Monopsony and Backloaded Compensation: Theory and Evidence from Public Accountants

Michael Rubens* Bernardo Silveira†

March 26, 2026

Abstract

In monopsony models, wage markdowns induce deadweight loss and are therefore inefficient. Yet markdowns also arise in models with backloaded efficiency pay, where they are designed to induce effort and are thus efficient. These frameworks are non-nested, assuming either exogenous effort or perfectly competitive labor markets. To bridge this gap, we build a model incorporating both imperfect labor market competition and endogenous effort—and use it to empirically distinguish the two forces. Estimating a team production model on novel, uniquely rich data covering 569 U.S. public accounting firms, we find that these firms carry substantial monopsony power: lifetime wages are marked down by 27% below lifetime MRPL. At the same time, short-run markdowns are even larger for junior employees, while senior employees are paid above their marginal revenue product—indicating the use of backloaded efficiency wage contracts.

Keywords: Monopsony, Wage Markdowns, Team Production, Implicit Contracts, Dynamics, Effort Incentives, Tournaments

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, Moritz Meyer-ter-Vehn, Charles Angelucci, and participants at the MIT Organizational Economics Lunch for useful feedback.

Disclosure Statement: Neither Michael Rubens or Bernardo Silveira have any financial interests, conflicts of interest, or research support to disclose that are relevant to this paper.

1 Introduction

There is increasing evidence of monopsony power in labor markets (Card, 2022; Azar and Marinescu, 2024; Caldwell et al., 2025; Kline, 2025). Monopsony models usually focus on how market power distorts how much labor is employed, but treat the effort exerted by employees as exogenous (Card et al., 2018; Berger et al., 2022). This implies that the wage 'markdown' below the marginal revenue product of labor is inefficient, as it induces deadweight loss and misallocation (Berger et al., 2022; Lamadon et al., 2022). In contrast, the literature on backloaded incentive payments features wage markdowns as a means to incentivize effort of forward-looking employees, while assuming competitive labor markets (Lazear, 1981; Lazear and Rosen, 1981; Malcomson, 1984). In these models, markdowns 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. To this end, we construct a partial-equilibrium theoretical model incorporating both forces, tailored to the context of professional services industries—such as law firms and consultancies—which are characterized by 'up-or-out' promotion policies. The model combines 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—if offered promotion—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.

Solving the model in steady state reveals two novel results on wage markdowns under dynamic incentive contracts, which can be tested empirically. First, we show that 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—and that this 'lifetime' wage markdown is a function of the residual *long-run* labor supply elasticity faced by 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.

Second, if effort is endogenous, firms pay junior employees below their marginal revenue product but senior employees above it. Strikingly, senior employees benefit from a

wage markup even though the firm retains market power in the senior labor market. In contrast, when effort is exogenous (or irrelevant for production), the firm marks down junior wages but pays senior employees exactly their marginal revenue product—again, despite the presence of monopsony power in that market. Thus, whether seniors' compensation differs from their marginal revenue product is determined entirely by the incentive structure, not by the degree of monopsony power.

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 over 500 U.S. public accounting firms, which includes billing hours, billing rates, worked hours, and hourly compensation for seven categories of workers that differ in terms of seniority and role within the firm. Our empirical strategy focuses on estimating wage markdowns for these different 'worker types' 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 type, and for team production. The team production feature is important because it captures, for instance, the effects of partners' client-attraction efforts on junior billing revenue, or mentoring externalities within the firm. As is common in the empirical production functions literature, we rely on timing assumptions on the variable and fixed production inputs for identification (Olley and Pakes, 1996; Ackerberg et al., 2015).

Using our estimated team production model, we compute marginal revenue products of labor for each worker type, which we use 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 employee wages are marked down by less than half of their marginal revenue product, whereas employees start earning substantial wage mark*ups* after approximately 15 years of service. This indicates backloaded incentive pay contracts are in use.[1,2]

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

[1] 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.

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

market competition, in which the NPV of wages equals the NPV of the MRPL.[3] 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. That is, 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 it remains far from perfectly elastic, firms still exert monopsony power.

Finally, we examine by how much junior wages could increase in order to achieve the perfect competition scenario in which the NPV of wages equals the NPV of the MRPLs, keeping senior wages and promotion rates (and hence, the effort incentives) fixed. We find that wages of non-partner employees could increase by 55% on average. As a result, the wage markdown for entry-level employees would shrink from 0.55 (wages 45% below MRPL) to 0.83 (wages 17% below MRPL). This exercise suggests that 62% of the junior wage markdown is due to monopsony power and 38% due to incentive-wage backloading.

Our empirical findings carry several important implications. First, estimated instantaneous wage markdowns no longer reflect the true extent of monopsony power in the presence of implicit contracts, so one needs to estimate the lifecycle markdown of the NPV of wages below the NPV of the marginal revenue product instead. Second, not all markdown dispersion is necessarily inefficient, as wage markdowns are in part designed to induce effort by junior employees—in contrast to spot market monopsony models, in which wage markdown variation necessarily leads to misallocation (Berger et al., 2022). Third, policies that restrict markdown dispersion within firms, such as pay-ratio caps, can either increase or decrease welfare depending on the relative magnitudes of the monopsony and incentive-wage mechanisms in driving wage markdowns.

This paper relates to five 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).[4] We contribute to this literature by incorporating

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

[4] Different strands of dynamic monopsony models with search frictions also feature wage markdowns in equilibrium (Burdett and Mortensen, 1998; Manning, 1987; Gottfries and Jarosch, 2023). Also related are

endogenous effort of forward-looking workers into the workhorse monopsony model. We derive a dynamic analogue to the static markdown rule of Card et al. (2018), linking long-run labor supply elasticities to lifetime wage markdowns. Another difference relative to the monopsony literature is that it has focused predominantly on lower-wage occupations and industries (Card and Krueger, 1994), whereas we study monopsony power in a high-wage industry, public accounting, which shares many features with other professional services industries, such as law firms and consultancies.

Second, we contribute to the literature on implicit labor contracts with endogenous effort, featuring backloaded pay as an incentive device (Lazear, 1981; Malcomson, 1984; Medoff and Abraham, 1980). We extend the tournament model by accounting for monopsony power, thereby generating lifetime wage markdowns. We further introduce nonlinear, multi-output production with productivity externalities across various worker types—a feature that is crucial for bringing the model to the data.

Third, we contribute to the prior literature on estimating wage markdowns using production models (Kotlikoff and Gokhale, 1992; Hellerstein et al., 1999; Morlacco, 2019; Brooks et al., 2021; Yeh et al., 2022; Mertens, 2022; Delabastita and Rubens, 2025). Whereas this literature has relied on single-output production functions, we estimate a multi-output production model allowing for team production and output differentiation by worker seniority. Moreover, while the production-based markdown literature has mainly focused on manufacturing industries, we extend this approach to professional services firms.[5]

Fourth, we relate to the literature on multi-product production models (De Loecker et al., 2016; Dhyne et al., 2017; Orr, 2022; Valmari, 2023). Our approach to estimating the multi-output production model involves estimating a full matrix of own- and cross-output elasticities across worker types, rather than the own-output elasticities typical of single-output models. We follow Dhyne et al. (2017) in treating inputs as nonrival within the firm, thereby avoiding the need to assign input usage to individual product lines, an approach that fits naturally with production in professional services firms.

Finally, this paper relates to prior work on competition and market power in auditing and accounting services (Numan and Willekens, 2012; Gerakos and Syverson, 2015). Whereas this literature has focused mainly on product market power of auditing firms, we highlight and quantify their monopsony power in labor markets.

The remainder of this paper is structured as follows. We start by describing the

models of implicit labor contracts such as Balke and Lamadon (2022).

[5] In contrast to prior empirical work on team production (Jarosch et al., 2021; Herkenhoff et al., 2024; Bonhomme, 2021), we do not rely on individual worker wage data, but rather on output, input, and price data aggregated by worker 'types' of different seniorities. Whereas this has the benefit of providing 'conduct-free' markdown estimates by worker type, it comes at the cost of assuming away within-type worker heterogeneity.

public accounting industry, our dataset, and key facts on billing-rate-to-salary spreads and promotion rates by worker seniority in Section 2. In Section 3, we set up a tournament model with monopsonistic competition and nonlinear multi-output production, and derive two empirical hypotheses on efficiency wages and monopsony power. In Section 4, we test these hypotheses for the U.S. public accounting industry by estimating wage markdowns using our production model, and convert 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" (Deloitte, EY, KPMG, PwC), it contains many smaller firms with collectively significant market shares. For instance, SEC filings show that the Big Four together accounted for 48.4% of public-company audits in 2024 (Ideagen Audit Analytics, 2024). As explained in Section 2.2, our analysis focuses on non–Big Four firms due to data availability. Accordingly, unless otherwise noted, our description of the public accounting industry below refers to this group.[6] Public accounting firm revenue comes from 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—most of whom hold Certified Public Accountant (CPA) licenses—who follow an 'up-or-out' career path that can culminate in 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 drawn 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.

