alt}] \tag{5}$$

where $\kappa^P(\tau_t)$ denotes the τ_t-quantile of the distribution Φ.

The following lemma follows directly from (5) and the convexity of $C(\cdot)$:

[14] If senior workers were regular employees of the firm, they would exert only the minimum effort level $\underline{e}$. Under the interpretation put forth in Section 3.5.1—that senior workers are the residual claimants—there are incentives to set $e^*_{2,t} > \underline{e}$. In any case, the choice of $e^*_{2,t}$ is exogenous at the time the firm determines the compensation–promotion policy $\{W_{1,t}, W_{2,t}, \tau_t\}$.

Lemma 1. *Assume that $\tau_t \in (0,1)$. Then junior workers' effort choice $\tilde{e}_t$ is strictly increasing in senior compensation $W_{2,t}$.*

In other words, the prospect of higher senior compensation induces junior workers to exert greater effort in order to raise their probability of promotion. For this result to hold, we need $\tau_t \in (0,1)$—that is, the firm must commit neither to dismiss all junior employees nor to promote all of them.

3.2.2 *Employment Dynamics*

The number of juniors willing to work for the firm is

$$L_{1,t} = G_1^{alt}(U_{1,t})\,\mathcal{L} \tag{6}$$

The number of seniors in $t+1$ is

$$L_{2,t+1} = (1-\tau_t) G_2^{alt}(U_{2,t})\, L_{1,t} \tag{7}$$

3.2.3 *Firm's Problem*

Given incumbent seniors $l_{2,t}$, the firm chooses $\{W_{1,t}, W_{2,t}, \tau_t\}$ to maximize the discounted value:[15]

$$\begin{aligned} V(l_{2,t}) = \max_{W_{1,t}, W_{2,t}\tau} \Big\{ & R_1(L_{1,t}, l_{2,t}, \tilde{e}_t, e^*_{2,t}) + R_2(l_{2,t}, L_{1,t}, e^*_{2,t}, \tilde{e}_t) \\ & - W_{1,t}L_{1,t} - \rho W_{2,t}(1-\tau_t) G_2^{alt}(U_{2,t}) L_{1,t} + \rho V\left[(1-\tau_t) G_2^{alt}(U_{2,t}) L_{1,t}\right] \Big\} \end{aligned} \tag{8}$$

Subject to the incentive-compatibility constraint (5). Under constant-elasticity demand in the product market, the marginal revenue associated with service s in period t, in equilibrium, is

$$MR_{s,t} = \frac{1+\eta}{\eta} d\left[F_s(L_{s,t}, L_{-s,t}, \bar{e}_{s,t}, \bar{e}_{-s,t})\right]^{\frac{1}{\eta}}, \quad s \in \{1,2\} \tag{9}$$

Then, the envelope condition from (8) gives the marginal value of seniors:

$$V'(l_{2,t}) = \frac{1+\eta}{\eta} d \left\{ \left[F_1(L_{1,t}, l_{2,t}, \tilde{e}_t, e^*_{2,t})\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_{2,t}} + \left[F_2(L_{1,t}, l_{2,t}, e^*_{2,t}, \tilde{e}_t)\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_{2,t}} \right\} \tag{10}$$

[15] Note that the compensation-promotion policy set in period t fully determines the number of type-1 workers employed by the firm in t, $L_{1,t}$; and the number of type-2 workers employed in $t+1$, $L_{2,t+1}$. It is thus without loss to write off the present discounted value of the entire wage bill—for periods t and $t+1$—of workers hired in period t when defining the firm's value in period t.

3.3 Testable Implications

We use our model to derive two testable implications linking wages to marginal revenue products. One concerns the compensation of senior workers, while the other concerns the lifecycle wedge between the present discounted value of wages and marginal products for a new hire. In Section 4, we verify these implications empirically.

3.3.1 Steady-State Equilibrium

To simplify the analysis, we focus on the model's steady-state equilibrium, in which all variables are stationary:

$$W_{1,t} = W_1, \quad W_{2,t} = W_2, \quad \tau_t = \tau, \quad L_{1,t} = L_1, \quad L_{2,t} = L_2$$

In steady state, the workers' conditions reduce to

$$U_2 = \alpha W_2 - C(e_2^*) \tag{11}$$

$$U_1 = \alpha W_1 - C(\tilde{e}) + \rho\left[(1-\tau)\bar{\bar{U}}_2 + \tau \bar{U}_2^{alt}\right] \tag{12}$$

$$C'(\tilde{e}) = \rho\, \phi(\kappa^P(\tau))\, [\bar{\bar{U}}_2 - \bar{U}_2^{alt}] \tag{13}$$

and employment satisfies

$$L_1 = G_1^{alt}(U_1)\mathcal{L}, \qquad L_2 = (1-\tau)G_2^{alt}(U_2)L_1 \tag{14}$$

3.3.2 Present-Discounted Values of Wages and Marginal Products

Before proceeding, we introduce notation that will be useful in deriving the testable implications. First, let the steady-state marginal revenue products of junior and senior labor, respectively, be

$$\text{MRPL}_1 \equiv \frac{1+\eta}{\eta} d \left\{ \left[F_1(L_1, L_2, \tilde{e}, e_2^*)\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_1} + \left[F_2(L_1, L_2, e_2^*, \tilde{e})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_1} \right\} \tag{15}$$

$$\text{MRPL}_2 \equiv \frac{1+\eta}{\eta} d \left\{ \left[F_1(L_1, L_2, \tilde{e}, e_2^*)\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_2} + \left[F_2(L_1, L_2, \tilde{e}, e_2^*)\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_2} \right\} \tag{16}$$

Next, define the firm's expected net present values (NPVs) of wages and marginal

products for a new hire:

$$\text{NPV}_W = W_1 + \rho(1-\tau)G_2^{alt}(U_2)\, W_2 \tag{17}$$

$$\text{NPV}_{\text{MRPL}} = \text{MRPL}_1 + \rho(1-\tau)G_2^{alt}(U_2)\, \text{MRPL}_2 \tag{18}$$

3.3.3 *Test for Efficiency Wages*

We first derive a test for whether the firm uses efficiency wages to incentivize workers' effort. To that end, we compare the case in which junior workers' effort increases output, $\partial F_1(L_1, L_2, \bar{e}_1)/\partial \bar{e}_1 > 0$, to that in which junior effort is irrelevant for production, $\partial F_1(L_1, L_2, \bar{e}_1)/\partial \bar{e}_1 = 0$. We refer to these cases as the "Productive Effort" and "Irrelevant Effort" scenarios, respectively. When effort is productive, firms optimally use efficiency wages to provide incentives; when effort is irrelevant for production, firms do not optimize on this margin.

Proposition 1. *Assume that the solution to the firm's satisfies $\tau \in (0,1)$. Then, the relationship between senior workers' compensation and their marginal revenue product of labor differs across effort regimes as follows:*

- ***Irrelevant Effort:*** $W_2 = MRPL_2$
- ***Productive Effort:*** $W_2 > MRPL_2$

See Appendix A.2.1 for the proof. In other words, when effort is productive, the firm pays senior employees more than their marginal contribution in the second period in order to compensate them for effort exerted in the first period. Conversely, when effort is irrelevant for production, senior compensation equals their marginal revenue product. The assumption $\tau \in (0,1)$ is essential for providing incentives to junior workers in the Productive Effort case, as per Lemma 1; in our empirical application, we verify that this condition holds.

Intuition. The proposition delivers a sharp result. In the Irrelevant Effort scenario, senior workers face no instantaneous wage markdown and therefore receive $W_2 = MRPL_2$, even though the firm is a monopsonist. The reason is that, absent incentive considerations, the sole role of wages is to attract and retain workers, so the monopsony distortion is fully absorbed by the junior wage. This result echoes Ioannides and Pissarides (1985), who develop a model of lifetime wage dynamics that is closely related to ours under the Irrelevant Effort scenario.[16]

[16]The model in Ioannides and Pissarides (1985) does not incorporate an endogenous effort choice by the worker, so we cannot compare its implications to ours in the Productive Effort case.

For intuition on the result, consider the firm's optimal choice of W_1 and W_2 (for the first-order conditions to the firm's problem in steady state, see Section A.1). In setting W_1, the firm trades off the marginal benefit of raising junior compensation—namely, attracting additional junior workers—against the marginal cost of increasing the wage bill for inframarginal juniors. At the optimum, these two forces balance each other.

When choosing W_2, the firm accounts for the same two forces: a higher senior wage increases the expected payoff from joining the firm as a junior worker, but it also raises the wage bill for inframarginal promoted workers. In addition, W_2 affects retention: a higher senior wage induces more senior workers to stay, generating marginal profits equal to $MRPL_2 - W_2$.

In the Irrelevant Effort scenario, these are the only considerations. Moreover, because the firm and workers discount time identically, the recruitment and inframarginal wage-bill effects enter symmetrically in the first-order conditions for W_1 and W_2. As a result, once W_1 is optimally chosen, these two components cancel out in the determination of W_2, leaving only the retention margin. Put differently, the firm's monopsony power is reflected in the junior wage, while senior wages are pinned down by the marginal profitability of retaining promoted workers. The firm therefore sets the senior wage so that the marginal profit from retaining an additional senior worker is zero, implying $W_2 = MRPL_2$.

By contrast, when junior effort is productive, senior compensation also provides incentives ex ante: higher promised wages upon promotion increase the return to exerting effort in the junior period. The firm then optimally backloads compensation by raising senior wages above their marginal revenue product, so that $W_2 > MRPL_2$.

Empirical Implication. Proposition 1 lays the groundwork for a formal test of the existence of backloaded efficiency wage contracts. We empirically assess the following condition:

$$\begin{cases} H_0 : W_2 = MRPL_2 \\ H_1 : W_2 > MRPL_2 \end{cases} \tag{19}$$

Rejecting the null hypothesis indicates that effort is productive and that firms account for effort incentives when determining compensation.

3.3.4 Test for Monopsony Power

Our second test concerns the firm's exercise of monopsony power. We begin by defining the short- and long-run elasticities of junior labor supply.

The firm-specific residual elasticity of junior labor supply is

$$\psi_1 \equiv \frac{\partial \ln L_1}{\partial \ln W_1} = \frac{g_1^{alt}(U_1)\, W_1}{G_1^{alt}(U_1)} \alpha_1 \tag{20}$$

Similarly, define the *long-run* firm-specific residual labor-supply elasticity as

$$\psi_{LR} \equiv \frac{\partial \ln L_1}{\partial \ln \mathrm{NPV}_W} \tag{21}$$

The latter elasticity measures how employment of newly hired junior workers responds to the total discounted wage package they are offered. Applying the chain rule, we have

$$\psi_{LR} = \frac{\partial \ln L_1}{\partial \ln W_1} \cdot \frac{\partial \ln W_1}{\partial \ln \mathrm{NPV}_W} = \psi_1 \cdot \frac{\mathrm{NPV}_W}{W_1} \tag{22}$$

where the second equality uses $\partial \mathrm{NPV}_W / \partial W_1 = 1$ and the definition of ψ_1. Hence, the long-run elasticity ψ_{LR} scales the static elasticity ψ_1 by (the inverse of) the share of the current wage in the present value of total expected wage payments.

Proposition 2. *The firm's monopsony power is characterized by the long-run firm-specific residual elasticity of labor supply. This elasticity determines the lifecycle wage markdown:*

$$\frac{NPV_W}{NPV_{MRPL}} = \frac{\psi_{LR}}{1 + \psi_{LR}}$$

Appendix A.2.2 contains the proof. The result follows from the first-order condition for W_1 in the solution to the firm's probem. Proposition 2 shows that the firm faces an intertemporal wage–productivity wedge arising from the long-run elasticity of junior labor supply, ψ_{LR}. The firm's monopsony power is exercised at the entry margin, since hiring additional junior workers implies higher discounted wage obligations for senior workers.[17]

The equation in the proposition is the dynamic analogue of the familiar relationship between wage markdowns and labor supply elasticities in static monopsony models. When long-run residual labor supply becomes perfectly elastic, $\psi_{LR} \to \infty$, the firm pays the competitive lifetime wage, $\mathrm{NPV}_W = \mathrm{NPV}_{\mathrm{MRPL}}$, and the model becomes an extension of Malcomson (1984) with nonlinear production. Under a finite labor supply elasticity ψ_{LR}, the ratio $\mathrm{NPV}_W / \mathrm{NPV}_{\mathrm{MRPL}}$ captures the extent to which lifetime wages are marked down

[17] One might wonder how this equation can be consistent with a zero-profit condition in steady state when seniors are the residual claimants of profits. The discount factor plays a key role. In the constant-MRPL case, steady-state zero profits imply $W_1 + W_2 = MRPL_1 + MRPL_2$. However, $W_1 + \rho W_2 \neq MRPL_1 + \rho MRPL_2$. With wage backloading ($W_2 > W_1$), we have $W_1 + \rho W_2 < MRPL_1 + \rho MRPL_2$.

relative to the lifetime marginal revenue product of labor.

Proposition 2 implies that the residual long-run labor supply elasticity for juniors, ψ_{LR}, can be recovered from NPV_W and NPV_{MRPL}. While this result is, up to our knowledge, new, it generalizes the firm-level labor supply elasticity central to Manning (2003). Empirical work such as Bassier et al. (2022) and Datta (2024) estimate separate recruitment and separation elasticities. Our dynamic formulation unifies these margins into a single, model-consistent long-run elasticity that governs intertemporal wage markdowns in overlapping-generations monopsony.

Empirical Implication. Based on Proposition 2, our empirical test for monopsony power, under the null hypothesis of no monopsony power, is:

$$\begin{cases} H_0 : \text{NPV}_\text{W} = \text{NPV}_\text{MRPL} \\ H_1 : \text{NPV}_\text{W} < \text{NPV}_\text{MRPL} \end{cases} \tag{23}$$

Rejecting H_0 indicates that the firm exerts monopsony power in the design of its wage scheme.

3.4 Welfare Implications of Wage Markdown Variation

The empirical tests above are relevant because the welfare consequences of wage markdowns depend critically on the degree of monopsony power and on whether wages play an efficiency-enhancing incentive role. In a competitive labor market—i.e., when long-run labor supply is perfectly elastic—eliminating wage markdowns would generally be inefficient, as it would weaken the firm's ability to provide incentives and may therefore reduce output. By contrast, in the presence of monopsony power but absent efficiency wage considerations, reducing lifetime wage markdowns can increase output by mitigating deadweight loss, provided that firms are still able to sustain the incentive structure embedded in the wage contract.

3.5 Model Discussion

3.5.1 Interpretation of Senior Compensation

In partnerships—including public accounting firms—the residual claimants are the equity partners, so newly promoted seniors do not receive a wage in the literal sense. Instead, upon promotion in period t+1, a new partner's present-discounted income is

$$-\, b^{in}_{t+1} \;+\; \pi^{s}_{t+1} \;+\; \xi^{lump} \;+\; \rho\, b^{out}_{t+2}$$

where b^{in}_{t+1} is the partnership buy-in, π^s_{t+1} is the expected share of one-period profits, ξ^{lump} is a lump-sum transfer to equity partners independent of firm performance, and b^{out}_{t+2} is the buy-out received upon retirement in the following period. The components b^{in}_{t+1}, π^s_{t+1}, and b^{out}_{t+2} are tied to the firm's present-discounted value and the partner's equity stake, and satisfy the accounting identity

$$b^{in}_{t+1} = \pi^s_{t+1} + \rho\, b^{out}_{t+2}$$

Therefore, the partner's present-discounted income simplifies to ξ^{lump}_{t+1}. In other words, the firm can implement any desired income for new partners by appropriately choosing the lump-sum transfer. This is what we denote by $W_{2,t}$ in the model.

After paying the buy-in, senior workers acquire ownership and control. Their incentives are therefore aligned with maximizing the firm's present-discounted value, since this increases both current profit shares and the future buy-out. Accordingly, we assume throughout that the firm maximizes the present-discounted value of profits.

3.5.2 Multiple Seniority Types

In Appendix C, we show that Proposition 1 extends to an N-period career. Under the Irrelevant Effort scenario, the instantaneous wage markdown is zero in every period except the first. By contrast, under the Productive Effort scenario, compensation in the second period exceeds the corresponding marginal revenue product of labor in present value terms.