Two-thirds of firms set prices using a billing-rate system, charging clients staff-specific hourly billing rates multiplied by billable hours. One-quarter of firms 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

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

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 do. Equity partners earn on average around twice as much as non-equity partners. Most firms impose a mandatory retirement age for equity partners (Inside Public Accounting, 2023b).

2.2 Data

Our dataset consists of annual surveys of U.S. public accounting firms, conducted by Inside Public Accounting from 2015 to 2024. The number of firms surveyed ranges from 484 to 569, depending on the year. The survey does not include the Big Four companies. For confidentiality reasons, we do not observe firm-level data; instead, we observe averages for firm bins grouped by location and size, typically containing around 20 firms, with the smallest bin containing 5. 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, though their compensation and billing rates are available. We deflate all monetary variables using the CPI for urban consumers, averaged across 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 69.5M. Firms employ on average 239 employees, of whom 11% are equity partners and 55% hold CPA licenses. Firms spend 48% of revenue on labor, 4.5% on technology, 2% on marketing, and 1% each on training and recruiting. Firm capital is on average around 19% 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	

Notes: Summary statistics on the IPA dataset, 2015–2024.

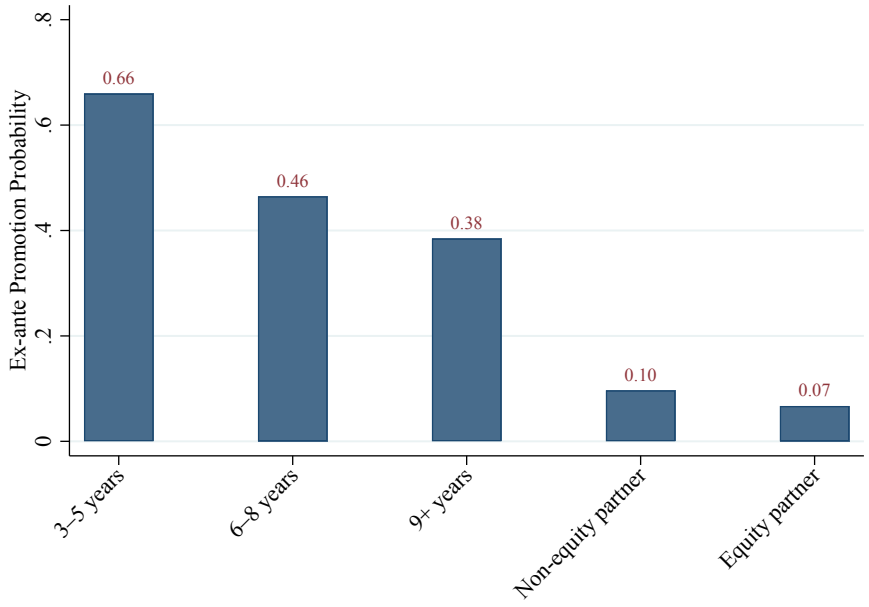
2.3 Stylized Facts

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

Public accounting firms use an 'up-or-out' promotion policy, as in many other professional services industries. We compute promotion probabilities by assuming a steady-state labor force at the industry level, while allowing these probabilities to differ across small, medium-sized, and large firms.[8] Figure 1 shows the resulting promotion probabilities for

[8]Details on the computation of promotion probabilities are in Appendix B.3.

Figure 1: Cumulative Promotion Probabilities



Notes: Bars indicate the cumulative probabilities of reaching every grade i for an employee with 0-2 years of experience. Details on computation are in Appendix B.3.

a career starter. 34% of starters exit after 2 years, another 20% exit after 5 years, and so on. The ex-ante probability of becoming a non-equity partner is 10%, whereas only 7% of starters eventually become equity partners. There is a sharp drop in the promotion probabilities right before reaching the non-equity partnership.

Labor markets are largely internal, with very low lateral mobility between public accounting firms (Inside Public Accounting, 2023a). Workers who are not promoted to the partnership leave the firm, either voluntarily or involuntarily, and usually transition to in-house accounting positions in other industries, rather than partnerships at competing public accounting firms. IPA reports that only around 14% of new equity partners are lateral hires rather than in-house promotions (Inside Public Accounting, 2023b).

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

We compute the spread between compensation and billing revenue for employees of seven different seniority levels, denoted by s:[9]

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

[9]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.

This spread measures the labor share of revenue by seniority level. If one were to assume independent and linear production by each worker i and no product market power, this margin would equal the wage markdown (or markup), $\delta_s \equiv W_s/MRPL_s$.[10]

Figure 3 shows the salary-billing rate spreads for the various employee levels. Upon joining the firm, professional staff in the 'partner-track' are paid on average 30% of their respective billing revenue. As seniority increases, this compensation rate slightly raises, up to 37% for staff with 9+ years of experience. Non-equity partners are paid 62% of the revenue they generate. In contrast, equity partners receive 138% of their billing revenue. Finally, paraprofessionals, who are outside the 'partner track' are paid on average 38% of their billing revenue.

However, three complications prevent a straightforward interpretation of these spreads. First, in reality, production is likely nonlinear, so average and marginal products of labor are not necessarily identical. Second, firms likely have product market power, which is another reason why marginal and average revenue products differ. Third, even if these spreads would identify wage markdowns, these could still either reflect monopsony power or backloaded compensation.

Fact 3: Billing Hours Decrease, Worked Hours Increase with Seniority

Finally, we show the average hours worked and billed by seniority for the 'partnership-track' employees, in Figure 2. As employees get more senior, they work more hours, but get billed less (except when moving from 0-2 to 3-5 years, which might be due to initial training). This suggests that team production is important: although partners do not directly generate that many billing hours, many of their worked hours likely generate billing hours indirectly through, for instance, business development and mentoring. Therefore, taking into account team production is a crucial ingredient into our model.

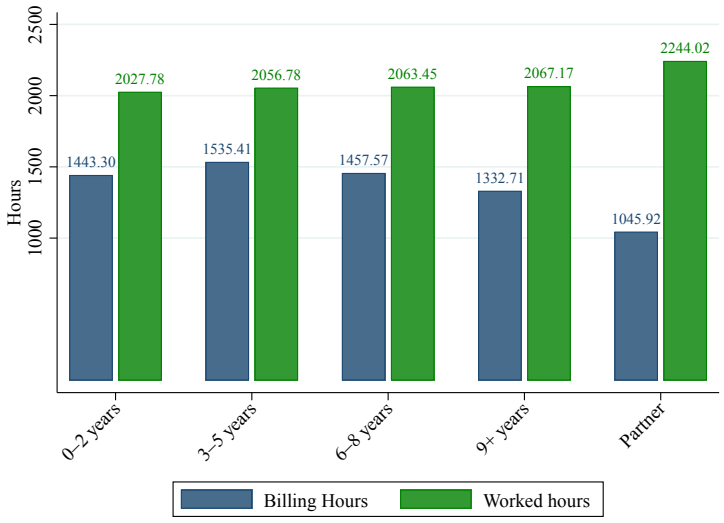
Roadmap

Taking stock, we want a model that incorporates both monopsony power and backloaded incentive pay, as both these forces can lead to wage markdowns below marginal revenue products of labor. The facts presented above suggest that key ingredients to incorporate into our model are (i) a tournament incentive system with forward-looking employees, (ii) nonlinear production and market power, and (iii) team production.

We start by combining these elements into a general model of production and labor supply in Section 3. Next, we bring this model to the data in Section 4.

[10] Under this assumption, production for type-s services would be $Q_s = a_s L_s$, implying $(W_s/MRPL_s) = (W_s L_s)/(P_s Q_s)$, where W_s denotes hourly wages, P_s billing rates, L_s hours worked, and Q_s billable hours.

Figure 2: Billed vs. Worked Hours, by Seniority



3 Model: Monopsony and Backloaded Compensation

We develop a partial-equilibrium 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.

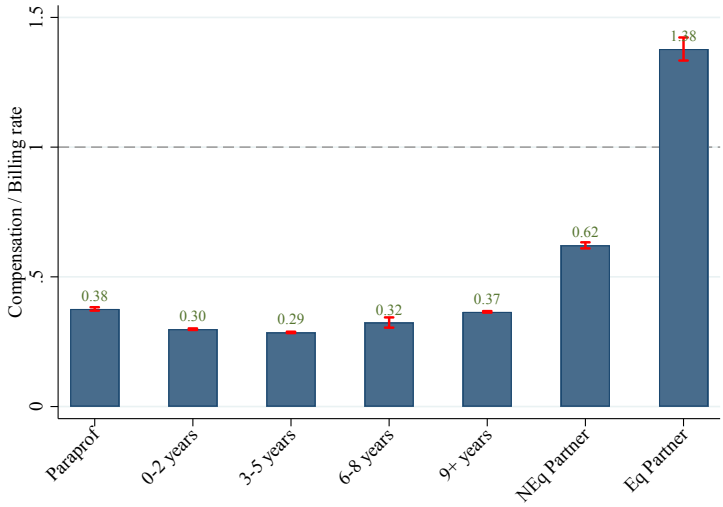
To further simplify the exposition, we assume that both junior and senior workers are employees, and that wages and promotion policies are chosen by a residual claimant who does not participate in the tournament. Section 3.6.1 discusses how to interpret the model when senior workers are themselves the 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, $\rho < 1$. Firms, indexed by $f \in \{1, \ldots, F\}$, are infinitely-lived. 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

Figure 3: 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.

the beginning of $t+2$.

Each worker i employed by firm f in period t exerts effort $e_{i,f,t} \in [\underline{e}, \infty)$, incurring cost $C(e_{i,f,t})$ with $C'(\cdot) > 0$ and $C''(\cdot) > 0$. Employees' flow utility increases linearly in wages (with coefficient α) and decreases in effort cost. Thus, if firm f pays compensation $W_{s,f,t}$ to its type-s workers in period t, the flow utility of worker i is

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

where $\xi_{s(i,t),f}$ and $\epsilon_{i,f,t}$ represent non-pecuniary aspects of worker i's utility from working at firm f in period t—such as the firm's amenities. The distinction between the two terms is that $\xi_{s(i,t),f}$ affects all type-$s(i,t)$ workers equally and is constant over time, whereas $\epsilon_{i,f,t}$ is idiosyncratic to worker i and period t. Section 3.1.3 presents further distributional assumptions that we impose on $\xi_{s(i,t),f}$ and $\epsilon_{i,f,t}$.[11] For simplicity, we abstract away from random wage coefficients in the workers' flow utility in this section, so α is treated as a constant. However, the empirical model in Section 4 allows α to vary randomly across firms. Additionally, the empirical model accommodates temporal variation in the firm-level non-pecuniary term, $\xi_{s(i,t),f}$.

[11] Previous studies typically assume that $\epsilon_{i,f,t}$ follows a type I extreme value distribution; see, for example, Card et al. (2018). Such an assumption is not necessary for our analysis, as the empirical tests we develop do not require a closed-form solutions for the workers' decisions of where to work.

Firms observe a noisy, non-contractible signal of effort of each of their workers,

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

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

3.1.2 *Production and Compensation Policy*

Each firm employs both juniors and seniors. Let $L_{s,f,t}$ and $\bar{e}_{s,f,t}$ denote the employment level and average effort of type-s workers in firm f. The firm produces two outputs,

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

where $L_{-s,f,t}$ and $\bar{e}_{-s,f,t}$ denote the employment and effort of the other type. In our empirical application, $Q_{s,f,t}$ corresponds to the billing hours produced by workers of seniority level $s \in \{1,2\}$. Effort weakly increases production; specifically, $\partial F_s / \partial \bar{e}_{k,f,t} \geq 0$ for all $s, k \in \{1,2\}$. To lighten notation, we define $\mathfrak{I}_{s,f,t} \equiv (L_{s,f,t}, L_{-s,f,t}, \bar{e}_{s,f,t}, \bar{e}_{-s,f,t})$, which allows us to write firm f's output of type s in period t as $F_s(\mathfrak{I}_{s,f,t})$.

At the time of hiring junior workers, firm f commits to a compensation–promotion policy, $\{W_{1,f,t}, W_{2,f,t}, \tau_{f,t}\}$, with wages for juniors and seniors (i.e., promoted juniors) $W_{1,f,t}$ and $W_{2,f,t}$, and a promotion cutoff $\tau_{f,t}$, which is defined as the share of juniors dismissed. Junior workers for which $\theta_{i,t} \geq \theta^P_{f,t}(\tau_{f,t})$ are promoted, where $\theta^P_{f,t}(\tau_{f,t})$ denotes the $\tau_{f,t}$-quantile of effort signals among junior workers hired by firm f in period t.

We assume that compensation of employees is set by the firm, rather than negotiated.[12] We also abstract from a lateral labor market within the industry. In particular, firms cannot hire senior workers directly; instead, workers must first be hired as juniors and attain seniority through internal promotion in the following period. As discussed in Section 3.6.3 and Appendix D, this assumption is not essential for our results: the theoretical mechanisms and testable implications remain unchanged in an extended version of the model in which firms can hire senior workers laterally.

[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) and 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 3.6.1.

3.1.3 *Workers' Outside Options*

When making their employment decisions regarding firm f, workers of type 1 and 2 hired in period t draw outside options $\ddot{U}^{alt}_{1,i,f,t}$ and $\ddot{U}^{alt}_{2,i,f,t}$, respectively. These are the outside options considered by juniors when deciding which firm to join and by seniors when deciding whether to stay in their current firm. They represent the utility a worker would derive from working at the most attractive accounting firm other than f or outside of public accounting—whichever is higher.[13]

To simplify the notation in the remainder of this section, define $U^{alt}_{1,i,f,t} \equiv \ddot{U}^{alt}_{1,i,f,t} - \xi_{1,f} - \epsilon_{i,f,t}$ and $U^{alt}_{2,i,f,t} \equiv \ddot{U}^{alt}_{2,i,f,t} - \xi_{2,f} - \epsilon_{i,f,t+1}$. That is, for $s \in \{1,2\}$, $U^{alt}_{s,i,f,t}$ is the outside option net of the firm-specific and idiosyncratic non-pecuniary terms. To determine whether worker i works at firm f, we can compare $U^{alt}_{s,i,f,t}$ to the wage- and effort-related components of worker i's utility, as explained in Section 3.2.2. We assume that, within firm f, the draws of $U^{alt}_{s,i,f,t}$ are i.i.d. across i. We further assume that $U^{alt}_{s,i,f,t}$ is independently distributed over s and t.[14] That said, the distribution of $U^{alt}_{s,i,f,t}$ can vary with f and s, reflecting cross-firm heterogeneity in amenities and systematic differences between the outside options of junior and senior workers. Denote the distribution of $U^{alt}_{s,i,f,t}$ by $G^{alt}_{s,f}$, and let $g^{alt}_{s,f}$ be the associated density.

Juniors observe $U^{alt}_{1,i,f,t}$ before deciding whether to accept employment and how much effort to exert, but they do not observe $U^{alt}_{2,i,f,t}$ in advance. A worker hired by firm f in period t and subsequently promoted observes $U^{alt}_{2,i,f,t}$ at the beginning of period $t+1$, before deciding whether to remain with the firm.

3.1.4 *Product Market*

Firms are monopolistic competitors in the product market. The demand for accounting services of type s is loglinear with elasticity $\eta < -1$. That is, denoting by $P_{s,f,t}$ the price of service s by firm f and by $\overline{Q}_t$ and $\overline{P}_t$ the market-level quantities and prices, we have

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

[13] In the case of senior workers, given our assumption of no lateral hiring by accounting firms, the outside option consists solely of the highest utility that the worker could obtain outside of the public accounting labor market.

[14] It is sufficient for these conditions to hold that $\ddot{U}^{alt}_{s,i,f,t}$ is independent over s and i, $\epsilon_{i,f,t}$ is i.i.d. over i, and both $\ddot{U}^{alt}_{s,i,f,t}$ and $\epsilon_{i,f,t}$ independently distributed over t. Recall our previous assumption that $\xi_{1,f}$ and $\xi_{2,f}$ are time-invariant.