Proposition 2 likewise generalizes to the N-period setting: the long-run labor supply elasticity governs the wedge between the present discounted values of wages and marginal products at the start of workers' careers.

Therefore, we can use (19) and (23) to test for efficiency wages and monopsony power in our empirical setting, which features more than two seniority levels.

3.5.3 Caveats and Limitations

Although we allow for the services of different employee types to be differentiated, we impose that employees of a given type s are homogeneous within firms, up to the effort term $e_{i,t}$. If employees of a certain type would be persistently different for reasons other than effort, promotion policies would be used not just as an incentive mechanism, but also as a means to sort high- from low-ability workers, which we do not allow for in our model.

4 Empirical Analysis

Our model in Section 3 shows that wage markdowns can be due to two different forces, backloaded efficiency-wages and monopsony power, and that both carry different efficiency consequences. To examine the existence and quantitative importance of both these forces, we estimate wage markdowns for public accountants using a 'production approach'. Doing so allows us to infer wage markdown patterns over the course of employee careers while maintaining a minimal set of assumptions on labor supply and demand.

4.1 Empirical Model

4.1.1 Demand for Accounting Services

Accounting firms f consists of 'billable' employees with seniority types $s \in \{1, ..., |S|\}$. Each level of employees produces $Q_{s,f}$ total billable hours that are billed to consumers at hourly rates $P_{s,f}$ and works $L_{s,f}$ hours for which they are compensated at an hourly salary rate $W_{s,f}$.[18] Consequently, each type s generates $R_{s,f} = P_{s,f}Q_{s,f}$ in revenues.

We denote the output vector produced by firm f as $\boldsymbol{Q}_f = \{Q_{s,f}\}$ and total firm revenue as $R_f = \sum_s R_{s,f}$. We model the firm as a multiproduct producer that sets individual line prices because the IPA reports reveal that two thirds of firms follow this pricing practice.[19]

We implement the loglinear demand curve for type-s accountants from (2) at the firm-level, adding an unobserved demand shifter $\zeta_{s,f}$ and, again, suppressing time notation:

$$Q_{s,f} = \overline{Q}_s \left(\frac{P_{s,f}}{\overline{P}_s} \right)^{\eta} \zeta_{s,f}. \tag{24}$$

We assume that the accounting services markets for non-big-4 firms are in monopolistic competition, so firms set billing rates at constant markups above accountant marginal costs $MC_{s,f}$, $\mu_{s,f} \equiv \frac{P_{s,f}}{MC_{s,f}} = \frac{\eta}{\eta+1}$.

4.1.2 Production of Accounting Services

The production of services by each level of employee type s (professional staff with 0-2, 3-5, 6-8, 9+ years of experience, paraprofessionals, partners) is given by the production model (25). The output of each type of employee is determined by hours worked of not just these employees, but also by all other types of employees, in a Cobb-Douglas function.

[18] We do not impose that wage contracts are written at the hourly level, W_s can be the result of a monthly or annual compensation contract as well.

[19] A quarter of firms instead charges a fixed fee for the bundle of accountants. In Appendix A.3, we extend the model to allow for bundle pricing, and discuss the implications for our estimates.

In addition, production involves a vector of $|O|$ other production inputs, denoted by $\boldsymbol{O}_f$, that do not directly generate billing revenue: interns, administrative staff, and intermediate inputs.

$$Q_{s,f} = \left(\prod_{k \in S} L_{kf}^{\beta^l_{s,k}} \right) \boldsymbol{O}_f{}^{\boldsymbol{\beta}^o_s} \Omega_{s,f}(\bar{e}_{s,f}, \bar{\boldsymbol{e}}_{-s,f}) \exp(\varepsilon_{s,f}). \tag{25}$$

Unfortunately, we do not observe fixed assets, such as office space rent or insurance expenses. In Appendix B.3.1, we impute fixed assets by imposing a zero-profit condition, and include them as an additional input in the production function. Doing so leads to very similar production estimates.

The (own-and-cross) output elasticities of the accountants are a matrix of coefficients $\boldsymbol{\beta}^l = \{\boldsymbol{\beta}^l_s\}$ with dimensions $|I|\text{x}|S|$, whereas the output elasticities of the other inputs are a matrix of coefficients $\boldsymbol{\beta}^o = \{\boldsymbol{\beta}^o_s\}$ with dimensions $|O|\text{x}|S$.

Seniority-specific productivity levels are denoted by $\Omega_{s,f}$. We decompose $\omega_{s,f} \equiv \ln(\Omega_{s,f})$ into an effort component $h_s(\bar{e}_{s,f}, \bar{\boldsymbol{e}}_{-s,f})$ and a component that is unrelated to effort, $\chi_{s,f}$.

$$\omega_{s,f} = \chi_{s,f} + h_s(\bar{e}_{s,f}, \bar{\boldsymbol{e}}_{-s,f})$$

By including effort of other worker types into $h_s(.)$, we allow effort of other worker types to affect the productivity of type-s workers. However, we impose that there are no productivity spillovers in terms of the non-effort productivity term $\chi_{s,f}$, although we allow this term to be correlated arbitrarily across worker types.[20] Finally, measurement error in output is denoted by $\exp(\varepsilon_{s,f})$.

Discussion Our production model is a multi-output production model with team production. These features are crucial for at least two reasons. First, our model allows for output produced by accountants of different seniority levels to be differentiated, whereas estimating a regular firm-level production model would require aggregating billing hours across seniority levels.[21] Second, our model allows for labor inputs to be non-rival: we do not require assigning the different services produced to inputs, similarly to, for instance, Dhyne et al. (2017).[22] For instance, our model allows partner hours and effort to increase billing revenue of different types of junior accountants through business development, which is a public good within the firm.

[20] For a model in which such productivity spillovers are allowed, we refer to Iyoha (2023).

[21] We compare our estimates with a more conventional firm-level revenue production function that aggregates output across employees in Appendix B.3.4.

[22] Doing so would require a production model with the other outputs, rather than just inputs, on the right-hand side, $Q_{s,f} = F(L_{s,f}, \bar{e}_{s,f}, L_{-s,f}, \bar{\boldsymbol{e}}_{-s,f}, Q_{-s,f})$ rather than $Q_{s,f} = F(L_{s,f}, \bar{e}_{s,f}, L_{-s,f}, \bar{\boldsymbol{e}}_{-s,f})$.

Caveats There are two important caveats to our production model. First, by letting hours worked by other employees—rather than their billable hours—enter the production function of each worker type, we assume away economies of scope and non-effort productivity spillovers across goods, in contrast to Dhyne et al. (2022). We think that this is a reasonable assumption.

Second, although we do allow the products of all employee types to be differentiated across seniority types within firms, we do not allow for differentiation across firms for each given type. For instance, Q billed hours of, say, equity partners are assumed to be the same for every firm.[23]

4.1.3 Wage Markdowns

The marginal revenue product of type-s labor, $\text{MRPL}_{s,f} \equiv \frac{\partial R_f}{\partial l_{s,f}}$, is given by:

$$\text{MRPL}_{s,f} = \frac{1}{L_{s,f}} \sum_k \beta^l_{s,k} R_{kf} \left(\frac{\eta + 1}{\eta} \right). \tag{26}$$

Each type of accountant contributes to revenue both by generating its own billing revenue and by contributing to the billing revenue of accountants of different seniority levels. The demand elasticities η_k also enter each marginal revenue product because changing an accountant's output also changes their billing rates.

Using the MRPL expression from Equation (26), we write the 'instantaneous markdown' $\delta_{s,f}$ at every seniority level as:

$$\delta_{s,f} = \frac{W_{s,f} L_{s,f}}{\sum_k \beta^l_{s,k} R_{kf} \left(\frac{\eta+1}{\eta} \right)}. \tag{27}$$

Hence, to recover these instantaneous markdowns, we need to observe level-specific billing revenues $r_{s,f}$ and earnings $W_{s,f} l_{s,f}$, estimate the output elasticity matrices $\boldsymbol{\beta}^l$ and $\boldsymbol{\beta}^v$, and infer the product demand elasticity η.

4.2 Production Function Estimation

4.2.1 Firm-Level Moment Conditions

We start by estimating the output elasticities, for which we need to observe billable hours $Q_{s,f}$, rather than just billing revenue $R_{s,f}$.Denoting lowercase variables as logs, we estimate

[23] In Appendix B.3.3, we relax this assumption by including a price control on the right-hand side of the production functions. We find that this results in very similar estimates.

$|S|$ production functions, one for each type of worker:

$$q_{s,f} = \sum_{k \in S} \left(\beta_{s,k} l_{s,kf}\right) + \beta_s^o \boldsymbol{o}_{s,f} + \chi_{s,f} + h_s(\bar{e}_{s,f}, \bar{e}_{-s,f}) + \varepsilon_{s,f}. \tag{28}$$

Denote the one-year lag of any variable X as $\hat{X}$. We impose an AR(1) assumption on the sum of non-effort productivity $\xi_{s,f}$ and measurement error $\varepsilon_{s,f}$:

$$\chi_{s,f} + \varepsilon_{s,f} = \sigma_s(\hat{\chi}_{s,f} + \hat{\varepsilon}_{s,f}) + \nu_{s,f} \tag{29}$$

We express the productivity shock as a function of data and the parameters $(\sigma_s, \boldsymbol{\beta}_s)$ to be estimated. Given that optimal effort is time-invariant in equilibrium, the effort term enters as an unobserved s-specific constant $c_s \equiv h_s(\bar{e}_{s,f}, \bar{e}_{-s,f})$.

$$\nu_{s,f} = q_{s,f} - \sigma_s \hat{q}_{s,f} - \left(\sum_k \beta_{s,k}^l l_{s,k,f} - \rho \sum_k \beta_{s,k}^l \hat{l}_{s,k,f} \right) - (\beta_s^o \boldsymbol{o}_f - \sigma_s \beta_s^o \hat{\boldsymbol{o}}_{s,f}) - (1 - \sigma_s) c_s$$

As is usual in the literature (Ackerberg et al., 2015), we separate the inputs into those that are flexibly chosen, which we denote $\boldsymbol{X}^{\text{flex}}$ and those that are predetermined, $\boldsymbol{X}^{\text{pre}}$. The flexible inputs are chosen prior to the arrival of the productivity shock $\nu_{s,f}$, whereas the predetermined inputs are chosen prior of observing this productivity shock.

We assume that all professionals of the different seniority levels i are dynamic inputs (facing adjustment costs), meaning that firms choose them before observing the transient productivity shocks ν_{st}, except for entry-level staff. This is consistent with the model in Section 3, in which junior employment and transition rates are chosen for each cohort, making senior employees predetermined.

In contrast, we assume that interns, paraprofessionals, administrative staff, intermediate inputs, and entry-level professional staff are variable inputs that can be flexibly adjusted. Consistent with these timing assumptions, we form the following moment conditions to estimate the parameters $(\rho_s, c_s, \boldsymbol{\beta}_s)$:

$$\begin{aligned} E(\nu_{s,f} | \boldsymbol{X}_f^{\text{pre}}, \hat{\boldsymbol{X}}_f^{\text{pre}}) &= 0 \\ \text{and} \quad E(\nu_{s,f} | \hat{\boldsymbol{X}}_f^{\text{flex}}) &= 0. \end{aligned} \tag{30}$$

4.2.2 *Estimation*

We measure outputs $Q_{s,f}$ as the product of billable hours and the number of employees of labor type i. For the 'billable-hour'-producing labor inputs $L_{s,f}$, we multiply each labor type's hours worked with the number of employees. Because we observe neither worked nor billed hours for non-equity partners, we assume their working and billing hours per worker are the same as those for equity partners. Thus, variation in the contribution of these two worker types stems only from their headcounts. Administrative staff and interns are measured in employee counts, as we do not observe hours for these employees. We define intermediate inputs as the sum of recurring training, recruiting, and technology expenses, which are reported in the data as shares of revenue. We compute the dollar value of intermediate input expenses by multiplying this share by total annual revenue, and include its logarithm in the production function.

Rather than observing the data at the firm f–level, we observe it at the level of bins of firms g, with each firm belonging to exactly one bin. In Appendix B.1, we show that we can still recover the production function parameters by estimating Equation (28) at the bin-year level, under two additional assumptions. First, we need to assume that input prices and residual input supply elasticities are identical within bin-year combinations. Second, the firm-level timing assumptions need to extend to the bin-level, which requires the additional assumption that exit and entry are orthogonal to the transient productivity shocks v.

We estimate the production function by specifying the moment conditions (30) at the bin-level. For the flexible inputs (interns, intermediate inputs, paraprofessionals), we include a single lag in the instruments vector. For the predetermined inputs (professional staff and partners), we include current log levels and a single lag into the instruments vector. We estimate the production function using one-step GMM and compute heteroskedacity-robust standard errors.

4.2.3 *Results: Production Estimates*

The production estimates are in Table 2. The diagonal elements report the own-output elasticities for each role. The own-output elasticity is the lowest for paraprofessionals and starting employees, and the highest for equity partners. This is intuitive: starters need much guidance so their billing hours are less driven by how much they work, as compared to more senior employees whose efforts independently generate billable output.

The off-diagonal elements measure spillovers in production. Two patterns stand out. First, equity partners have strong positive spillovers on all other employees. This makes

sense as partners attract clients, and provide general management and mentorship. The positive spillovers from equity partners play an important role in shaping the marginal revenue product of labor for partners, which demonstrates the importance of modelling multi-output production.

Second, the off-diagonal elements are mostly asymmetric, with more positive effects of seniors on juniors than vice-versa. For instance, entry-level employees receive a positive spillover effect from nearly all more senior employees (the first column), but impose negative spillovers on their senior colleagues (the first row). This could be due to mentoring relationships, which are beneficial for juniors but costly to seniors.

Turning to the non-partner-track staff, paraprofessionals seem to impose negative spillovers on the other employee types, which could be due to task duplication or congestion effects. Interns, materials, administrative staff, and capital all enter with coefficients that are close to zero, which indicates that they are not crucial to accounting services production, at least not at the intensive margin.

Given that our production model does not imply constant returns to scale, we can estimate employee-specific returns to scale. We find returns-to-scale parameters that are close to unity, indicating constant returns to scale, for all employee types.

Table 2: Production Estimates

	(I) 0-2 years		(II) 3-5 years		(III) 6-8 years		(IV) 9+ years		(V) Non-eq. Part.		(VI) Eq. Part.		(VII) Paraprof.	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
0-2 years	0,952	0,055	0,004	0,040	-0,035	0,061	-0,053	0,056	-0,070	0,085	-0,070	0,085	0,086	0,155
3-5 years	0,046	0,027	1,027	0,022	0,004	0,033	0,022	0,024	0,032	0,052	0,032	0,052	0,029	0,070
6-8 years	0,040	0,041	0,014	0,021	1,048	0,042	-0,041	0,033	0,057	0,052	0,057	0,052	0,057	0,117
9 years	-0,023	0,034	-0,025	0,017	-0,025	0,036	0,974	0,031	-0,018	0,042	-0,018	0,042	-0,058	0,085
Non-eq. Partners	0,012	0,019	0,009	0,012	0,007	0,017	-0,011	0,014	0,981	0,024	-0,019	0,024	-0,014	0,046
Eq. Partners	0,084	0,057	0,085	0,033	0,161	0,057	0,121	0,064	0,231	0,083	1,231	0,083	0,121	0,151
Paraprof.	-0,138	0,041	-0,073	0,025	-0,092	0,040	-0,036	0,045	-0,129	0,058	-0,129	0,058	0,747	0,095
Interns	0,013	0,029	0,007	0,022	-0,046	0,027	-0,009	0,021	-0,025	0,039	-0,025	0,039	0,027	0,077
Materials	0,009	0,015	-0,011	0,013	0,009	0,020	0,018	0,015	-0,031	0,024	-0,031	0,024	0,030	0,040
Admin. staff	-0,004	0,023	-0,044	0,017	-0,030	0,026	0,023	0,016	-0,040	0,032	-0,040	0,032	-0,015	0,079
Ser. corr.	0,806	0,067	0,710	0,107	0,816	0,083	0,951	0,058	0,813	0,068	0,813	0,068	0,796	0,079
Obs.	174		174		174		174						174	

4.2.4 Results: TFP vs. MRPL

We recover instantaneous wage markdowns from Equation (27), using the estimated output elasticities together with observed billing rates and compensation. Identifying the markdown requires knowing the product-market residual demand elasticity, η. Given the

CES demand system for all inputs, firms set a common markup across all accountant types (Feenstra and Ma, 2007; Orr, 2022). Rather than estimating η ourselves, we draw from Gerakos and Syverson (2015), who estimate residual demand elasticities for auditing services. We use their median own-price elasticity for non-big-4 auditing companies, $\eta = -2.389$ (implying a Lerner index of 42%), which we impose across all levels of accountant categories.[24] Given the instantaneous markdown estimates and observed wages, we can then back out the implied MRPL straightforwardly. We trim the estimated instantaneous markdowns at the 1st and 99th percentiles, to remove outliers.