Then, denoting by $P_s(Q_{s,f,t})$ 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

$$R_s(\mathcal{I}_{s,f,t}) \equiv P_s\left(F(\mathcal{I}_{s,f,t})\right) F(\mathcal{I}_{s,f,t}) = d_t\left[F(\mathcal{I}_{s,f,t})\right]^{\frac{1+\eta}{\eta}} \quad s \in \{1,2\}, \tag{3}$$

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 3.4—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 $e^*_{2,f,t}$.[15] Let $\bar{U}_{2,f,t}$ denote the senior workers' value of working at firm f in period t, net of the non-pecuniary components $\xi_{2,f}$ and $\epsilon_{i,f,t}$. We thus have

$$\bar{U}_{2,f,t} = \alpha W_{2,f,t} - C(e^*_{2,f,t}). \tag{4}$$

Juniors choose effort to maximize the expected discounted utility of working at firm f. In setting up the juniors' problem, it is without loss to consider the utility net of the non-pecuniary terms $\xi_{1,f}$ and $\epsilon_{i,f,t}$, which we denote by $\bar{U}_{1,f,t}$. Therefore,

$$\bar{U}_{1,f,t} = \max_e \ \alpha W_{1,f,t} - C(e) + \rho\left\{\left[1 - \Phi(\theta^P_{f,t}(\tau_{f,t}) - e)\right]\bar{\bar{U}}_{2,f,t} + \Phi(\theta^P_{f,t}(\tau_{f,t}) - e)\bar{U}^{alt}_{2,f}\right\}, \tag{5}$$

where

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

and $\bar{U}^{alt}_{2,f}$ is the expected value of outside options for seniors at firm f, $\mathbb{E}[U^{alt}_{2,i,f,t}]$. The term $\bar{\bar{U}}_{2,f,t}$ is the expected continuation value of a worker who is promoted in period t, prior to the realization of the senior outside option and period-$t+1$ non-pecuniary terms. A worker exerting effort e is promoted with probability $1 - \Phi(\theta^P_{f,t}(\tau_{f,t}) - e)$ and dismissed

[15] 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 3.6.1—that senior workers are the residual claimants—there are incentives to set $e^*_{2,f,t} > \underline{e}$. In any case, the choice of $e^*_{2,f,t}$ is exogenous at the time the firm determines the compensation–promotion policy $\{W_{1,f,t}, W_{2,f,t}, \tau_{f,t}\}$.

with the complementary probability. Accordingly, the term in braces in (5) 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 $\bar{\tilde{U}}_{2,f,t}$ and $\bar{U}^{alt}_{2,f}$.

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

$$C'(\tilde{e}_{f,t}) = \rho\phi(\kappa^P(\tau_{f,t}))[\bar{\tilde{U}}_{2,f,t} - \bar{U}^{alt}_{f,2}], \tag{6}$$

where $\kappa^P(\tau_{f,t})$ denotes the $\tau_{f,t}$-quantile of the distribution Φ.

The following lemma follows directly from (6) and the convexity of $C(\cdot)$:

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

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 $\tau_{f,t} \in (0,1)$—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 firm f is:

$$L_{1,f,t} = G^{alt}_{f,1}(\bar{U}_{1,f,t})\,\mathcal{L}. \tag{7}$$

Meanwhile, the number of seniors in $t+1$ is the number of promoted juniors that choose not to leave the firm:

$$L_{2,f,t+1} = (1-\tau_{f,t})G^{alt}_{2,f}(\bar{U}_{2,f,t})\,L_{1,f,t}. \tag{8}$$

3.2.3 *Firm's Problem*

Given incumbent seniors $l_{2,f,t}$, firm f[16] chooses $\{W_{1,f,t}, W_{2,f,t}, \tau_{f,t}\}$ to maximize the discounted value:[17]

$$\begin{aligned} V(l_{2,f,t}) = \max_{W_{1,f,t}, W_{2,f,t}, \tau_{f,t}} \Big\{ & R_1(L_{1,f,t}, l_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t}) + R_2(l_{2,f,t}, L_{1,f,t}, e^*_{2,f,t}, \tilde{e}_{f,t}) \\ & - W_{1,f,t} L_{1,f,t} - \rho W_{2,f,t}(1-\tau_{f,t}) G^{alt}_{2,f}(U_{2,f,t}) L_{1,f,t} \\ & + \rho V\left[(1-\tau_{f,t}) G^{alt}_{2,f}(U_{2,f,t}) L_{1,f,t}\right] \Big\}, \end{aligned} \tag{9}$$

subject to the incentive-compatibility constraint (6), and with $L_{1,f,t}$ given by (7). Under constant-elasticity demand in the product market, the marginal revenue associated with service s in period t, in equilibrium, is

$$MR_{s,t} = \frac{1+\eta}{\eta} d \left[F_s(\mathcal{I}_{s,f,t})\right]^{\frac{1}{\eta}}, \quad s \in \{1,2\}.$$

Then, the envelope condition from (9) gives the marginal value of seniors:

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

3.3 Equilibrium

Firms are monopsonistic competitors in the labor market and monopolistic competitors in the product market, taking market-level wages, employment, prices, and output as given. Formally, we define an equilibrium of our model as a starting senior employee quantity $L_{2,f,0}$; and sequences of compensation–promotion policies, $\{W_{1,f,t}, W_{2,f,t}, \tau_{f,t}\}_{t\in\mathbb{N}^+}$; junior and senior employee quantities $\{L_{1,f,t}, L_{2,f,t}\}_{t\in\mathbb{N}^+}$; and junior effort levels $\{\tilde{e}_{f,t}\}_{t\in\mathbb{N}^+}$ for each firm $f \in \{1,\ldots,F\}$, such that:

- Given $L_{2,f,t-1}$, the policy $\{W_{1,f,t}, W_{2,f,t}, \tau_{f,t}\}$ solves the firm's problem (9),

[16] In our application, 'the firm chooses' will mean 'the equity partners choose'. Given that monitoring the signals θ_i is presumably costly, the firm needs residual claimants of profits to set the promotions policy (and wages), as shown by Alchian and Demsetz (1972).

[17] Note that the compensation-promotion policy set in period t fully determines the number of type-1 workers employed by the firm in t, $L_{1,f,t}$; and the number of type-2 workers employed in $t+1$, $L_{2,f,t+1}$. 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.

- Given $\{W_{1,f,t}, W_{2,f,t}, \tau_{f,t}\}$, the effort level $\tilde{e}_{f,t}$ solves the junior worker's incentive-compatibility constraint (6),
- Given $\{W_{1,f,t}, W_{2,f,t}, \tau_{f,t}\}$, the labor supply quantities $\{L_{1,f,t}, L_{2,f,t}\}$ satisfy the transition rules (7) and (8),

for each f and t. Moreover, we require

$$\sum_{f=1}^{F} L_{1,f,t} \leq \mathcal{L}, \quad \sum_{f=1}^{F} L_{2,f,t} \leq \mathcal{L},$$

for every period t; that is, the aggregate number of workers employed at accounting firms at any level cannot exceed the total number of workers in the respective cohort.[18] Finally, for any firm f and period t, the distribution of workers' outside options $\ddot{U}^{alt}_{1,i,f,t}$ and $\ddot{U}^{alt}_{2,i,f,t}$ must be consistent with the compensation–promotion policies set by competing firms.

3.4 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.

3.4.1 Symmetric Steady-State Equilibrium

We make two assumptions to simplify the exposition of our results. First, the equilibrium is in steady-state, so all variables are stationary. Second, we assume that the distributions of outside options for junior and senior workers, $G^{alt}_{1,f}$ and $G^{alt}_{2,f}$, are the same across all f, denoted by G^{alt}_{1} and G^{alt}_{2}, respectively. Since outside option distributions are the only model primitives that vary at the firm level, the latter assumption implies that all firms set the same compensation–promotion policies in equilibrium. Accordingly, junior workers exert the same level of effort regardless of the firm. We refer to this equilibrium as *symmetric steady state*. Importantly, neither the steady-state nor the symmetry assumptions are crucial for the derivation of the testable implications established here, and we do not impose either of them in our empirical implementation in Section 4.

In a symmetric steady-state equilibrium, for any t and f, the compensation–promotion policies are

$$W_{1,f,t} = W_1, \quad W_{2,f,t} = W_2, \quad \tau_{f,t} = \tau,$$

[18] This constraint is automatically met for senior workers if it holds for juniors, given the employment transition equation.

the number of workers employed by the firm and the equilibrium effort levels are

$$L_{1,f,t} = L_1, \quad L_{2,f,t} = L_2, \quad \tilde{e}_{f,t} = \tilde{e}, \quad e^*_{2,f,t} = e^*_2,$$

and the net workers' values become

$$\bar{U}_{1,f,t} = \bar{U}_1, \quad \bar{U}_{2,f,t} = \bar{U}_2, \quad U^{alt}_{2,f} = U^{alt}_2, \quad \bar{\bar{U}}_{2,f,t} = \bar{\bar{U}}_2.$$

The workers' conditions collapse to

$$\bar{U}_2 = \alpha W_2 - C(e^*_2), \tag{11}$$

$$\bar{U}_1 = \alpha W_1 - C(\tilde{e}) + \rho\left[(1-\tau)\bar{\bar{U}}_2 + \tau \bar{U}^{alt}_2\right], \tag{12}$$

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

and employment satisfies

$$L_1 = G^{alt}_1(\bar{U}_1)\mathcal{L}, \qquad L_2 = (1-\tau)G^{alt}_2(U_2)L_1. \tag{14}$$

Since all firms are identical in a symmetric steady-state equilibrium, we refer to any firm as *the firm* throughout this Section, unless otherwise noted.