In Figure 3, we plot the distributions of log productivity ω and of MRPL across employee types. Whereas partners rank the lowest in terms of productivity, they rank near the top of the employee types in terms of MRPL.[25] This contrast is again a feature of multi-output production: using traditional TFP measures, partners look unproductive because they work relatively many hours for little billable hour production. However, much of these worked hours generate billing revenue for more junior accountants, which shows up in the MRPL measure but not in the productivity residuals.

When comparing MRPL distributions, we find evidence for an 'MRPL ladder'. Paraprofessionals have by far the lowest MRPL. Among the 'partner-track' staff, starters have the lowest MRPL, which then gradually increases until 9+ years of service. This human capital accumulation leads to dynamic labor supply independently of backloaded incentive pay, as analyzed in Jungerman (2023).

Interestingly, equity partners have actually a slightly lower MRPL than non-equity partners and 9+year employees, and with a subtantial left tail, which is consistent with our model given that they are no longer incentivized through the tournament.[26]

4.3 Testing for Incentive Contracts and Monopsony Power

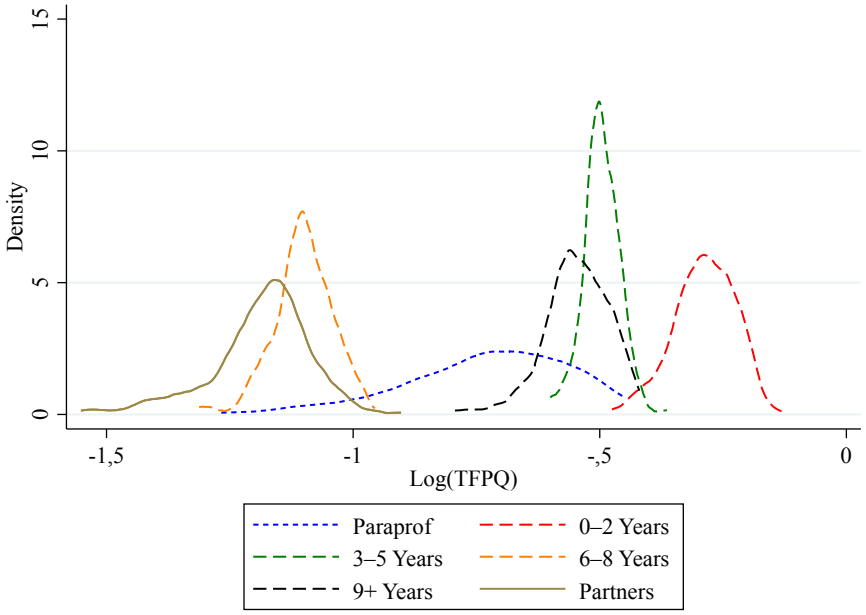
4.3.1 Backloaded Compensation Contracts

The estimated average instantaneous wage markdowns/markups are summarized in Figure 4. They are statistically different from one at conventional levels for all worker categories. Paraprofessional wages, which are outside the 'partner-track', earn on average 68% of their marginal revenue product. Turning to the 'career' professional staff, employees with 0-2 years of experience are paid only 55% of their MRPL, a markdown that remains relatively constant up to 8 years of seniority. That junior professional staff face a much

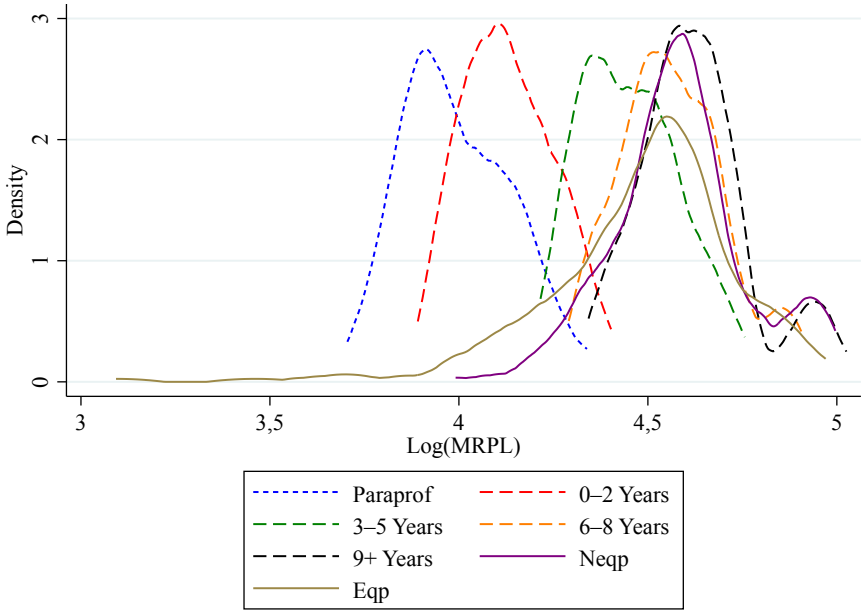
[24] Footnote on bundle vs individual demand elasticities here, again refer to Appendix A.3.
[25] This holds for both TFPQ and TFPR, as we show in Appendix Figure OA-5.
[26] Although partners are partly incentivized through profit-sharing, they only internalize their share of profits, which can lead to free-riding behavior.

Figure 3: Productivity vs. MRPL Dispersion

(a) log(TFP)

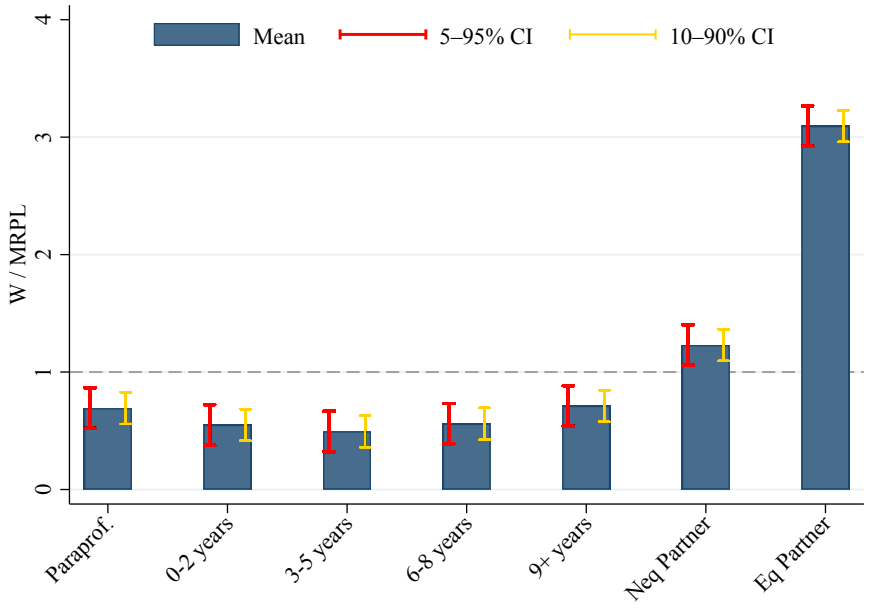


(b) log(MRPL)



Notes: Panel (a) shows the distributions of log(TFPQ) by employee type, panel (b) shows the distributions of log(MRPL). Both distributions are trimmed at the 2nd and 98th percentile.

Figure 4: Instantaneous Wage Markdowns and Seniority



Notes: Average instantaneous wage markdowns by seniority level. Block-bootstrapped confidence intervals with 250 iterations.

larger wage markdown compared to paraprofessionals is, once again, consistent with our model in Section 3: in contrast to paraprofessionals, junior professional staff find it acceptable to receive a relatively small compensation in the present in anticipation of future growth in earnings.

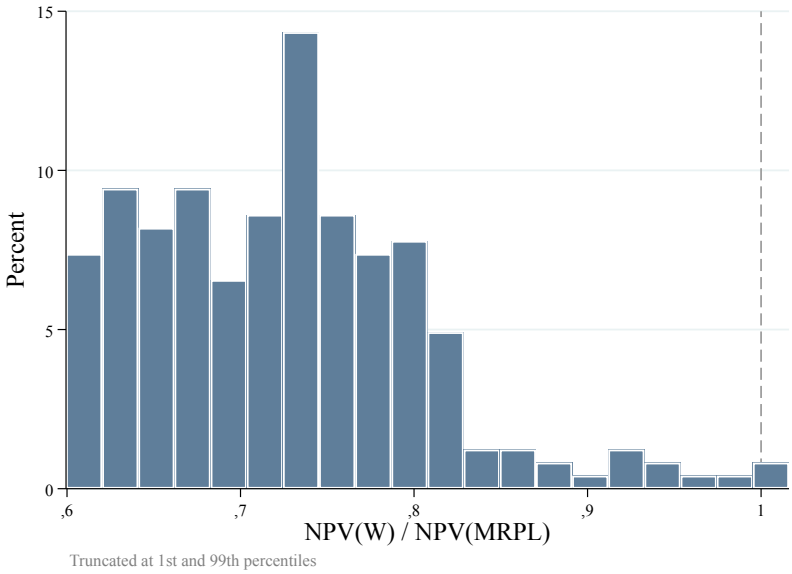
Wages of senior employees with 9+ years of experience are 71% of their MPRL. As soon as accountants are promoted to non-equity partners—who, in practice, still count as employees rather than firm owners—they start earning wage markups, with wages being 22% *above* their marginal revenue product. Equity partners, finally, earn 3 times more than their marginal revenue product.

Given our Proposition **??**, we find that we can reject the null hypothesis of irrelevant effort: wages are signficantly above the MRPL of both non-equity and equity partners. Therefore, public accounting firms are implementing backloaded efficiency wage contracts, and junior employee wage markdowns are (at least partly) due to wage backloading.

4.3.2 Monopsony Power

We compute the net present values (NPVs) of both wages and marginal revenue products from the start of a worker's career until retirement at T. Let λ_t denote the ex-ante probability

Figure 5: Distribution of Expected Wage Markdowns



of still being employed at the firm in year t. We define the NPVs of the wage

$$\text{NPV}_W = \sum_{t=0}^{T} \rho^{t-1} \lambda_t W_t \tag{31}$$

and similarly for the NPV of the marginal revenue product of labor, NPV(MRPL).

We estimate the survival probabilities λ_t using worker stocks and by assuming a steady-state labor market, as explained in Appendix B.2. We impose an annual discount factor $\rho = 0.95$, which corresponds to a real interest rate of around 5%.[27]

Figure 5 presents the distribution of the resulting estimated NPV markdowns. Most of the distribution lies below one, which implies that the NPV of the wage lies below the NPV of the marginal revenue product of labor. The average expected markdown is 0.73, with a 10-90% confidence interval [0.657,0.979], so the NPV of compensation is significantly below the NPV of the MRPL on average. Therefore, in accordance with Proposition 2, our results reject the competitive labor markets assumption that is usually made in tournament models. This implies that the observed wage markdowns for junior employees are not only due to backloaded compensation, but also to monopsony power.

[27] In Appendix B.3.2, we examine robustness of the results to different discount factors.

4.4 Labor Supply Elasticities

We turn our estimated wage markdowns into short- and long-run labor supply elasticities using Equations (20) and (21). Key moments from both elasticity distributions are reported in Table 3. Labor supply is very inelastic in the short-run—with the average elasticity across firms being 0.224 and the median 0.222. This is due to the forward-looking nature of labor supply: changes to starter wages conditional on senior wages do not change residual labor supply much, because workers supply labor in expectation of higher future wages, which remain unaltered. These elasticity estimates are in line with the prior literature that assesses short-run labor supply elasticities using quasi-experimental variation, such as Staiger et al. (2010).

On the long-run, however, labor supply is more elastic, with an elasticity of 2.7 on average and 2.6 at the median firm. This is similar in magnitude to 'steady-state' labor supply elasticities that are estimated using dynamic specifications that take into account separations, such as Bassier et al. (2022). Taken together, our estimated elasticities are consistent with a long literature documenting the discrepancy between short- and long-term elasticities (Manning, 2003; Sokolova and Sorensen, 2021). Notably, we obtain this pattern despite not estimating labor supply in the conventional way; instead, we infer it by inverting our estimated markdowns.

Table 3: Long-term vs. Short-term Labor Supply Elasticities

	Elasticities of Labor Supply			
	Short-run		Long-run	
	Est.	S.E.	Est.	S.E.
Average	0,224	0,133	2,728	0,572
1st Quartile	0,184	0,105	1,875	0,390
Median	0,222	0,126	2,604	0,527
3rd Quartile	0,260	0,164	3,363	0,717
Observations	174		174	

Notes: Selected moments of the long-run and short-run elasticity distributions. When computing the average, distributions are trimmed at the 5st and 95th percentiles to remove outliers. Standard errors are bootstrapped with 250 iterations.

4.4.1 Covariates of Long-Run Labor Supply Elasticities

There is substantial heterogeneity in the long-run labor supply elasticity across firms. Column (I) of Table 4 reports regressions of the log elasticity on two revenue-based firm-

size indicators, controlling for year and region fixed effects. We find that labor supply is substantially more inelastic at small firms than at medium and large ones. This suggests that smaller firms are more horizontally differentiated from the workers' point of view. In terms of the time-series variation of labor supply elasticity, 2019 and—especially—2024 are outliers: long-run elasticity in 2024 is more than 60% higher than that in 2015.

In column (II), we add intermediate input expenditures and the share of CPA-certified personnel to the right-hand side, with the caveat that these are endogeneous variables. Firms with higher training expenditure (relative to revenue) face more elastic labor supply curves, which could indicate that firms are less likely to invest in human capital when they know workers are less likely to switch firms. Firms with a higher share of CPA-certified workers also face more elastic supply, which could indicate that CPA-certified accountants tend to have more attractive outside options than their non-CPA co-workers.

5 Conclusion

In this paper, we examine the exertion of monopsony power when firms use implicit long-term efficiency wage contracts to incentive worker effort. We find that wage markdowns can originate from both the exertion of monopsony power and from the usage of backloaded incentive wage contracts, which have different welfare implications. Using a tournament model with multi-output team production and monopsonistic competition, we develop empirical tests for both monopsony power and backloaded incentive pay. First we find that a sufficient test for the use of backloaded incentive pay is that the wage of senior employees exceeds their marginal revenue product of labor. Second, the degree of monopsony power is characterized by a 'lifecycle' markdown of the find that firms set the NPV of wages at a markdown below the NPV of the marginal revenue product of labor, and that this expected markdown is a function of the long-run residual labor supply elasticity. Intergenerational markdown variation around this long-run supply elasticity is due to backloaded efficiency wages. Our estimates for the U.S. public accounting industry indicate that residual labor supply is very inelastic on the short-run, and more elastic, ranging around 2.7, on the long-run. This implies substantial monopsony power in this industry. However, the very large wage markdowns observed for junior workers would suggest even higher monopsony power if interpreted through the lense of a static model: they are not just the consequence of monopsony power, but also of backloaded incentive pay. In fact, most variation in wage markdowns across workers is due to backloaded incentive pay, rather than monopsony power.