3.4.2 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

$$\text{MRPL}_1 \equiv \frac{1+\eta}{\eta} d \left\{ \left[F_1(L_1, L_2, \tilde{e}, e^*_2)\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_1} + \left[F_2(L_1, L_2, e^*_2, \tilde{e})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_1} \right\}, \tag{15}$$

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

Next, define the firm's expected net present values (NPVs) of wages and marginal products for a new hire:

$$\text{NPV}_W = W_1 + \rho(1-\tau)G^{alt}_2(U_2)\, W_2, \tag{17}$$

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

3.4.3 *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 problem 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 has monopsony power. The reason is that, absent incentive considerations, the sole role of wages is to attract and retain workers, so the monopsony effect is fully absorbed by the junior wage.[19] 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.[20]

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,

[19] Even though only junior wages are marked down, monopsony still distorts employment at both the junior *and* senior levels.

[20] 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.

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.

3.4.4 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 *short-run* residual elasticity of junior labor supply is

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

Similarly, define the *long-run* 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

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

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 problem. 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.

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 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 $\mathrm{NPV}_{\mathrm{MRPL}}$. 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.

3.5 Effects of Eliminating Wage Markdown Variation

We now characterize the effects of eliminating the junior wage markdown, $W_1 < MRPL_1$, on output and prices across different environments. The proofs of the propositions and the corollary below are in Appendix A.2.2.

Proposition 3 (Irrelevant Effort and Monopsony)**.** *Assume that: (i) junior effort is irrelevant for production, so $\partial F_s/\partial \bar{e}_1 = 0$ for all $s \in \{1, 2\}$; (ii) the firm has monopsony power, so $\psi_{LR} < \infty$ and $NPV_W < NPV_{MRPL}$; (iii) $\tau \in (0, 1)$.*

Then, a policy that imposes $W_1 = MRPL_1$ increases employment at both levels, raises output, and lowers prices.

Proposition 4 (Productive Effort and Perfect Competition)**.** *Assume that: (i) junior effort is productive, so $\partial F_s/\partial \bar{e}_1 > 0$ for at least one $s \in \{1, 2\}$; (ii) labor markets are perfectly competitive at both junior and senior margins, so $NPV_W = NPV_{MRPL}$; (iii) $\tau \in (0, 1)$;*

Then, imposing $W_1 = MRPL_1$ lowers junior effort, decreases employment and output, and raises output prices.

The two propositions highlight that the output and price effects of eliminating the junior wage markdown depend on its source. Under monopsony with irrelevant effort, the markdown is purely distortive, so removing it increases employment and output. Under perfect competition with productive effort, the markdown is part of an optimal backloaded incentive contract, so removing it weakens effort incentives and reduces employment and output.

It follows that, in the presence of both productive effort and monopsony power, the effect of eliminating junior markdowns on employment, output, and prices is ambiguous, as it depends on the net effect of the elimination of the monopsony distortion vs. the effort-incentive mechanism.

Corollary 1 (Productive Effort and Monopsony)**.** *Assume that: (i) junior effort is productive, so* $\partial F_s/\partial \bar{e}_1 > 0$ *for at least one* $s \in \{1,2\}$*; (ii) the firm has monopsony power, so* $\psi_{LR} < \infty$ *and* $NPV_W < NPV_{MRPL}$*; (iii)* $\tau \in (0,1)$*.*

Then the effects of imposing $W_1 = MRPL_1$ *are, in general, ambiguous: the effects on employment, output, and prices can be either positive or negative.*

Real-life policies are unlikely to impose firms to pay workers their MRPL. However, the results above help to understand policies that compress the wage distribution within firms, such as pay-ratio caps, which have been proposed in various countries Caselli (2025), mandatory executive pay disclosure, and collective wage bargaining. Experimental evidence on the incentive effects of disclosing executive pay in Cullen and Perez-Truglia (2022) is, for instance, consistent with Proposition 3.

3.6 Model Discussion and Extensions

3.6.1 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

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

where b^{in}_{t+1} is the partnership buy-in, π^{s}_{t+1} is the expected share of one-period profits, W^{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 W^{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.[21]

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.

3.6.2 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, second-period compensation exceeds the corresponding marginal revenue product in present value. Assuming quadratic effort costs, the latter result extends to all periods: from the second seniority level onwards, discounted wages always strictly exceed discounted marginal revenue product. Consequently, wage schedules are inherently backloaded: higher-level workers are paid above marginal product to sustain effort incentives over the career.

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.

3.6.3 Lateral Hiring and Retaining Workers who Failed Promotion

For expositional simplicity, the baseline model assumes that the firm employs only senior workers who were initially hired as juniors and subsequently promoted internally. In practice, firms can also hire senior workers laterally—although such hiring is infrequent in our empirical application, as was mentioned earlier. As shown in Appendix D, our main theoretical results extend directly to an environment in which the firm can hire senior workers from an external market.

More broadly, lateral hiring can be reinterpreted as the retention of juniors who fail to obtain promotion, but under a compensation scheme that differs from that of internally promoted seniors. Under this interpretation, the extension in Appendix D effectively enlarges the firm's contract space by allowing distinct compensation structures for different

[21] One might wonder how equation (2) 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, with discounting it becomes possible that $W_1 + \rho W_2 < MRPL_1 + \rho MRPL_2$.

categories of senior workers. Importantly, the core mechanisms generating the incentive and monopsony distortions remain unchanged, so that the model's testable implications are robust to this richer contracting environment.

3.6.4 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 arise from two distinct forces: backloaded efficiency wages and monopsony power. Crucially, each carries different efficiency consequences. To examine the existence and quantitative importance of both 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

We implement empirical specifications for the model presented in Section 3. The most important change is that we impose functional form assumptions on the production function, in order to estimate it. Moreover, firms are subject to demand and productivity shocks. For expositional reasons, we suppress time notation throughout the empirical model.

4.1.1 Market Structure, Product Demand, Labor Utility

A set of accounting firms $\mathcal{F}$ competes on both the product and labor market. A firm $f \in \mathcal{F}$ hires 'billable' employees with seniority types $s \in \{1, ..., |S|\}$. We follow the type categorization provided in the dataset: professional staff with 0–2, 3–5, 6–8, and 9+ years of experience, paraprofessionals, and partners. Each type 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 $W_{s,f}$.[22] Consequently, each type generates $R_{s,f} = P_{s,f}Q_{s,f}$ in revenues.

[22] We do not impose that wage contracts are written at the hourly level. $W_{s,f}$ can be the result of a monthly or annual compensation contract as well.

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 prices each service line separately since the IPA data shows that two thirds of firms follow this pricing practice.[23]

We implement the loglinear demand curve for type-s accountants from (2) at the firm level with a common demand elasticity η, 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, $MC_{s,f}$:

$$\mu_{s,f} \equiv \frac{P_{s,f}}{MC_{s,f}} = \frac{\eta}{1+\eta}.$$

We maintain the linear labor flow utility specification from Equation (1), but now allow for a random wage coefficient α_f, to allow for flexible variation in labor supply elasticities across firms.[24]

$$U_{i,f} = \alpha_f W_{s(i),f} - C(e_{i,f}) + \xi_{s(i),f} + \epsilon_{i,f} \tag{25}$$

4.1.2 *Production of Accounting Services*

The production of services is given by Equation (26). Billable hours produced by each employee type s depend on hours worked by all employee types, via a Cobb-Douglas production function

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.

$$Q_{s,f} = \left(\prod_{k \in S} L_{k,f}^{\beta^l_{s,k}} \right) \boldsymbol{O}_f^{\boldsymbol{\beta}^o_s} \Omega_{s,f}(\bar{e}_{s,f}) \exp(c_s + v_{s,f}). \tag{26}$$

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

Type-specific productivity levels are denoted by $\Omega_{s,f}$. In our main empirical specification, we let productivity of type-s workers be only a function of average effort of the

[23] A quarter of firms instead charges a fixed fee for the bundle of accountants. In Appendix B.5, we extend the model to allow for bundle pricing, and discuss the implications for our estimates.