Our results highlight the need to incorporate forward-looking behavior and incentive

Table 4: Covariates of Long-Run Supply Elasticities

	(I)		(II)	
	log(Labor Supply Elast.)			
	Est.	S.E.	Est.	S.E.
Medium-size firm	0,145	0,072	0,085	0,068
Small-size firm	-0,396	0,055	-0,521	0,060
Year = 2016	0,029	0,062	0,028	0,054
Year = 2017	0,007	0,057	-0,020	0,052
Year = 2018	0,110	0,064	0,083	0,059
Year = 2019	0,072	0,069	0,044	0,073
Year = 2020	0,062	0,066	-0,059	0,074
Year = 2021	0,116	0,165	0,037	0,135
Year = 2022	0,340	0,115	0,232	0,126
Year = 2023	0,307	0,076	0,066	0,117
Year = 2024	0,482	0,089	0,239	0,130
log(Marketing exp.)			-0,411	0,115
log(Training exp.)			0,344	0,089
log(Recruiting exp.)			0,081	0,055
log(Technology exp.)			0,163	0,130
CPA share			-0,552	0,481
R-squared	.356		.418	
Observations	238		238	

Notes: Omitted categories are Large firms and the year 2015.

contracts into monopsony models. Markdown variation can be partly efficient, by increasing effort, but can also induce deadweight loss and misallocation, through employer rationing. When evaluating policies that change the markdown distribution within and across firms, such as minimum wages, collective bargaining, or maximum executive compensation, one therefore needs to take into account both the effects of such policies on worker effort and on monopsony distortions.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*(6), 2411–2451.

Agostinelli, F., D. Ferraro, G. Sorrenti, and L. Treuren (2025). Employment relationships, wage setting, and labor market power. Technical report, National Bureau of Economic Research.

Alchian, A. A. and H. Demsetz (1972). Production, information costs, and economic organization. *American Economic Review 62*(5), 777–795.

Azar, J. and I. Marinescu (2024). Monopsony Power in the Labor Market. In *Handbook of Labor Economics*, Volume 5, pp. 761–827. Elsevier.

Azar, J. A., S. T. Berry, and I. Marinescu (2022). Estimating Labor Market Power. *NBER Working Paper, No. 30365*.

Balke, N. and T. Lamadon (2022). Productivity shocks, long-term contracts, and earnings dynamics. *American Economic Review 112*(7), 2139–2177.

Bassier, I., A. Dube, and S. Naidu (2022). Monopsony in movers: The elasticity of labor supply to firm wage policies. *Journal of Human Resources 57*(S), S50–s86.

Berger, D., K. Herkenhoff, A. R. Kostøl, and S. Mongey (2024). An anatomy of monopsony: Search frictions, amenities, and bargaining in concentrated markets. *NBER Macroeconomics Annual 38*(1), 1–47.

Berger, D., K. Herkenhoff, and S. Mongey (2022). Labor Market Power. *American Economic Review 112*(4), 1147–1193.

Brooks, W. J., J. P. Kaboski, Y. A. Li, and W. Qian (2021). Exploitation of labor? classical monopsony power and labor's share. *Journal of Development Economics 150*, 102627.

Burdett, K. and D. T. Mortensen (1998). Wage differentials, employer size, and unemployment. *International economic review*, 257–273.

Caldwell, S., A. Dube, and S. Naidu (2025). Monopsony makes it big. Technical report, Tech. rep.

Card, D. (2022). Who Set Your Wage? *American Economic Review 112*(4), 1075–1090.

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018). Firms and Labor Market Inequality: Evidence and Some Theory. *Journal of Labor Economics 36*(S1), 13–70.

Card, D. and A. B. Krueger (1994). Minimum wages and employment: A case study of

the fast food industry in new jersey and pennsylvania. *American Economic Review 84*(4), 772–793.

Datta, N. (2024). Local monopsony power.

De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016). Prices, markups, and trade reform. *Econometrica 84*(2), 445–510.

Delabastita, V. and M. Rubens (2025). Colluding against workers. *Journal of Political Economy 133*(6), 1796–1839.

Dhyne, E., A. Petrin, V. Smeets, and F. Warzynski (2017). Multi product firms, import competition, and the evolution of firm-product technical efficiencies. Technical report, National Bureau of Economic Research.

Dhyne, E., A. Petrin, V. Smeets, and F. Warzynski (2022). Theory for extending single-product production function estimation to multi-product settings. Technical report, National Bureau of Economic Research.

Dobbelaere, S. and J. Mairesse (2013). Panel data estimates of the production function and product and labor market imperfections. *Journal of Applied Econometrics 28*(1), 1–46.

Feenstra, R. and H. Ma (2007). Optimal choice of product scope for multiproduct firms under monopolistic competition. Technical report, National Bureau of Economic Research.

Gerakos, J. and C. Syverson (2015). Competition in the audit market: Policy implications. *Journal of Accounting Research 53*(4), 725–775.

Gottfries, A. and G. Jarosch (2023). *Dynamic monopsony with large firms and noncompetes*. Number 31965. National Bureau of Economic Research.

Hellerstein, J. K., D. Neumark, and K. R. Troske (1999). Wages, productivity, and worker characteristics: Evidence from plant-level production functions and wage equations. *Journal of Labor Economics 17*(3), 409–446.

Ideagen Audit Analytics (2024). Who audits public companies – 2024 edition.

Inside Public Accounting (2023a). 2023 Firm Administration Report.

Inside Public Accounting (2023b). 2023 Practice Management Report.

Ioannides, Y. M. and C. A. Pissarides (1985). Monopsony and the lifetime relation between wages and productivity. *Journal of Labor Economics 3*(1, Part 1), 91–100.

Iyoha, E. (2023). *Estimating productivity in the presence of spillovers: Firm-level evidence from the us production network*. Harvard Business School.

Jarosch, G., J. S. Nimczik, and I. Sorkin (2024). Granular search, market structure, and wages. *Review of Economic Studies 91*(6), 3569–3607.

Jungerman, W. (2023). Dynamic monopsony and human capital. Technical report, Working Paper.

Kline, P. M. (2025). Labor market monopsony: Fundamentals and frontiers.

Kotlikoff, L. J. and J. Gokhale (1992). Estimating a firm's age-productivity profile using the present value of workers' earnings. *The Quarterly Journal of Economics 107*(4), 1215–1242.

Lamadon, T., M. Mogstad, and B. Setzler (2022). Imperfect Competition, Compensating Differentials, and Rent Sharing in the US Labor Market. *American Economic Review 112*(1), 169–212.

Lazear, E. P. (1981). Agency, earnings profiles, productivity, and hours restrictions. *The American Economic Review 71*(4), 606–620.

Lazear, E. P. and S. Rosen (1981). Rank-order tournaments as optimum labor contracts. *Journal of political Economy 89*(5), 841–864.

Malcomson, J. M. (1984). Work incentives, hierarchy, and internal labor markets. *Journal of political economy 92*(3), 486–507.

Manning, A. (1987). An Integration of Trade Union Models in a Sequential Bargaining Framework. *The Economic Journal 97*(385), 121–139.

Manning, A. (2003). *Monopsony in Motion: Imperfect Competition in Labor Markets*. Princeton University Press.

Medoff, J. L. and K. G. Abraham (1980). Experience, performance, and earnings. *Quarterly Journal of Economics 95*(4), 703–736.

Mertens, M. (2022). Micro-mechanisms behind declining labor shares: Rising market power and changing modes of production. *International Journal of Industrial Organization 81*, 102808.

Morlacco, M. (2019). Market Power in Input Markets: Theory and Evidence From French Manufacturing. *Working Paper*.

Orr, S. (2022). Within-firm productivity dispersion: Estimates and implications. *Journal of Political Economy 130*(11), 2771–2828.

Roussille, N. and B. Scuderi (2021). Bidding for talent. *Unpublished manuscript*.

Sokolova, A. and T. Sorensen (2021). Monopsony in labor markets: A meta-analysis. *ILR*

Review 74(1), 27–55.

Staiger, D. O., J. Spetz, and C. S. Phibbs (2010). Is there monopsony in the labor market? evidence from a natural experiment. *Journal of Labor Economics 28*(2), 211–236.

Treuren, L. (2022). Wage markups and buyer power in intermediate input markets. *FEB Research Report Department of Economics*.

Valmari, N. (2023). Estimating production functions of multiproduct firms. *Review of Economic Studies 90*(6), 3315–3342.

Volpe, O. (2024). Job preferences, labor market power, and inequality. Technical report, Discussion paper, Working Paper.

Yeh, C., C. Macaluso, and B. Hershbein (2022). Monopsony in the US Labor Market. *American Economic Review 112*(7), 2099–2138.

Wage Markdowns and Dynamic Labor Supply

Michael Rubens, Bernardo Silveira

Online Appendix

Contents

A Theory Appendix . . . . . **OA - 2**
A.1 Firm's Problem: First-Order Conditions OA - 2
A.2 Proofs of Propositions in the Main Text OA - 3
A.3 Bundle Pricing OA - 4

B Empirical Analysis Appendix . . . . . **OA - 7**
B.1 Aggregation to the Bin-Level OA - 7
B.2 Promotion Probabilities OA - 8
B.3 Alternative Specifications and Robustness OA - 10

C Tournament Model with Monopsonistic Competition and N Employee Levels . . . . . **OA - 16**
C.1 Environment and Timing OA - 16
C.2 Production and Prices OA - 16
C.3 Compensation–Promotion Policy OA - 17
C.4 Outside Options OA - 17
C.5 Workers' Values and Incentives OA - 17
C.6 Employment Dynamics OA - 18
C.7 Firm's Problem OA - 18
C.8 Steady State OA - 19
C.9 Incentive Contract Implications OA - 20
C.10 Entry Elasticity and the Exact Intertemporal Wedge OA - 22

D Tournament Model with Lateral Hiring . . . . . **OA - 24**
D.1 Primitives OA - 24
D.2 Worker and Firm Behavior OA - 24

E Appendix Tables and Figures . . . . . **OA - 26**

Online Appendix

A Theory Appendix

A.1 Firm's Problem: First-Order Conditions

We start from the recursive value function (8). We derive the first-order conditions associated with W_1, W_2, and τ in steady state. To simplify notation and improve readability, we define

$$B_1 = G_1^{\text{alt}}(U_1)\mathcal{L}, \qquad B_2 = G_2^{\text{alt}}(U_2).$$

Using these definitions, the first-order conditions for W_1, W_2, and τ can be written as:

$$W_1: \ \frac{\partial U_1}{\partial W_1} g_1^{alt}(U_1)\mathcal{L}\Big[MRPL_1 - W_1 - \rho W_2(1-\tau)B_2 + \rho(1-\tau)B_2 V'\Big((1-\tau)B_2B_1\Big)\Big] - B_1 = 0,$$

In (OA.1), the expression within square brackets captures how the firm's value changes as the number of juniors increases, in response to a raise in W_1. The term B_1 outside the brackets represents the marginal increase in the firm's wage bill resulting from the higher junior compensation.

$$\begin{aligned} W_2: \ & \frac{\partial U_1}{\partial W_2} g_1^{alt}(U_1)\mathcal{L}\Big[MRPL_1 - W_1 - \rho W_2(1-\tau)B_2 + \rho(1-\tau)B_2\ V'\big((1-\tau)B_2B_1\big)\Big] \\ & - \rho(1-\tau)B_1B_2 + \rho(1-\tau)B_1\frac{\partial U_2}{\partial W_2} g_2^{alt}(U_2)\Big[-W_2 + V'\big((1-\tau)B_2B_1\big)\Big] \\ & + \frac{1+\eta}{\eta}\frac{\partial \tilde{e}}{\partial W_2}\left\{d\left[F_1(B_1, l_2, \tilde{e}, e_2^*)\right]^{\frac{1}{\eta}}\frac{\partial F_1}{\partial \tilde{e}} + d\left[F_2(l_2, B_1, e_2^*, \tilde{e})\right]^{\frac{1}{\eta}}\frac{\partial F_2}{\partial \tilde{e}}\right\} = 0, \end{aligned}$$

As in the previous case, the term inside the square brackets reflects the change in the number of juniors attracted by a higher W_2. In the second line, the first term represents the marginal (discounted) increase in the wage bill due to the higher W_2, while the second term captures the additional number of senior employees the firm retains. Finally, the term in the third line corresponds to the greater effort exerted by junior employees in response to the change in W_2.

$$\begin{aligned} \tau: \ & \frac{\partial U_1}{\partial \tau} g_1^{alt}(U_1)\mathcal{L}\Big[MRPL_1 - W_1 - \rho W_2(1-\tau)B_2 + \rho(1-\tau)B_2\ V'\big((1-\tau)B_2B_1\big)\Big] \\ & + \rho W_2B_2B_1 - \rho B_2B_1V'\big((1-\tau)B_2B_1\big) \\ & + \frac{1+\eta}{\eta}\frac{\partial \tilde{e}}{\partial \tau}\left\{d\left[F_1(B_1, l_2, \tilde{e}, e_2^*)\right]^{\frac{1}{\eta}}\frac{\partial F_1}{\partial \tilde{e}} + d\left[F_2(l_2, B_1, e_2^*, \tilde{e})\right]^{\frac{1}{\eta}}\frac{\partial F_2}{\partial \tilde{e}}\right\} = 0. \end{aligned}$$

Once again, the term in square brackets captures how the number of juniors hired by the firm responds to a change in τ. The term in the second line reflects the impact on the retention of senior employees, while the term in the third line accounts for the change in equilibrium effort induced by the variation in τ. Note that this last effect can be positive or negative, depending on the shape of $\phi(\cdot)$ and the value of τ. For instance, if $\phi(\cdot)$ is symmetric around zero, the equilibrium effort increases with τ when $\tau < 0$ and decreases when $\tau > 0$. It is also worth recalling that, from the envelope theorem, $V'(l_2)$ is given by (10).

A.2 Proofs of Propositions in the Main Text

A.2.1 Proof of Proposition 1

From (4), we have

$$\frac{\partial U_1}{\partial W_1} = \alpha.$$

Meanwhile, from (3), (4), and the definition of $\bar{\bar{U}}_2$, we have that

$$\frac{\partial U_1}{\partial W_2} = \rho(1-\tau)B_2\alpha.$$

From the firm's first-order condition with respect to W_1,

$$g_1^{alt}(U_1)\mathcal{L}\Big[MRPL_1 - W_1 - \rho W_2(1-\tau)B_2 + \rho(1-\tau)B_2V'\big((1-\tau)B_2B_1\big)\Big] = \frac{B_1}{\alpha}.$$

Substituting the left-hand side for the right-hand side on the firm's first-order condition with respect to W_2,

$$\begin{aligned}
&\rho(1-\tau)B_2B_1 - \rho(1-\tau)B_2B_1 \\
&\qquad + \rho(1-\tau)B_1\alpha g_2^{alt}(U_2)\left[-W_2 + MRPL_2\right] \\
&\qquad + \frac{1+\eta}{\eta}\frac{\partial \tilde{e}}{\partial W_2}\left\{d\left[F_1(B_1, l_2, \tilde{e}, e_2^*)\right]^{\frac{1}{\eta}}\frac{\partial F_1}{\partial \tilde{e}} + d\left[F_2(l_2, B_1, e_2^*, \tilde{e})\right]^{\frac{1}{\eta}}\frac{\partial F_2}{\partial \tilde{e}}\right\} = 0. \qquad \text{(OA.1)}
\end{aligned}$$

The terms in the first line of (OA.1) cancel each other out. Then, we can conclude the following:

- In the Irrelevant Effort scenario, the term in the third line of (OA.1) equals zero. Under the assumption that $\tau \in (0,1)$, the prefactor in the second line is positive. Therefore, we have $W_2 = MRPL_2$.
- In the Productive Effort scenario, the term in the third line of (OA.1) is strictly positive, from Lemma 1. We must then have $W_2 > MRPL_2$.

A.2.2 Proof of Proposition 2

From (20) and $\partial U_1/\partial W_1 = \alpha$, we obtain

$$\alpha g_1^{alt}(U_1) = \frac{\psi_1}{W_1} G_1^{alt}(U_1). \tag{OA.2}$$

Moreover, from (10), we have

$$V'\Big((1-\tau)B_2 B_1\Big) = \text{MRPL}_2. \tag{OA.3}$$

Substituting the expressions in (OA.2) and (OA.3) into the first-order condition for W_1, and rearranging, yields

$$\text{MRPL}_1 + \rho(1-\tau)G_2^{alt}(U_2)\,\text{MRPL}_2 = W_1 + \rho(1-\tau)G_2^{alt}(U_2)\,W_2 + \frac{W_1}{\psi_1}.$$

Finally, combining (17), (18), and (21), we can rewrite this condition as

$$\frac{\text{NPV}_\text{W}}{\text{NPV}_\text{MRPL}} = \frac{\psi_{LR}}{1+\psi_{LR}},$$

which corresponds to the statement of the proposition.