[24] This parallels Volpe (2024), though we include firm-specific rather than worker-specific random coefficients.

same type of workers, not of other worker types.[25] Finally, measurement error in output is denoted by $\exp(v_{s,f})$, and we include a seniority-level intercept c_s.

Unfortunately, we do not observe fixed assets, such as office space rent or insurance expenses. In a robustness analysis in Appendix B.4.1, 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.

Discussion. Our model features multiple outputs and team production. These features are valuable for at least two reasons. First, our model allows for output produced by accountants of different seniority levels to be differentiated, whereas estimating a standard firm-level production model would require aggregating billing hours across worker types.[26] Second, our model allows for labor inputs to be non-rival: we do not require assigning the different services produced to inputs, as in Dhyne et al. (2017). 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.

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).[27] Although we could in principle allow for this, a specification including the entire output and input vectors would be hard to estimate in practice given the number of products and our dataset size. However, we do allow type-s productivity to be correlated arbitrarily across worker types.

Second, although we allow output to be differentiated across seniority types within firms, we do not allow for differentiation across firms. For instance, Q billed hours of equity partners are assumed to be homogeneous across firms. As we show in Appendix B.4.3, this assumption can be relaxed by including a price control on the right-hand side of the production functions, yielding very similar estimates.

4.1.3 Wage Markdowns

The marginal revenue product of type-s labor, $\text{MRPL}_{s,f} \equiv \partial R_f / \partial L_{s,f}$, can be expressed as:

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

[25] We extend this in Appendix B.2.

[26] We compare our estimates with a more conventional firm-level revenue production function that aggregates output across employees in Appendix B.4.4.

[27] Doing so would require including all of the other outputs on the right-hand side, $Q_{s,f} = F(L_{s,f}, \bar{e}_{s,f}, L_{-s,f}, \bar{\boldsymbol{e}}_{-s,f}, Q_{-s,f})$.

Each type of accountant contributes to revenue both directly, through its own billing, and indirectly, through its effect on the billing revenue of other seniority types. The demand elasticity η enters each marginal revenue product because changing an accountant's output affects their billing rates.

Using the MRPL expression from Equation (27), we write the 'instantaneous markdown' $\delta_{s,f}$ at every seniority level as:

$$\delta_{s,f} = \frac{W_{s,f} L_{s,f}}{\sum_k \beta^l_{s,k} R_{k,f} \left(\frac{\eta+1}{\eta}\right)}. \tag{28}$$

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 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 require data on billable hours $Q_{s,f}$ rather than billing revenue $R_{s,f}$ alone. Denoting lowercase variables as logs, we estimate $|S|$ production functions, one for each type of worker:

$$q_{s,f} = \sum_{k \in S} \left(\beta^l_{s,k} l_{k,f}\right) + \beta^o_s \boldsymbol{o}_{s,f} + \omega_{s,f} + c_s + \upsilon_{s,f}. \tag{29}$$

Denote the one-year lag of any variable X as $\hat{X}$. We impose an AR(1) assumption on the sum of productivity $\omega_{s,f} \equiv \ln(\Omega_{s,f})$ and measurement error $\upsilon_{s,f}$, with worker type-specific serial correlation σ_s and a constant c_s.

$$\omega_{s,f} + \upsilon_{s,f} = \sigma_s(\hat{\omega}_{s,f} + \hat{\upsilon}_{s,f}) + \nu_{s,f}. \tag{30}$$

We express the productivity shock as a function of data and the parameters $(\sigma_s, \boldsymbol{\beta}_s)$ to be estimated.

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

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, $\boldsymbol{X}^{\text{pre}}$.

Flexible and predetermined inputs are respectively chosen after and prior to the arrival of the productivity shock, $\nu_{s,f}$.

We assume that professionals at all seniority levels s, except entry-level staff, are dynamic inputs (facing adjustment costs), meaning that firms choose them before observing the transient productivity shocks $\nu_{s,f}$. This is consistent with the model in Section 3, in which junior employment and transition rates are chosen for each cohort, implying that senior employees are 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 $(\sigma_s, c_s, \boldsymbol{\beta}_s)$:

$$\begin{aligned} E(\nu_{s,f}|\boldsymbol{X}_f^{\text{pre}}, \hat{\boldsymbol{X}}_f^{\text{pre}}) &= 0, \\ E(\nu_{s,f}|\hat{\boldsymbol{X}}_f^{\text{flex}}) &= 0. \end{aligned} \tag{31}$$

4.2.2 Estimation

We measure outputs $Q_{s,f}$ as total billable hours, computed as billable hours per worker multiplied by the number of employees of type s. Similarly, labor inputs $L_{s,f}$ are measured as hours worked per employee multiplied by headcount. 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 of equity partners. Thus, variation in the contribution of these two worker types stems only from their headcounts. Relatedly, 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 firm bins g, where each firm belongs to exactly one bin. In Appendix B.1, we show that we can still recover the production function parameters by estimating Equation (29) at the bin-year level under two additional assumptions. First, we require that input prices and residual input supply elasticities are identical across firms within each bin-year. 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 $\nu_{s,f}$.

We estimate the production function by specifying the moment conditions (31) 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 in the instruments vector. We estimate the production function using one-step GMM and compute heteroskedasticity-robust standard errors.

4.2.3 Results: Production Estimates

Table 2 reports the production estimates. The diagonal elements report the own-output elasticities for each role. The own-output elasticity is roughly around one for the professional staff categories, but substantially below one for paraprofessionals (at 0.747), indicating. In contrast, the own-output elasticity is significantly above one (at 1.231) for equity partners, indicating increasing returns to equity partners. This is consistent with the intuition that equity partners are the main responsable for attracting new clients, which implies that their worked hours disproportionally generate additional billing hours compared to other labor types, whereas there seem to be decreasing returns to the usage of paraprofessionals.

The off-diagonal elements measure spillovers in production. Three patterns stand out. First, the upper-triangular elements measure externalities of junior employees on more senior colleagues. For employees with 0-2 years of experience (the first row), these are mostly negative: these workers still require training, which is costly to their senior colleagues.

Second, the lower-triangular elements are the effects of senior hours on junior output. Again looking at employees with 0-2 years of experience (the first column), these coefficients are mostly positive, which reflects that juniors gain from mentoring by their senior colleagues.

Third, the row with the highest positive coefficients are the equity partners (the sixth row), these have strong positive, and mostly significant, spillovers on all other employees. This highlights the crucial role of equity partners in providing general management and mentorship, and in business development. These positive spillovers from equity partners are crucial in measuring their marginal revenue products, which highlights the importance of modeling multi-output team production.

Turning to the non-partner-track staff, paraprofessionals seem to generate negative spillovers for the other employee types, which could be due to task duplication or congestion effects. Interns, materials, administrative staff, and capital all have small coefficients, which is in line with their small revenue shares.

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		174		174	

Notes: This table shows the estimated production functions for the seven billing revenue-generating employee types in the dataset (the columns). The own- and cross-output elasticities for all inputs are reported in the rows. The number of observations shrinks to 174 because interns are only observed on 2017-2024 and because we include first lags, effectively estimating the production parameters on the period 2018-2024.

4.2.4 Results: TFP vs. MRPL

We recover instantaneous wage markdowns from Equation (28), using the estimated output elasticities together with observed billing rates and compensation. Identification of the markdown requires the product-market residual demand elasticity, η. Under a CES demand system, 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 accountant types.[28] Given the instantaneous markdown estimates and observed wages, we can then recover the implied MRPL directly. We trim the estimated instantaneous markdowns at the 1st and 99th percentiles to remove outliers.

Figure 4 plots the distributions of log productivity ω and of MRPL across employee types. Whereas partners rank lowest in terms of productivity, they rank near the top across employee types in terms of MRPL.[29] This contrast is again a feature of multi-output production: using traditional TFP measures, partners look unproductive because they work relatively long hours for little billable hour production. However, much of these

[28] Footnote on bundle vs individual demand elasticities here, again refer to Appendix B.5.
[29] This holds for both TFPQ and TFPR.

worked hours generate billing revenue for more junior accountants, which shows up in the MRPL measure but not in the productivity residual.

When comparing MRPL distributions, we find evidence of an 'MRPL ladder'. On average, paraprofessionals have by far the lowest MRPL. Among the 'partner-track' staff, starters have the lowest MRPL, which then gradually increases through 9+ years of service. This pattern is consistent with human capital accumulation, which generates dynamic labor supply independently of backloaded incentive pay (Jungerman, 2023).

Interestingly, equity partners actually have a slightly lower MRPL than non-equity partners and 9+year employees, and exhibit a substantial left tail. This pattern is consistent with our model, given that equity partners face no tournament incentives.

4.3 Testing for Incentive Contracts and Monopsony Power

4.3.1 Backloaded Compensation Contracts

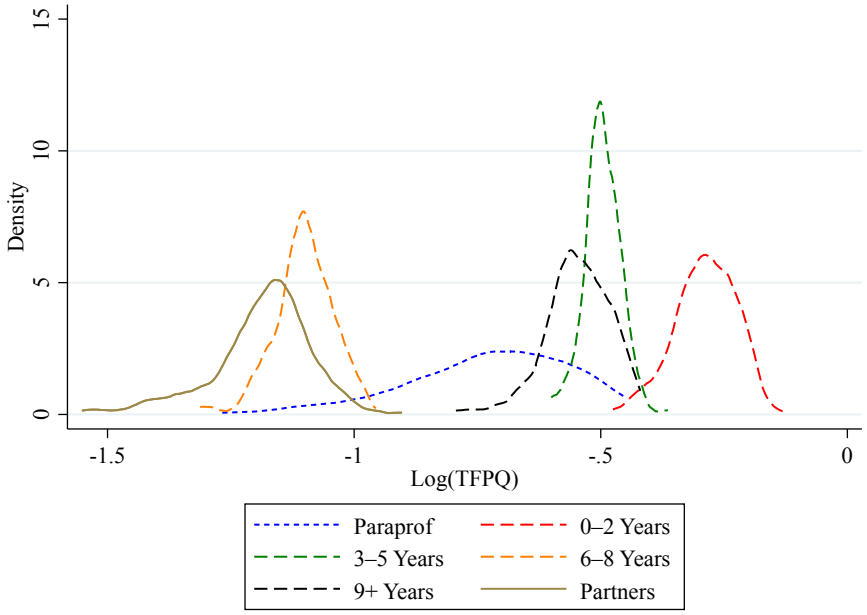
The estimated average instantaneous wage markdowns/markups are summarized in Figure 5. They are significantly different from one at conventional levels for all worker categories. Paraprofessionals, who are outside the 'partner-track', earn on average 68% of their marginal revenue product. Turning to the 'career' professional staff, employees with 0-2 years of experience are paid only 55% of their MRPL, a markdown that remains relatively constant throughout the first 8 years of seniority. That junior professional staff face a 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 MRPL. 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.

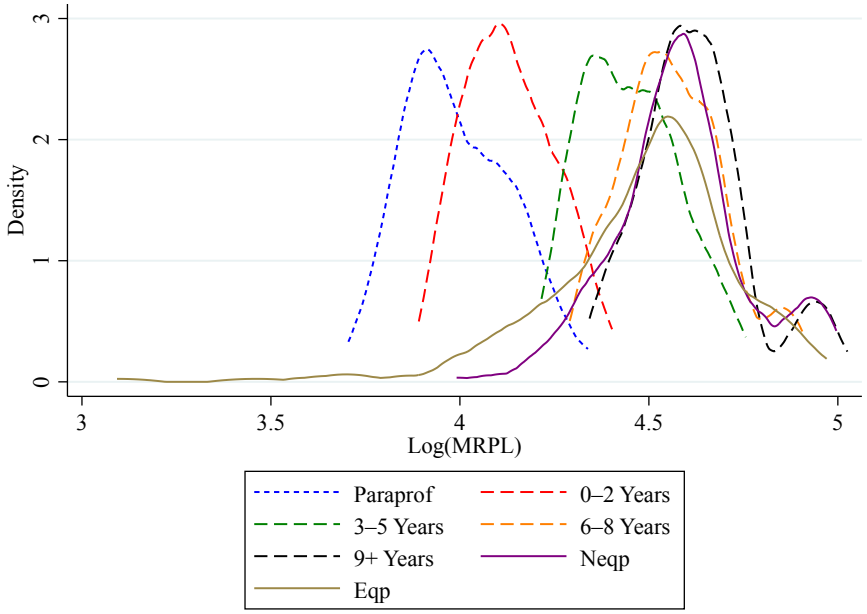
Given our Proposition 1, we reject the null hypothesis of irrelevant effort: wages are significantly above the MRPL for both non-equity and equity partners. Therefore, our evidence indicates that public accounting firms are implementing backloaded efficiency wage contracts, and junior employee wage markdowns are (at least partly) due to wage backloading.

Figure 4: Productivity vs. MRPL Dispersion

(a) log(TFP)

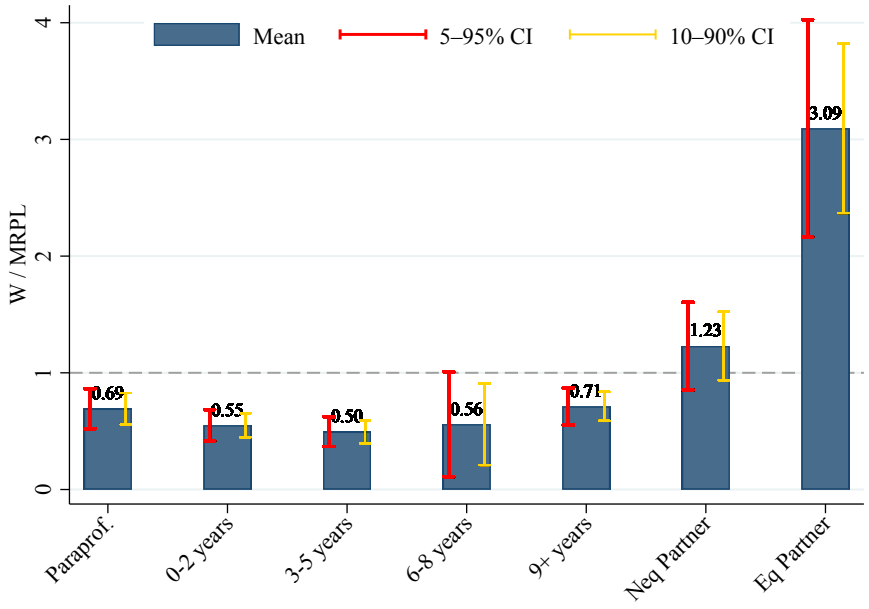


(b) log(MRPL)



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 5: Instantaneous Wage Markdowns and Seniority



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

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

$$\text{NPV}_{W_f} = \sum_{t=1}^{T} \rho^{t-1} \lambda_{ft} W_{ft}, \tag{32}$$

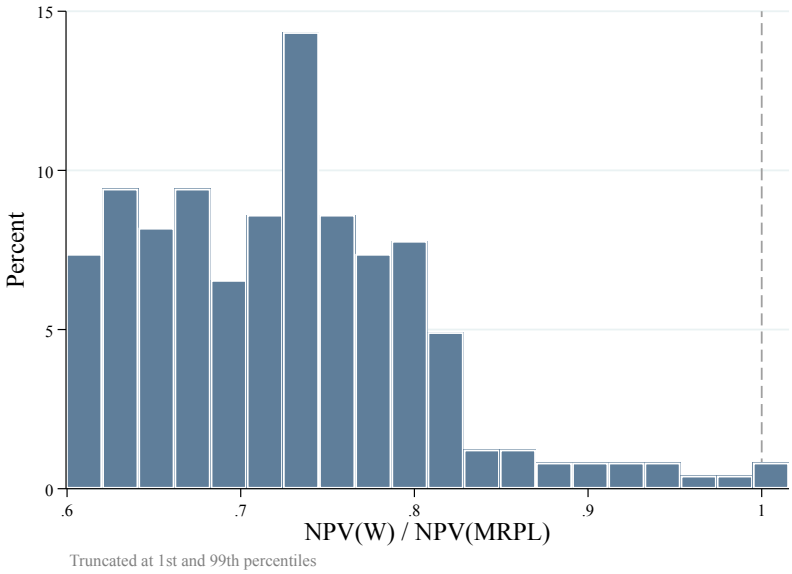
and similarly for the NPV of the marginal revenue product of labor, NPV_{MRPL}.

We estimate the survival probabilities λ_t using worker stocks and assuming a steady-state labor market, as explained in Appendix B.3. We impose an annual discount factor $\rho = 0.95$, which corresponds to a real interest rate of around 5%.[30]

Figure 6 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 0.732, with a 5-95% confidence interval [0.698,0.767], so, on average, the NPV of compensation is significantly below that of the MRPL. Therefore, in accordance with Proposition 2, our

[30] In Appendix B.4.2, we examine robustness of the results to different discount factors.