A.3 Bundle Pricing

A.3.1 MRPL formula for bundle pricing

In the main text, we assumed that firms use a line-pricing model, in which clients are charged based on individual accountants' billing rates, rather than a bundled price for the entire accounting service. This was motivated by the fact that the IPA reports indicate that the vast majority of surveyed firms use such line-pricing practices. In this appendix, we work out the markdown and MRPL expressions if firms would price bundles instead.

We impose that clients consume a CES bundle of accountants, with shares a_i and elasticity of substitution σ:

$$Q_f = [\sum_{i=1}^{I} a_i Q_{if}^{\frac{\sigma-1}{\sigma}}]^{\frac{\sigma}{\sigma-1}} \tag{OA.4}$$

The implied bundle price is:

$$P_f = [\sum_{i=1}^{I} a_i^{\sigma} P_{if}^{1-\sigma}]^{\frac{1}{1-\sigma}}$$

Demand for the bundle is

$$Q_f = \overline{Q}\left(\frac{P_f}{\overline{P}}\right)^{\eta} \zeta_f. \tag{OA.5}$$

The marginal revenue product of labor for each accountant type becomes:

$$\text{MRPL}_{if} = P_f \left(1 - \frac{1}{\eta}\right) \sum_{k=1}^{I} \alpha_k \left(\frac{Q_f}{Q_{kf}}\right)^{\frac{1}{\sigma}} \beta_{ki} \frac{Q_{kf}}{L_{if}}$$

A.3.2 Recovering Product Weights

The first-order condition implies, for any i and j,

$$\frac{Q_i}{Q_j} = \left(\frac{a_i}{a_j}\right)^{\sigma} \left(\frac{P_i}{P_j}\right)^{-\sigma}. \tag{OA.6}$$

Choose good 7 as the comparison good. Define the observable constants

$$\kappa_i \equiv \left(\frac{Q_i}{Q_7}\right)^{1/\sigma} \frac{P_i}{P_7}, \qquad i = 1, \ldots, 6. \tag{OA.7}$$

The relative demand conditions then become

$$a_i = \kappa_i a_7, \qquad i = 1, \ldots, 6. \tag{OA.8}$$

We impose the normalization

$$\sum_{i=1}^{7} a_i = 1. \tag{OA.9}$$

Thus, we obtain a system of 7 equations with 7 unknowns, which has the solution:

$$a_i = \frac{\kappa_i}{1 + \sum_{k=1}^{6} \kappa_k}. \tag{OA.10}$$

for each $i = 1, \ldots, 6$, and

$$a_7 = \frac{1}{1 + \sum_{i=1}^{6} \kappa_i}, \tag{OA.11}$$

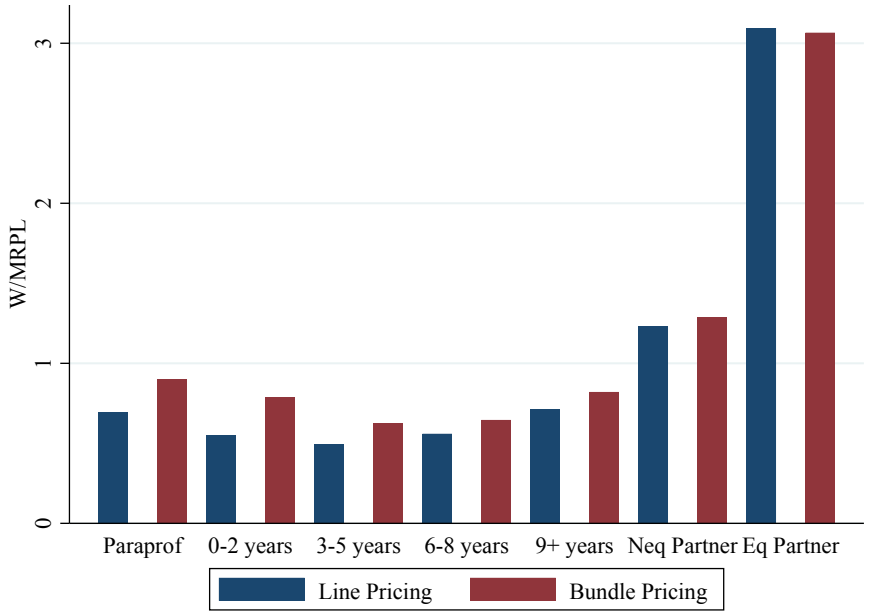
A.3.3 Markdown Estimation

We estimate the seniority-specific wage markdowns under bundle pricing. We solve the product weights a_i as explained above and calibrate two different values $\sigma = 0.5$ and $\sigma = 2$,to allow accountants of different seniority levels to be complements and substitutes, respectively.

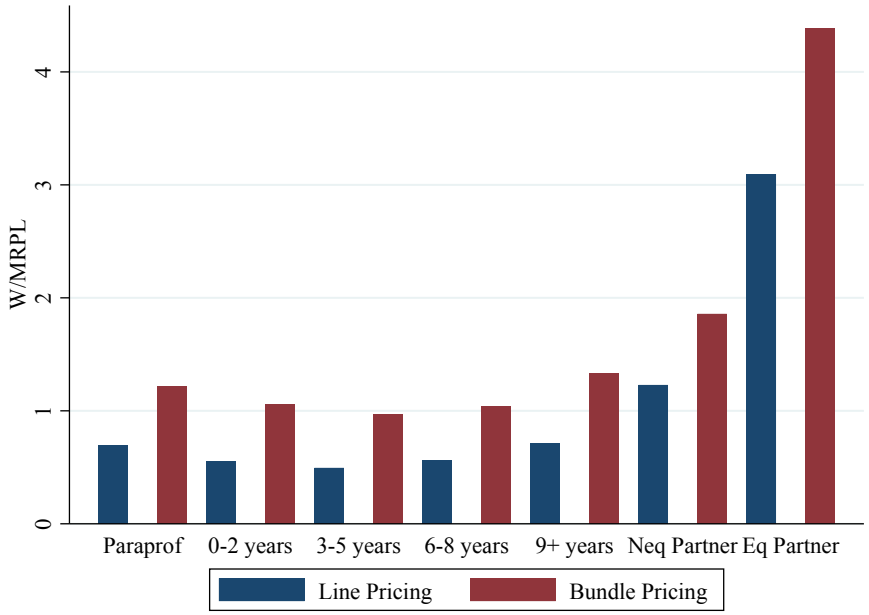
The results are in Figure OA-1. We find similar markdowns for line pricing and bundle pricing when the accountants are gross complements ($\sigma = 0.5$), although equity partner wage markups are slightly smaller under the bundle specification. However, when assuming that the accountants are gross substitutes ($\sigma = 2$), all employees are estimated to earn wage markups, which is impossible as this would imply loss-making firms.

Figure OA-1: Wage Markdowns: Bundle Pricing vs. Line Pricing

(a) $\sigma = 0.5$



(b) $\sigma = 2$



Notes: Average wage markdowns by seniority level under bundle pricing, for $\sigma = 1.5$ and $\sigma = 2.5$.

B Empirical Analysis Appendix

B.1 Aggregation to the Bin-Level

Suppose each firm in bin g faces the same input prices and residual input supply elasticities for all labor and intermediate inputs, although these prices and elasticities may vary across bins. With a Cobb–Douglas technology and cost minimization, identical within-bin marginal costs imply proportional input bundles across firms:

$$L_{if} = s_{fg} L_{ig}, \qquad \boldsymbol{V}_f = s_{fg} \boldsymbol{V}_g, \tag{OA.12}$$

for weights s_{fg} satisfying $\sum_{f \in g} s_{fg} = 1$. Under these conditions, total output in bin g is

$$\begin{aligned} Q_{ig} &= \sum_{f \in g} \left[\exp(\omega_{if} + \varepsilon_{if}) \prod_k (L_{ikf})^{\beta_{ik}} \boldsymbol{V}_{if}^{\beta_i^v} \right] \\ &= \left[\sum_{f \in g} \exp(\omega_{if} + \varepsilon_{if}) \, s_{fg}^{\sum_k \beta_{ik} + \beta_i^v} \right] \prod_k L_{ikg}^{\beta_{ik}} \boldsymbol{V}_{ig}^{\beta_i^v}. \end{aligned} \tag{OA.13}$$

Taking logs yields a Cobb–Douglas relationship at the bin level:

$$q_{ig} = \sum_k \beta_{ik} l_{ikg} + \beta_i^v \boldsymbol{v}_{ig} + \omega_{ig}, \tag{OA.14}$$

where the group productivity term is $\omega_{ig} = \ln\big(\sum_{f \in g} \exp(\omega_{if} + \varepsilon_{if}) s_{fg}^{\sum_k \beta_{ik} + \beta_i^v} \big)$. Hence, the output elasticities (β_{ik}, β_i^v) are preserved under aggregation.

If each firm's productivity follows the AR(1) process above with a common σ_i, then bin-level productivity satisfies

$$\omega_{ig} = \sigma_i \tilde{\omega}_{ig} + \nu_{ig}, \tag{OA.15}$$

for a bin-level innovation ν_{ig} that aggregates the firm-level shocks ν_{if}. Replacing firm-level variables with bin-level aggregates, the residual used in estimation is

$$\nu_{ig} = q_{ig} - \sigma_i \tilde{q}_{ig} - \Big(\sum_k \beta_{ik} l_{ikg} - \sigma_i \sum_k \beta_{ik} \tilde{l}_{ikg} \Big) - \Big(\beta_i^v \boldsymbol{v}_{ig} - \sigma_i \beta_i^v \tilde{\boldsymbol{v}}_{ig} \Big). \tag{OA.16}$$

The timing assumptions that underpin the firm-level moments carry over directly to the bin level. All professional labor inputs $\boldsymbol{L}_{ig}$ are dynamic inputs chosen prior to observing the transitory productivity shock ν_{ig}, implying

$$E(\nu_{ig} \mid \boldsymbol{L}_{ig}, \tilde{\boldsymbol{L}}_{ig}) = 0. \tag{OA.17}$$

Intermediate and intern inputs $\boldsymbol{V}_{ig}$ are variable inputs chosen after shocks are realized, so their lags remain valid instruments:

$$E(\nu_{ig} \mid \tilde{\boldsymbol{V}}_{ig}) = 0. \tag{OA.18}$$

A key assumption for the exclusion restriction to carry over to the bin-level is that group compositions are exogenous w.r.t. firm-level productivity shocks. If entry into and exit from bins would be correlated with the transitory productivity shocks, this would violate the exclusion restriction. Given that exit takes time, it is likely that exit decisions are more to be predetermined, similar to capital investment decisions, which would validate the imposed timing assumptions.

Equation (OA.14) shows that, under the assumptions of (i) common technology parameters, (ii) identical input prices within each bin-year (so that inputs are proportional across firms), and (iii) the same timing and exogeneity restrictions as at the firm level, the output elasticities estimated using bin-level data are identical to those that would be obtained at the firm level. Price variation across bins affects only the bin-specific productivity term ω_{ig} and not the estimated elasticities. Violations of proportionality (for example, heterogeneous within-bin input ratios or factor price dispersion) would break this equivalence, in which case additional composition controls would be required.

B.2 Promotion Probabilities

B.2.1 Career Profile for the Representative Worker

We assume that the representative worker in our dataset joins the public accounting firm at the age of 23, having taken a 4-year college degree and passing CPA licensing, which usually happens through a one-year master's degree. The seniority grades listed in the reports are "0-2 years of experience", "3-5 years of experience", "6-8 years of experience","9+ years of experience", "non-equity partner", and "partner". We assume that the rank "9+ years of experience" is held for 5 years before being promoted to non-equity partner, and that non-equity partnership is being held for 3 years before promotion to equity partner, which implies promotion to non-equity partner at 36 and to equity partner at 39. We choose 5 and 3 years to match the IPA reports, which report average experience to reach partnership at 12-15 years, and to match the fact that substantially more people are reported in the 9+ years rank than in the 6-8 years rank, which indicates that people spend more than 3 years in the 9+ years rank.

We assume that for the first four promotions, employees leave the firm when not being promoted, in line with the 'up-or-out' policies. Non-equity partners that are not promoted to equity partner after three years are assumed to have the option to stay at the firm until retirement as a non-equity partner, with an annual probability of leaving the firm (voluntarily or otherwise) τ^{neqp}. Similarly, equity partners stay at the firm until retirement, with a different annual probability of separating prior the the retirement age τ^{eqp}. We estimate these separation probabilities below. The vast majority of the surveyed firms have mandatory retirement for partners at age 65 (Inside Public Accounting, 2023). For equity partners reaching the mandatory retirement age, this implies being an equity partner for 26 years, and puts the entire career at 42 years.

B.2.2 Estimating Promotion Probabilities

Let $i = 1, 2, .., 6$ denote the 'partner-track' grades in increasing seniority, with $i = 1$ being the entry-level grade and $i = 6$ being equity partnership. We assume that non-partner employees never leave within a grade band i. Denote the number of employees in each grade by N_i and the length in years of grade i by T_i. Assuming a steady-state labor force, the ex-ante cumulative probability of being promoted to grade i when starting the career is denoted by λ_i. For all grades except non-equity and equity partners, we compute these probabilities using averages of the employee counts across firms and years.

$$\lambda_i = \frac{N_i/T_i}{N_1/T_1}, \quad i \in \{1, \ldots, 4\}.$$

We allow these averages to differ by three size groups of firms (small, medium, and large, as defined by the IPA reports), in order to allow firms of different sizes to have different promotion structures.

For both non-equity and equity partners, we directly observe the number of newly promoted partners, N_5^{new} and N_6^{new}. Hence, we compute the probability of reaching non-equity and equity partnership as:

$$\lambda_i = \frac{N_i^{new}}{N_1/T_1}, \quad i \in \{5, 6\}.$$

The resulting cumulative transition rates are shown in Figure 1 in the main text.

B.2.3 Partner Turnover

Once reaching partnership, we assume that individuals have an annual separation probability τ_5 and τ_6 for non-equity and equity partners, respectively. Individuals who remain at the firm until age 65 are assumed to retire mandatorily. We estimate these annual separation probabilities. In steady state, the total number of non-equity and equity partners, N_5 and N_6, is related to the inflow of new partners in each category, both of which are observed in the data:

$$N_{\mathrm{i}} = N_{\mathrm{i}}^{\mathrm{new}} \sum_{t=0}^{T_i} (1 - \tau_{\mathrm{i}})^t \quad i \in \{5, 6\}$$

Using the summational of a finite geometric series, the expression above simplifies to:

$$\frac{N_{\mathrm{i}}}{N_{\mathrm{i}}^{\mathrm{new}}} = \frac{1 - (1 - \tau_{\mathrm{i}})^{T_i}}{\tau_{\mathrm{i}}} \tag{OA.19}$$

Since the latter equation lacks a closed-form solution for τ_{i}, we solve it numerically. We find $\tau_5 = .170$ and $\tau_6 = .041$. This meaning that non-equity partners have an annual probability of leaving their grade of 17%, which includes leaving the firm and being promoted to equity partner, and that equity partners have an annual probability of separating from the firm of 4.1%.

B.3 Alternative Specifications and Robustness

B.3.1 Fixed Costs

Although the data does not report fixed costs and overhead, such as office rent and insurance costs, we can impute fixed costs by imposing a zero-profit condition. A zero-profit condition makes sense in steady state for public accounting firms, given that equity partners are the residual claimants of profits.

We write total observed costs as VC_{ft}, which includes labor expenses for the seven occupations listed in the reports and reported intermediate input expenses. We include administrative and intern expenses using the observed headcounts and assuming administrative and intern salaries of 100K and 62K, respectively.The administrative salary benchmark is based on the administrator wages observed in the IPA reports, the internship salaries is based on glassdoor.[28]

$$FC_{ft} = R_{ft} - VC_{ft}$$

We find a ratio of fixed costs to revenue of 26.5% on average, which is slightly higher than the 20-25% rate we found in industry reports.[29].

We include the logarithm of fixed costs as an additional inputs to the team production model. We include lagged but not current fixed costs in the instrument vectors, as fixed costs can reasonably be thought of as being predetermined.

The resulting production estimates are in Table OA-1. The produciton estimates are very similar to the baseline model in which fixed costs are excluded from the production function. One way to rationalize this would be that fixed overhead costs are mostly perfect complements to the aggregator of the other inputs. This seems to make sense when considering fixed expenses such as office rent and insurance expenses.

B.3.2 Discount Factors

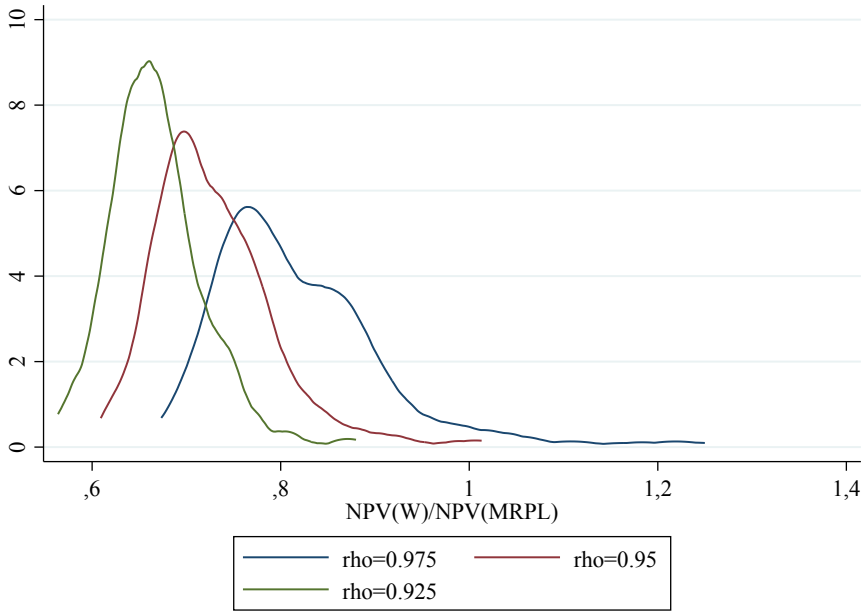
We examine robustness of our empirical results to different values of the discount factor, which was calibrated to $\rho = 0.95$ in the main text. We re-estimate our model for $\rho = 0.925$ and $\rho = 0.975$. The resulting distribution of the expected markdown (NPV(W)/NPV(MRPL)) is in Figure OA-2. A higher discount factor pushes the distribution closer to one, as employees are more patient and, therefore, value future wage markups more. Nevertheless, the expected markdown distribution still lies almost entirely below one for all of the discount factors considered.

[28] https://www.glassdoor.com/Salaries/accounting-intern-salarySRCH_KO0,17.htm
[29] https://www.cpapracticeadvisor.com/2014/12/16/how-much-admin-expense-should-accounting-firms-incur/17560/

Table OA-1: Production Estimates with Fixed Costs in Production

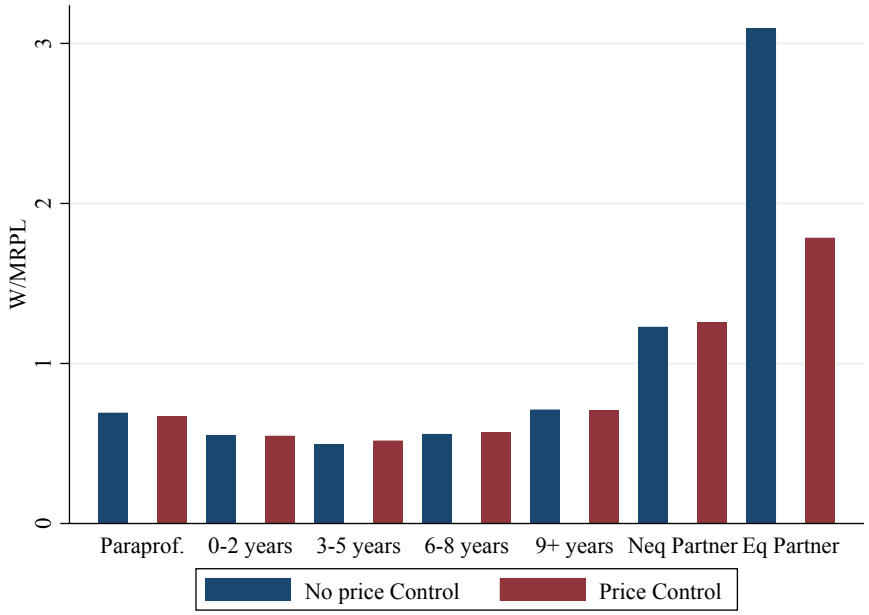
	(I) 0-2 years		(II) 3-5 years		(III) 6-8 years		(IV) 9+ years		(V) Non-eq. Part.		(VI) Eq. Part.		(VII) Paraprof.	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
0-2 years	0,969	0,045	-0,026	0,045	-0,024	0,042	-0,050	0,049	-0,069	0,078	-0,069	0,078	0,089	0,149
3-5 years	0,031	0,032	1,011	0,025	-0,037	0,030	-0,011	0,022	0,015	0,041	0,015	0,041	-0,002	0,095
6-8 years	0,016	0,030	0,011	0,021	1,021	0,029	-0,031	0,024	0,045	0,041	0,045	0,041	0,065	0,088
9 years	-0,060	0,055	-0,036	0,033	-0,023	0,049	0,985	0,030	-0,044	0,056	-0,044	0,056	-0,094	0,138
Non-eq. Partners	0,006	0,018	0,002	0,011	0,008	0,020	-0,007	0,012	0,968	0,023	-0,032	0,023	-0,057	0,045
Eq. Partners	0,070	0,040	0,088	0,050	0,133	0,043	0,106	0,047	0,174	0,085	1,174	0,085	0,075	0,145
Paraprof.	-0,106	0,036	-0,075	0,030	-0,062	0,028	-0,053	0,034	-0,114	0,052	-0,114	0,052	0,753	0,103
Interns	0,010	0,033	0,020	0,016	-0,018	0,025	0,010	0,017	-0,011	0,028	-0,011	0,028	0,048	0,070
Materials	0,001	0,019	0,010	0,014	-0,015	0,027	0,004	0,013	-0,017	0,029	-0,017	0,029	0,067	0,046
Admin. staff	0,008	0,021	-0,025	0,012	-0,016	0,024	0,019	0,014	-0,029	0,026	-0,029	0,026	0,045	0,065
Capital	0,037	0,040	0,027	0,032	0,030	0,041	0,020	0,030	0,062	0,060	0,062	0,060	0,018	0,127
Ser. corr.	0,773	0,072	0,908	0,064	0,796	0,092	0,921	0,074	0,856	0,060	0,856	0,060	0,828	0,077
Obs.	165		165		165		165							

Figure OA-2: Robustness: Discount Factors



Notes: We re-estimate the model with discount factors being $\rho = 0.975$, $\rho = 0.95$, and $\rho = 0.925$.

Figure OA-3: Including Price Controls



Notes: Average wage markdowns by seniority level when including a price control in the production function.

B.3.3 Product Differentiation

To allow for differentiation in accountant characteristics between firms, we introduce a price control into the right-hand side of the production function, similarly to De Loecker et al. (2016).

$$q_{if} = \sum_{k \in I} \left(\beta_{ik} l_{ikf}\right) + \beta_i^v \boldsymbol{v}_{if} + \beta^p \ln(p_{if}) + \omega_{if} + \varepsilon_{if}. \tag{OA.20}$$

The resulting wage markdowns are very similar to those estimated without price control, as shown in Figure OA-3.

B.3.4 Firm-level Production Function

How important is it to specify a team production model as opposed to a more conventional production function that is aggregated at the firm level? To answer this question, we estimate a firm-level production model and compare its implied markdowns to the team production model. Given that it is not obvious to aggregate physical output (billable hours), we include log firm-level revenue on the left-hand side of the production function, and all log input quantities on the right.

We estimate both a Cobb-Douglas specification, Equation (OA.21), and a translog specification, Equation (OA.22)

$$r_f = \sum_{k \in I} \left(\beta_k l_{kf} \right) + \beta^o \boldsymbol{o}_f + \omega_f + \varepsilon_f. \tag{OA.21}$$

$$\begin{aligned} r_f = \beta_0 &+ \sum_{k \in I} \beta_k \, l_{kf} + \beta^o \, o_f \\ &+ \tfrac{1}{2} \sum_{k \in I} \sum_{j \in I} \gamma_{kj} \, l_{kf} l_{jf} \\ &+ \omega_f + \varepsilon_f. \end{aligned} \tag{OA.22}$$

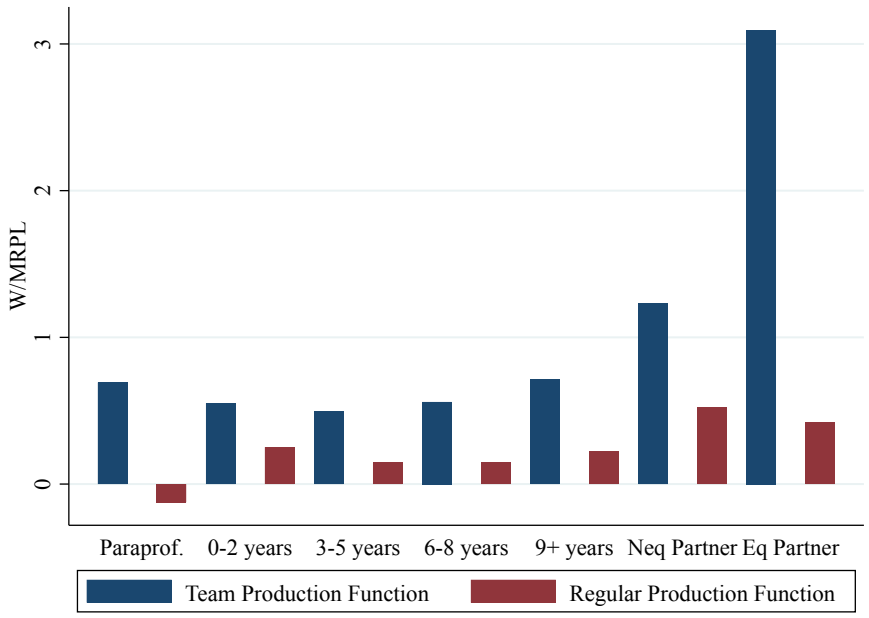
We estimate these revenue production functions with the same moment conditions and TFP transition process as assumed in the main text.

The production function estimates are in Table OA-2, and the resulting markdowns in Figure OA-4. The firm-level production model implies unrealistically large wage markdowns across the board for both the Cobb-Douglas and translog specifications. Also, the productivity spillovers from equity partners to all other labor inputs that we documented in the team production model are not picked up by the interaction effects in the translog specification.

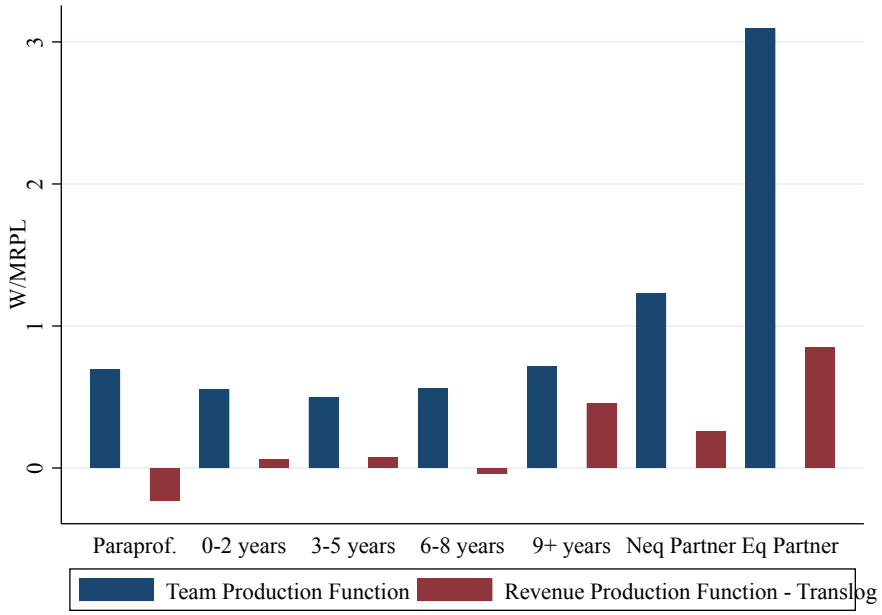
We think that the failure of the firm-level production model to generate realistic marginal revenue products stems from two sources. First, the input-specific team production functions do not aggregate to a firm-level Cobb-Douglas or translog specification, even if the product market would be in perfect competition. Second, given that the product market is not in perfect competition, the revenue production function suffers from the usual bias due to endogenous product prices. Controlling for prices is not obvious without imposing more structure on the demand-side, given that we observe product-level, rather than firm-level, prices.

Figure OA-4: Revenue Production Function

(a) Cobb-Douglas



(b) Translog



Notes: Average wage markdowns by seniority level when using a team production function and a revenue production function.

Table OA-2: Firm-level Production Function

	(I) Cobb-Douglas		(II) Translog	
	Est.	S.E.	Est.	S.E.
0-2 years	0,094	0,045	-1,506	2,241
3-5 years	0,200	0,026	1,361	1,199
6-8 years	0,179	0,023	-1,573	0,772
9 years	0,254	0,025	0,452	0,730
Non-eq. Partners	0,051	0,012	0,539	0,594
Eq. Partners	0,371	0,045	2,632	1,816
Paraprof.	-0,074	0,034	-0,228	0,390
Interns	-0,023	0,016	-0,037	0,026
Materials	0,011	0,008	0,011	0,010
Admin. staff	-0,036	0,017	-0,007	0,094
Ser. corr.	0,908	0,077	0,989	0,039
0-2 years x 0-2 years			0,300	0,178
0-2 years x 3-5 years			0,261	0,244
0-2 years x 6-8 years			-0,764	0,289
0-2 years x 9+ years			0,174	0,370
0-2 years x paraprof.			0,127	0,130
0-2 years x eq. part.			-0,153	0,613
0-2 years x neq. part.			-0,108	0,115
3-5 years x 3-5 years			-0,045	0,128
3-5 years x 6-8 years			0,388	0,235
3-5 years x 9+ years			-0,367	0,181
3-5 years x paraprof.			-0,055	0,100
3-5 years x eq. part.			-0,224	0,433
3-5 years x neq. part.			-0,016	0,161
6-8 years x 6-8 years			-0,133	0,178
6-8 years x 9+ years			0,130	0,170
6-8 years x paraprof.			0,137	0,091
6-8 years x eq. part.			0,490	0,342
6-8 years x neq. part.			0,081	0,119
9+ years x 9+ years			-0,060	0,115
9+ years x paraprof.			-0,293	0,115
9+ years x eq. part.			0,446	0,382
9+ years x neq. part.			-0,019	0,154
9+ years x paraprof.			-0,293	0,115
paraprof. x eq. part.			0,119	0,188
paraprof. x neq. part.			-0,025	0,053
eq. part. x eq. part.			-0,431	0,463
eq. part. x neq. part.			-0,056	0,213
neq. part. x neq. part.			0,056	0,030
Obs.	174		174	

C Tournament Model with Monopsonistic Competition and N Employee Levels

We extend the rank-order tournament framework presented in Section 3 to an environment with N employee levels: a junior level $s = 0$ and $N-1$ senior levels $s = 1, \ldots, N-1$. Workers are potentially productive for N periods. If promoted at each stage, a worker who enters at time t moves from level s in period $t+s$ to level $s+1$ in period $t+s+1$, and retires at the beginning of period $t+N$.

To simplify the notation, we assume that the firm operates in a perfectly competitive product market. Extending the analysis to a monopolistically competitive firm, as we do in Section 3.1.4, is straightforward.

For expositional simplicity, we also assume that there are no effort spillovers across employee levels. That is, the effort exerted by workers at level s affects their own output, but not the output produced at other levels. We do, however, allow for cross-level production spillovers in employment: the number of level-s workers may affect production at other levels. Allowing for spillovers both in effort and in employment, as we do in Section 3.1.4, would be conceptually direct, albeit at the cost of heavier notation.

C.1 Environment and Timing

Time is discrete and indexed by t. The discount rate, ρ, is the same to all agents. Each period, a mass $\mathcal{L}$ of juniors enters the market. A worker from the cohort hired in period t is at level s in period $t+s$ if (and only if) she has been promoted at every prior stage.

Each employed worker i at level $s \in \{0, 1, \ldots, N-1\}$ in period t chooses effort $e_{i,s,t} \in [\underline{e}, \infty)$ and pays cost $C(e_{i,s,t})$ with $C'(\cdot) > 0$ and $C''(\cdot) \geq 0$. The firm observes a noisy, non-contractible performance signal

$$\theta_{i,s,t} \;=\; e_{i,s,t} + \kappa_{i,s,t}, \qquad \kappa_{i,s,t} \sim \Phi(\cdot) \text{ i.i.d.},$$

with density $\phi(\cdot)$.

C.2 Production and Prices

Let $L_{s,t}$ denote employment at level s in period t, and $\bar{e}_{s,t}$ the average effort of level-s workers. The firm produces N outputs,

$$Q_{s,t} \;=\; F_s\big(L_{s,t}, L_{-s,t}, \bar{e}_{s,t}\big), \qquad s = 0, 1, \ldots, N-1,$$

where $L_{-s,t}$ stacks the other levels to allow cross-level spillovers. Product prices $\{p_s\}_{s=0}^{N-1}$ are taken as given at the firm level.

C.3 Compensation–Promotion Policy

At hiring (period t), the firm precommits to a policy

$$\Big\{W_{0,t}, W_{1,t}, \ldots, W_{N-1,t};\ \tau_{0,t}, \tau_{1,t}, \ldots, \tau_{N-2,t}\Big\},$$

where $W_{s,t}$ is the wage paid to the cohort hired at t when at level s (in period $t+s$), and $\tau_{s,t}$ is the dismissal rate (percentile cutoff) between level s and level $s+1$. In period $t+s$, a level-s worker is promoted to level $s+1$ for $t+s+1$ iff

$$\theta_{i,s,t+s} \ \geq\ \theta^{P}_{s,t}(\tau_{s,t}),$$

where $\theta^{P}_{s,t}(\tau)$ is the τ-quantile of the within-level signal distribution among level-s workers originally hired in period t—and thus working in period $t+s$. Workers who are not promoted in any given period are fired and cannot be hired by the firm in future periods.

C.4 Outside Options

At level s, a worker draws an outside option $U^{alt}_{i,s}$ from G^{alt}_{s} with density g^{alt}_{s}. Draws are i.i.d. across workers and independent across levels. New juniors observe $U^{alt}_{i,0}$ before joining and choosing effort; they do not observe future outside options when making effort.

For any random outside option $U^{alt}_{i,s}$ and employee value $U_{s,t}$ at level s, define

$$\bar{U}^{alt}_{s} \equiv \mathbb{E}[U^{alt}_{i,s}], \qquad \bar{\bar{U}}_{s,t} \equiv \mathbb{E}\Big[\max\{U_{s,t}, U^{alt}_{i,s}\}\Big] = G^{alt}_{s}(U_{s,t})U_{s,t} + \int_{U_{s,t}}^{\infty} u\, dG^{alt}_{s}(u).$$

C.5 Workers' Values and Incentives

Let $U_{s,t}$ denote the expected lifetime utility of a cohort-t worker who is at level s in period $t+s$ and remains with the firm. For the terminal level $s = N-1$, there is no further promotion, so $\tilde{e}_{N-1,t} = e^{*}_{N-1}$ and

$$U_{N-1,t} \ =\ \alpha_{N-1}W_{N-1,t} - C(e^{*}_{N-1}), \tag{OA.23}$$

where e^{*}_{N-1} is determined exogenously.[30] For any $s = 0, 1, \ldots, N-2$, a level-s worker chooses effort e_s to maximize

$$U_{s,t} \ =\ \max_{e_s}\ \alpha_s W_{s,t} - C(e_s) + \rho\left\{\left[1 - \Phi(\theta^{P}_{s,t}(\tau_{s,t}) - e)\right]\bar{\bar{U}}_{s+1,t} + \Phi(\theta^{P}_{s,t}(\tau_{s,t}) - e)\bar{U}^{alt}_{s+1}\right\}. \tag{OA.24}$$

In symmetric equilibrium, with within-level effort $\tilde{e}_{s,t}$, we have

$$\theta^{P}_{s,t}(\tau_{s,t}) = \tilde{e}_{s,t} + \kappa^{P}(\tau_{s,t}),$$

[30] See Footnote 14 for a discussion of possible values of e^{*}_{N-1}

where $\kappa^P(\tau)$ is the τ-quantile of Φ. The incentive compatibility (IC) conditions reduce to

$$C'(\tilde{e}_{s,t}) \;=\; \rho\,\phi(\kappa^P(\tau_{s,t}))\Big[\bar{\tilde{U}}_{s+1,t} - \bar{U}^{alt}_{s+1}\Big], \qquad s = 0, 1, \ldots, N-2, \tag{OA.25}$$

and $\tilde{e}_{N-1,t} = e^*_{N-1}$.

C.6 Employment Dynamics

The number of juniors willing to work for the firm is

$$L_{0,t} \;=\; G^{alt}_0(U_{0,t})\,\mathcal{L}. \tag{OA.26}$$

For $s = 0, 1, \ldots, N-2$, the next level's employment evolves as

$$L_{s+1,t+1} \;=\; (1 - \tau_{s,t-s})\,G^{alt}_{s+1}(U_{s+1,t-s})\,L_{s,t}. \tag{OA.27}$$

C.7 Firm's Problem

We specify the firm's problem recursively, which requires defining the state space. From (OA.26) and (OA.27), at any point in time, the stock of senior workers in the current period and the trajectory of those workers in future periods is pre-determined (through the firms' compensation-promotion choices at the moment those workers were hired). For the same reason, the effort levels of incumbent senior workers in the current and future periods are pre-contracted. The firm's state must account for all these employment and effort levels—both in the present and in the future. Formally, for $k \in \{0, \ldots, N-2\}$, let the vector $\vec{l}_k \equiv (l_{k+1,k}, \ldots, l_{N-1,k})$ collect the numbers of incumbent workers scheduled to work for the firm at levels $k+1$ to $N-1$ exactly k periods from the current one. Also, for $k \in \{0, \ldots, N-3\}$, let the vector $\vec{\epsilon}_k \equiv (\epsilon_{k+1,k}, \ldots, \epsilon_{N-2,k})$ collect the effort levels to be exerted by workers of level $k+1$ to $N-2$ exactly k periods from the current one.[31] Finally, define $l \equiv \{\vec{l}_k\}_{k=0}^{N-1}$ and $\epsilon \equiv \{\vec{\epsilon}_k\}_{k=0}^{N-2}$, for conciseness.[32] We can then write the firm's problem as choosing $\{W_s, \tau_s\}_{s=0}^{N-1}$ (with τ_{N-1} unused) to maximize the discounted value

$$V(l,\epsilon) \;=\; \max_{\{W_s,\tau_s\}_{s=0}^{N-1}} \left\{ \sum_{s=0}^{N-1} p_s\,F_s(L_0, \vec{l}_0, e_s) \;-\; W_0 L_0 \;-\; \sum_{s=1}^{N-1} \rho^s\,W_s\,L^{\text{cohort}}_s \;+\; \rho\,V(l^+,\epsilon^+) \right\}, \tag{OA.28}$$

[31] Recall that workers of level $N-1$ always set effort equal to e^*_{N-1}, so there is no need to keep track of their pre-contracted effort as a state variable.

[32] For example, if $N = 4$, we have: $\vec{l}_0 = (l_{1,0}, l_{2,0}, l_{3,0})$, $\vec{l}_1 = (l_{2,1}, l_{3,1})$, and $\vec{l}_2 = l_{3,1}$, which keep track of the pre-contracted senior employment levels in the current period, the next one, and the one after the next, respectively. Similarly, $\vec{\epsilon}_0 = (\epsilon_{1,0}, \epsilon_{2,0})$ and $\vec{\epsilon}_1 = \epsilon_{2,1}$ are the pre-contracted effort levels for non-terminal senior workers in the current period and in the next one.

where L_0 is given by (OA.26), the future employment path for the cohort hired this period is

$$L_s^{\text{cohort}} = \left(\prod_{m=0}^{s-1} (1-\tau_m)\, G_{m+1}^{alt}(U_{m+1}) \right) L_0 \qquad \text{for } s = 1, \ldots, N-1,$$

while e_s is $\epsilon_{s,0}$ for $s \in \{1, \ldots, N-2\}$ (e_{N-1}^* if $s = N-1$) and it is determined by (OA.25) if $s = 0$.[33] The stock of incumbent employed workers evolve as follows:

$$l_{s,k}^+ = \begin{cases} l_{s,k+1} & \text{if} \quad s \geq k+2, \\ L_s^{\text{cohort}} & \text{if} \quad s = k+1. \end{cases} \tag{OA.29}$$

Meanwhile, $\epsilon_{s,k}^+$ is equal to $\epsilon_{s,k+1}$ for $s \geq k+2$ and is given by (OA.25) if $s = k+1$. We then define $l^+ \equiv \{\vec{l}_k^+\}_{k=0}^{N-1}$ and $\epsilon^+ \equiv \{\vec{\epsilon}_k^+\}_{k=0}^{N-2}$.

C.8 Steady State

In a stationary equilibrium, $(W_{s,t}, \tau_{s,t}, L_{s,t}, \tilde{e}_{s,t}) = (W_s, \tau_s, L_s, \tilde{e}_s)$ for all s. Equations (OA.23)–(OA.27) specialize to

$$U_{N-1} = \alpha_{N-1} W_{N-1} - C(e_{N-1}^*), \tag{OA.30}$$

$$U_s = \alpha_s W_s - C(\tilde{e}_s) + \rho \left[(1-\tau_s) \bar{\tilde{U}}_{s+1} + \tau_s \bar{U}_{s+1}^{alt} \right], \quad s = 0, \ldots, N-2, \tag{OA.31}$$

$$C'(\tilde{e}_s) = \rho\, \phi(\kappa^P(\tau_s)) \left[\bar{\tilde{U}}_{s+1} - \bar{U}_{s+1}^{alt} \right], \quad s = 0, \ldots, N-2, \tag{OA.32}$$

$$L_0 = G_0^{alt}(U_0)\mathcal{L}, \qquad L_{s+1} = (1-\tau_s) G_{s+1}^{alt}(U_{s+1})\, L_s, \quad s = 0, \ldots, N-2. \tag{OA.33}$$

[33] Note that, as in Section 3.2.3, the firm pays immediately the present-discounted value of the entire wage bill of workers hired in the current period. We write the firm's problem in this manner to keep the state space as small as possible.

C.9 Incentive Contract Implications

We now show that the results from Proposition 1 in the main text generalize to the case of N employee seniority levels. For $s \geq 1$, the FOC of (OA.28) with respect to W_s is

$$\begin{aligned}
&\frac{\partial U_0}{\partial W_s} g_0^{alt}(U_0)\mathcal{L}\left\{MRPL_0 - W_0 - \sum_{k=1}^{N-1} \rho^k \prod_{m=0}^{k-1} \left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right](MRPL_k - W_k)\right\} \\
&\quad + \sum_{j=1}^{s} \rho^j \prod_{m=1}^{j} \left[(1-\tau_{m-1})G_{m-1}^{alt}(U_{m-1})\right] \mathcal{L}\frac{\partial U_j}{\partial W_s} g_j^{alt}(U_j)\Bigg\{MRPL_j - W_j \\
&\quad + \sum_{k=j+1}^{N-1} \rho^{k-j} \prod_{m=j}^{k-1} \left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right](MRPL_k - W_k)\Bigg\} \\
&\quad - \rho^s \prod_{j=0}^{s-1} \left[(1-\tau_j)G_{j+1}^{alt}(U_{j+1})\right] G_0^{alt}(U_0)\mathcal{L} \\
&\quad + \sum_{j=0}^{s-1} \frac{\partial F_j}{\partial \tilde{e}_j}\frac{\tilde{e}_j}{\partial W_s} = 0. \qquad \text{(OA.34)}
\end{aligned}$$

The term in the first line captures the effect of an increase in W_s on junior workers' entry decisions. Because entrants may remain with the firm until retirement, the wedge between $MRPL$ and compensation must be accounted for at all levels from 0 to N-1. The terms in the second and third lines capture the effect of W_s on the retention decisions of workers at levels 1 through s; for levels above s, changes in W_s do not affect the decision to stay. The fourth-line term is the marginal effect of W_s on the discounted payroll. Finally, the fifth-line term captures the incentive effect of W_s on effort for workers at levels 0 through $s-1$; incentive effects are zero for workers at level s and above.

As for W_0, the the FOC of (OA.28) is

$$\frac{\partial U_0}{\partial W_0} g_0^{alt}(U_0)\mathcal{L}\left\{MRPL_0 - W_0 + \sum_{s=1}^{N-1} \rho^s \prod_{m=0}^{s-1} \left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right](MRPL_s - W_s)\right\} = L_0. \qquad \text{(OA.35)}$$

Meanwhile,

$$\frac{\partial U_s}{\partial W_s} = \alpha, \quad \text{for } s \geq 1 \qquad \text{(OA.36)}$$

$$\text{and} \quad \frac{\partial U_k}{\partial W_s} = \rho^{(s-k)} \prod_{j=k}^{s-1} \left[(1-\tau_j)G_{j+1}^{alt}(U_{j+1})\right]\alpha, \quad \text{for } s > k \geq 1. \qquad \text{(OA.37)}$$

From (OA.33), (OA.35) and (OA.36), we obtain

$$g_0^{alt}(U_0)\mathcal{L}\left\{MRPL_0 - W_0 - \sum_{k=1}^{N-1} \rho^k \prod_{m=0}^{k-1} \left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right](MRPL_k - W_k)\right\} = \frac{G_0^{alt}(U_0)\mathcal{L}}{\alpha}.$$

Together with (OA.37), this allows us to rewrite the first line of (OA.34) as

$$\rho^s \prod_{j=0}^{s-1} \left[(1-\tau_j) G_{j+1}^{alt}(U_{j+1})\right] G_0^{alt}(U_0)\mathcal{L},$$

which is the additive inverse of the fourth line of the same equation. Hence, for every $s \in \{1,\ldots,N-1\}$, the first and fourth lines of (OA.34) cancel each other out, and we can rewrite the FOC as

$$\begin{aligned}\sum_{j=1}^{s} \rho^j \prod_{m=1}^{j} \left[(1-\tau_{m-1}) G_{m-1}^{alt}(U_{m-1})\right] \mathcal{L} \frac{\partial U_j}{\partial W_s} g_j^{alt}(U_j) \Bigg\{ MRPL_j - W_j \\ + \sum_{k=j+1}^{N-1} \rho^{k-j} \prod_{m=j}^{k-1} \left[(1-\tau_m) G_{m+1}^{alt}(U_{m+1})\right] (MRPL_k - W_k) \Bigg\} \\ + \sum_{j=0}^{s-1} \frac{\partial F_j}{\partial \tilde{e}_j} \frac{\tilde{e}_j}{\partial W_s} = 0. \end{aligned} \tag{OA.38}$$

We can now assess how the need to incentivize workers affects equilibrium wages. We consider two scenarios—first assuming that workers' effort play no role in production, and then assuming that effort positively affects output.

C.9.1 Effort if Irrelevant for Production

If $\partial F_s / \partial \tilde{e}_s = 0$ for all s, the term in the third line of (OA.38) vanishes. Evaluating the condition for $s = 1$, we obtain

$$\rho(1-\tau_0) L_0 \frac{\partial U_1}{\partial W_1} g_1^{alt}(U_1) \Bigg\{ MRPL_1 - W_1 + \sum_{k=2}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G_{m+1}^{alt}(U_{m+1})\right] (MRPL_k - W_k) \Bigg\} = 0.$$

Since the term outside the braces is strictly positive, it follows that

$$MRPL_1 - W_1 + \sum_{k=2}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G_{m+1}^{alt}(U_{m+1})\right] (MRPL_k - W_k) = 0. \tag{OA.39}$$

That is, the present-discounted value of the wedge between $MRPL$ and compensation from level $s = 1$ through retirement is zero.

Next, evaluate (OA.38) for $s = 2$, which yields

$$\rho(1-\tau_0)L_0\frac{\partial U_1}{\partial W_2}g_1^{alt}(U_1)\Bigg\{MRPL_1 - W_1 + \sum_{k=2}^{N-1}\rho^{k-1}\prod_{m=1}^{k-1}\left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right](MRPL_k - W_k)\Bigg\}$$
$$+\rho^2(1-\tau_0)(1-\tau_1)L_0G_1(U_1)\frac{\partial U_2}{\partial W_2}g_2^{alt}(U_2)\Bigg\{MRPL_2 - W_2$$
$$+\sum_{k=3}^{N-1}\rho^{k-2}\prod_{m=2}^{k-1}\left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right](MRPL_k - W_k)\Bigg\} = 0.$$

By (OA.39), the first term is zero. Since the remaining prefactor is strictly positive, we obtain

$$MRPL_2 - W_2 + \sum_{k=3}^{N-1}\rho^{k-2}\prod_{m=2}^{k-1}\left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right](MRPL_k - W_k) = 0.$$

Thus, starting from $s = 2$, the present-discounted value of the $MRPL$–compensation wedge until retirement is also zero. Applying this argument iteratively up to $s = N-1$ implies $MRPL_{N-1} = W_{N-1}$. Moving backward, since the present-discounted wedge is zero at every level, we conclude that $MRPL_s = W_s$ for every $s \geq 1$.

C.9.2 Productive Effort

Now assume that $\frac{\partial F_0}{\partial \tilde{e}_0} > 0$. Evaluating (OA.38) at $s = 1$ gives

$$\rho(1-\tau_0)L_0\frac{\partial U_1}{\partial W_1}g_1^{alt}(U_1)\Bigg\{MRPL_1 - W_1 + \sum_{k=2}^{N-1}\rho^{k-1}\prod_{m=1}^{k-1}\left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right](MRPL_k - W_k)\Bigg\} < 0,$$

from which it follows that

$$MRPL_1 + \sum_{k=2}^{N-1}\rho^{k-1}\prod_{m=1}^{k-1}\left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right]MRPL_k < W_1 + \sum_{k=2}^{N-1}\rho^{k-1}\prod_{m=1}^{k-1}\left[(1-\tau_m)G_{m+1}^{alt}(U_{m+1})\right]W_k.$$

That is, from the perspective of a type-1 worker, the present-discounted value of compensation until retirement exceeds than that of the $MRPL$.

C.10 Entry Elasticity and the Exact Intertemporal Wedge

Here we generalize the results from Proposition 2 to the N-types case.

Define the firm-specific *entry* (junior) labor-supply elasticity

$$\psi_0 \equiv \frac{g_0^{alt}(U_0)\,W_0}{G_0^{alt}(U_0)}\cdot\frac{\partial U_0}{\partial W_0}, \qquad \frac{\partial U_0}{\partial W_0} = \alpha_0, \tag{OA.40}$$

where we normalize $\alpha_0 = 1$.

Also, define the steady-state *marginal revenue product of labor* of a type-s worker as

$$MRPL_s \equiv \sum_{j=0}^{N-1} p_j \frac{\partial F_j}{\partial L_s}.$$

From the envelope condition (and the iterated application of the chain rule), the steady-state derivatives of the value function with respect to the pre-contracted stocks of senior labor conveniently simplify to

$$\frac{\partial V(l,\epsilon)}{\partial l_{s,k}} = \rho^k MRPL_s, \quad k \in \{0,\dots,N-2\}, \quad s \in \{k+1,\dots,N-1\}. \tag{OA.41}$$

Substituting (OA.33) and (OA.40) on (OA.35), we obtain

$$\begin{aligned} MRPL_0 + \sum_{s=1}^{N-1} \rho^s \prod_{m=0}^{s-1} \left[(1-\tau_m) G^{alt}_{m+1}(U_{m+1})\mathcal{L}\right] MRPL_s \\ = W_0 + \sum_{s=1}^{N-1} \rho^s W_s \prod_{m=0}^{s-1} \left[(1-\tau_m) G^{alt}_{m+1}(U_{m+1})\mathcal{L}\right] + \frac{W_0}{\psi_0}. \end{aligned} \tag{OA.42}$$

Now define the cohort-level NPVs (from the firm's perspective) for a new hire:

$$\text{NPV}_W \equiv W_0 + \sum_{s=1}^{N-1} \rho^s W_s \prod_{m=0}^{s-1} \left[(1-\tau_m) G^{alt}_{m+1}(U_{m+1})\mathcal{L}\right], \tag{OA.43}$$

$$\text{NPV}_{\text{MRPL}} \equiv MRSP_0 + \sum_{s=1}^{N-1} \rho^s \prod_{m=0}^{s-1} \left[(1-\tau_m) G^{alt}_{m+1}(U_{m+1})\mathcal{L}\right] MRPL_s. \tag{OA.44}$$

Then (OA.42) implies the exact wedge

$$\text{NPV}_{\text{MRPL}} = \text{NPV}_W + \frac{W_0}{\psi_0}. \tag{OA.45}$$

By defining the long-run elasticity of labor supply as

$$\psi_{LR} = \psi_0 \frac{NPV_W}{W_0} \tag{OA.46}$$

and substituting on (OA.45), we obtain an equation equivalent to that in the statement of Proposition 2.

D Tournament Model with Lateral Hiring

We extend the tournament model from Section 3 to allow for lateral hiring of senior employees. For tractability, we continue to focus on an environment with two seniority levels—junior and senior. In addition to promoting juniors from the previous period, the firm can now hire senior workers directly from the external labor market in each period.

D.1 Primitives

The environment is otherwise identical to that in Section 3, including workers' preferences, the production technology, and time discounting. The key new feature is that, in every period t, there are $\mathcal{H}$ senior workers available for lateral hiring, in addition to the mass $\mathcal{L}$ of new junior workers entering the labor market and the workers promoted in period $t-1$.

Let $l_{2,t}$ denote the mass of workers promoted in period $t-1$ (who attain senior status in period t), and let lat_t denote the mass of senior workers hired laterally in period t. The total number of senior workers employed by the firm in period t is therefore

$$L_{2,t} = l_{2,t} + lat_t.$$

Lateral hires in period t receive a flat wage $W_{lat,t}$; that is, their compensation does not depend on performance, and the firm cannot use incentive pay to elicit effort from them. Thus, s laterally hired individual employed by the firm in period t has utility As in Section 3, junior workers hired in period t are offered a compensation–promotion policy $\{W_{1,t}, W_{2,t}, \tau_t\}$.

Potential lateral hires have outside-option utility distributed according to G^{alt}_{lat}, with density g^{alt}_{lat}. This distribution may differ from that of promoted seniors, whose outside-option utility is distributed according to G^{alt}_2. We maintain the assumption that outside-option draws are independent across workers.

D.2 Worker and Firm Behavior

Junior workers' effort choices, given the compensation–promotion policy offered by the firm, continue to be characterized by (5). For simplicity, we assume that all senior workers—regardless of whether they were promoted internally or hired laterally—exert effort $\underline{e}$. The resulting flow utilities in period t are

$$U_{2,t} = \alpha W_{2,t-1} - C(\underline{e}) \quad \text{and} \quad U_{lat,t} = \alpha W_{lat,t} - C(\underline{e}),$$

for promoted and laterally hired seniors, respectively, reflecting that promoted seniors' wages are determined upon promotion in period $t-1$, whereas lateral hires are contracted in period t.

The number of junior workers hired by the firm remains given by (6). The total number of senior workers employed in period t is now

$$L_{2,t} = (1-\tau_{t-1}) G^{alt}_2(U_{2,t-1})\, L_{1,t-1} + G^{alt}_{lat}(U_{lat,t})\, \mathcal{H}.$$

The first term captures seniors promoted in the previous period who both win the promotion tournament and accept the senior position, while the second term represents senior workers hired laterally from the external market.

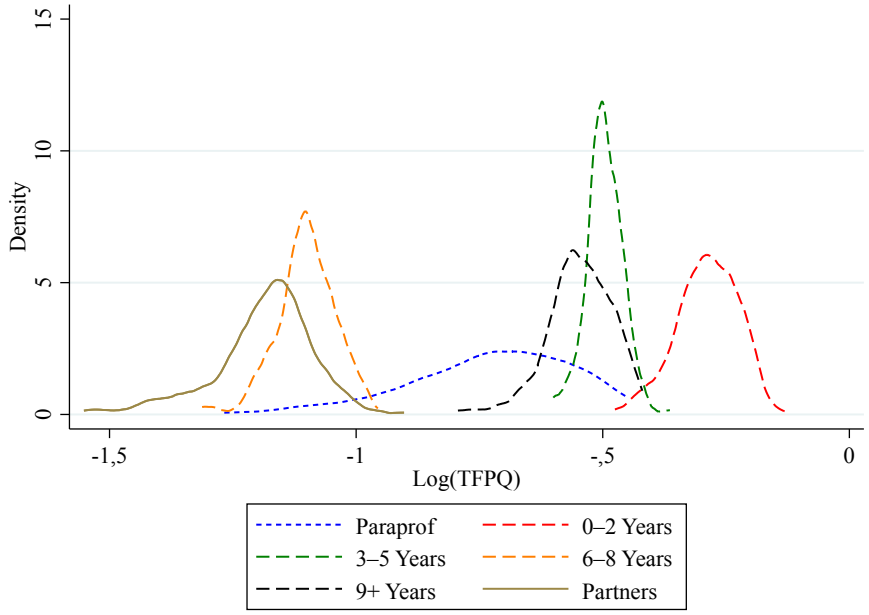
E Appendix Tables and Figures

Table OA-3: List of Symbols and Definitions

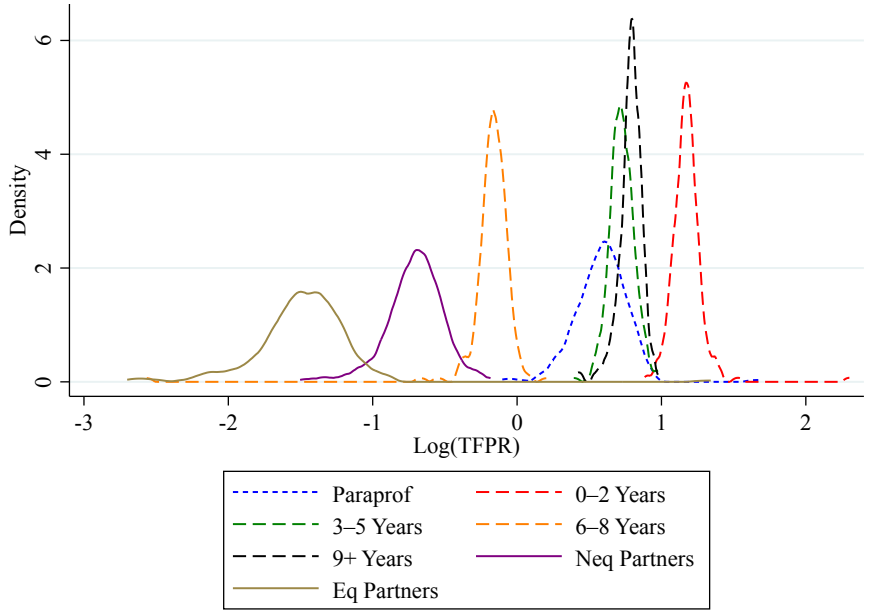
Symbol	Definition / Meaning
Demand, Revenues, and Costs	
f	Firm index
i	Employee types for billing-revenue-producing inputs (e.g. professional staff 0–2 yrs, 3–5 yrs, etc.)
Q_{if}	Billable hours produced by employees of type i in firm f
P_{if}	Billing rate (price per billable hour) for employee type i
$R_{if} = P_{if}Q_{if}$	Revenue generated by employee type i
L_{if}	Hours worked by employee type i
W_{if}	Hourly wage (or equivalent rate) for employee type i
MC_{if}	Marginal cost of providing accounting services of type i
$\mu_{if} = \frac{P_{if}-C_{if}}{C_{if}}$	Price markup for accountant type i
η	Demand elasticity
ζ_{if}	Firm-level demand residual
$\text{MRPL}_{if} = \frac{\partial R_f}{\partial L_{if}}$	Marginal revenue product of labor for worker type i
$\delta_{if} = \frac{W_{if}}{\text{MRPL}_{if}}$	Wage markdown (ratio of wage to MRPL)
$w_{if}l_{if}$	Total wage bill for workers of type i
Production Function	
Ω_{if}	Productivity level for employee type i in firm f
ε_{if}	Measurement error in output
$\boldsymbol{\beta}^l = \{\beta^l_{ik}\}$	Output elasticities (own and cross) across employee types for billing-revenue-producing workers
$\boldsymbol{\beta}^v$	Output elasticities with respect to other inputs V
$\boldsymbol{V}_f$	Vector of non-billing-revenue producing inputs (interns, materials, tech, etc.)
$Q_{if} = \Omega_{if}e^{\varepsilon_{if}} \prod_k L_{kf}^{\beta_{ik}} \prod_m V_{mf}^{\gamma_{im}}$	Cobb–Douglas production with team effects
ν_{if}	Innovation in productivity
σ_i	Serial correlation of productivity for type i
Dynamic Labor Market / Tournament Model	
t	Time period
$e_{i,t}$	Effort exerted by worker i in period t
$C(e)$	Cost of effort (increasing and convex)
$\theta_{i,t} = e_{i,t} + \kappa_{i,t}$	Noisy performance signal
$\kappa_{i,t} \sim \Phi(\cdot)$	Noise term with CDF Φ, density ϕ
τ_t	Promotion cutoff: share of juniors not promoted
$\theta_t^P(\tau_t)$	Promotion threshold corresponding to τ_t
$U_{i,s}^{alt}$	Worker's outside option (utility)
G_s^{alt}, g_s^{alt}	CDF and PDF of outside options for type-s workers
ρ	Discount factor
α_i	Utility scaling factor on wages
$U_{s,t}$	Expected utility of worker of type s
$\bar{U}_2^{alt}$	Expected value of outside option for seniors
$\bar{U}_{2,t}$	Expected maximum of $U_{2,t}$ and outside options
$\mathcal{L}$	Mass of potential new junior workers per period
$V(l_2)$	Firm's value function given senior employment l_2
$V'(l_2)$	Marginal value of senior labor to the firm
MRPL_2	Marginal revenue product of senior labor
$\psi_1 = \frac{g_1^{alt}(U_1)W_1}{G_1^{alt}(U_1)}$	Firm-specific (short-run) elasticity of junior labor supply
ψ_{LR}	Long-run (intertemporal) labor-supply elasticity
NPV_W	NPV of wages for a new hire
NPV_{MRPL}	NPV of marginal revenue product for a new hire

Figure OA-5: Productivity Dispersion within Firms: TFPQ vs. TFPR

(a) log(TFPQ)



(b) log(TFPR)



Notes: Panel (a) shows the distributions of log(TFPQ) by employee type, panel (b) shows the distributions of log(TFPR).

Regularity

Assumption OA-1 (Regularity and interiority)**.** *The cost function $C(\cdot)$ is twice continuously differentiable with $C' > 0$ and $C'' > 0$. The noise density $\phi(\cdot)$ is continuous and strictly positive in a neighborhood of $\kappa^P(\tau)$. The outside-option distributions G_1^{alt} and G_2^{alt} admit continuous densities $g_1^{alt} > 0$ and $g_2^{alt} > 0$ on the relevant supports. All choice variables admit interior optima unless stated otherwise.*

Employment responses to W_2. In steady state,

$$L_1 = \mathcal{L}\, G_1^{alt}(U_1), \qquad U_1 = \alpha_1 W_1 - C(\tilde{e}) + \rho\big[(1-\tau)\bar{\bar{U}}_2 + \tau \bar{U}_2^{alt}\big],$$

and

$$L_2 = (1-\tau) G_2^{alt}(U_2)\, L_1, \qquad U_2 = \alpha_2 W_2 - C(e_2^*).$$

Because U_1 depends on $\bar{\bar{U}}_2$, changes in W_2 generally affect both L_1 and L_2.

References for Online Appendix

De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016). Prices, markups, and trade reform. *Econometrica 84*(2), 445–510.

Inside Public Accounting (2023). 2023 Human Resources Report.