Figure 6: Distribution of Expected Wage Markdowns



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.

4.3.3 Observed vs. Model-Implied Seniority Levels

When interpreting the results through the lens of the model with $S > 2$ seniority levels, the complication arises that the seniority levels measured in the dataset are not necessarily the relevant seniority levels in the model; that is, the levels at which firms set wages. In Appendix B.6, we show that we can use our N-level model results to determine how the observed seniority levels map into the model-implied seniority levels.

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 across firm bins 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 entry-level 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).

In 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 elasticities reported in meta-analyses in Card (2022) and Azar and Marinescu (2024) , and similar to 'steady-state' labor supply elasticities 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	2.726	0.572
1st Quartile	0.184	0.105	1.875	0.390
Median	0.222	0.126	2.602	0.527
3rd Quartile	0.259	0.164	3.359	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 firm bins. 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 in the eyes of workers. 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 endogenous variables. Firms with higher training expenditure (relative to revenue) face more elastic labor supply curves, which could indicate that employers 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 be interpreted as evidence that CPA-certified accountants tend to have more attractive outside options than their non-CPA co-workers.

Table 4: Covariates of Long-Run Supply Elasticities

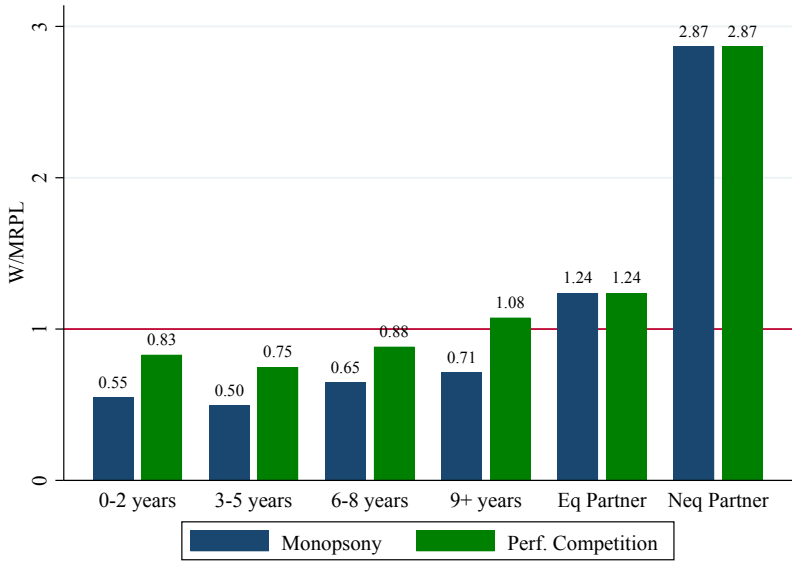
	(I)		(II)	
	log(Labor Supply Elast.)			
	Est.	S.E.	Est.	S.E.
Medium-size firm	0.145	0.072	0.085	0.067
Small-size firm	-0.395	0.055	-0.519	0.060
Year = 2016	0.029	0.062	0.028	0.054
Year = 2017	0.007	0.057	-0.020	0.052
Year = 2018	0.110	0.064	0.083	0.059
Year = 2019	0.072	0.069	0.044	0.073
Year = 2020	0.062	0.066	-0.059	0.074
Year = 2021	0.115	0.164	0.036	0.134
Year = 2022	0.339	0.114	0.232	0.126
Year = 2023	0.306	0.076	0.067	0.116
Year = 2024	0.482	0.089	0.239	0.129
log(Marketing exp.)			-0.410	0.115
log(Training exp.)			0.343	0.089
log(Recruiting exp.)			0.081	0.055
log(Technology exp.)			0.162	0.130
CPA share			-0.549	0.479
R-squared	.358		.419	
Observations	238		238	

Notes: Size categories (small, medium, large) are provided by the data provide and subdivide each location into three approximately equal groups of firms, ranked by annual revenue. Omitted categories are the 'Large firms' category in the year 2015.

4.5 Decomposition Exercise

How much of the junior markdown is driven by monopsony power, as opposed to wage backloading? Under incentive contracts, junior wages need not equal marginal revenue

Figure 7: Markdown Decomposition



products even in the absence of monopsony power. To disentangle the two, we quantify the factor χ_f by which non-partner wages would need to increase so that $\text{NPV}_w = \text{NPV}_{MRPL}$. We hold partner wages (the tournament prize) and promotion probabilities fixed, as both of these together determine junior incentives, and we want to leave the incentive structure intact. Letting T^{sen} denote the time of promotion to partnership, χ_f satisfies:

$$\sum_{t=1}^{T^{sen}} \chi_f \rho^{t-1} \lambda_t W_{ft} + \sum_{t=T^{sen}+1}^{T} \rho^{t-1} \lambda_t W_{ft} = \text{NPV}_{MRPL}$$

We compute χ_f at the firm-bin level by rearranging this equation, using observed wages, estimated marginal revenue products, and transition probabilities. On average, junior wages would need to rise by 52% to close the gap. Isolating the backloading component substantially compresses the markdowns: rather than ranging from 0.55 to 0.71, they would range from 0.83 (rookies) to 1.07 (associates with 9+ years of tenure), as shown in Figure 7.

This allows us to decompose junior markdowns into their monopsony and incentive contract components. For instance, 62% of the markdown for workers with 0–2 years of experience is attributable to monopsony power, with the remainder due to wage backloading.

One important caveat is that by holding senior wages and promotion probabilities fixed, we do not allow senior workers to grasp the rents that result from exercising monopsony power. We therefore view this decomposition as a back-of-the-envelope calculation rather

than a full counterfactual exercise, which would require resolving for optimal junior-senior employment ratios and senior wages under perfect labor market competition.

5 Conclusion

In this paper, we examine the exercise of monopsony power when firms use implicit long-term efficiency wage contracts to incentivize worker effort. We find that wage markdowns can originate both from monopsony power and from the use of backloaded incentive wage contracts—two mechanisms with 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 NPV of wages below the NPV of the marginal revenue product of labor. 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 in the short run, and more elastic—with elasticity around 2.7—in 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 lens of a static model: they are not just the consequence of monopsony power, but also of backloaded incentive pay.

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

Bonhomme, S. (2021). Teams: Heterogeneity, sorting, and complementarity. *arXiv preprint arXiv:2102.01802*.

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.

Caselli, F. (2025). Macroeconomic implications of executive pay caps.

Cullen, Z. and R. Perez-Truglia (2022). How much does your boss make? the effects of salary comparisons. *Journal of Political Economy 130*(3), 766–822.

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.

Herkenhoff, K., J. Lise, G. Menzio, and G. M. Phillips (2024). Production and learning in teams. *Econometrica 92*(2), 467–504.

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.

Jarosch, G., E. Oberfield, and E. Rossi-Hansberg (2021). Learning from coworkers. *Econometrica 89*(2), 647–676.

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*.

Numan, W. and M. Willekens (2012). An empirical test of spatial competition in the audit market. *Journal of Accounting and Economics 53*(1-2), 450–465.

Olley, S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–1297.

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 Proofs and Derivations . . . . . **OA - 3**
A.1 Firm's Problem: First-Order Conditions OA - 3
A.2 Proofs of Propositions in the Main Text OA - 4
A.3 Proof of Corrolary 1 OA - 7

B Empirical Analysis Appendix . . . . . **OA - 7**
B.1 Aggregation to the Bin-Level OA - 7
B.2 Productivity spillovers OA - 8
B.3 Promotion Probabilities OA - 10
B.4 Alternative Specifications and Robustness OA - 11
B.5 Bundle Pricing OA - 18
B.6 Observed vs. Model-implied Seniority Levels OA - 19

C Tournament Model with Monopsonistic Competition and N Employee Levels . . . . . **OA - 22**
C.1 Environment and Timing OA - 22
C.2 Production and Prices OA - 22
C.3 Compensation–Promotion Policy OA - 23
C.4 Outside Options OA - 23
C.5 Workers' Values and Incentives OA - 23
C.6 Employment Dynamics OA - 24
C.7 Firm's Problem OA - 24
C.8 Steady State OA - 25
C.9 Incentive Contract Implications OA - 26
C.10 Entry Elasticity and the Exact Intertemporal Wedge OA - 30

D Tournament Model with Lateral Hiring . . . . . **OA - 32**
D.1 Primitives OA - 32

D.2 Worker and Firm Behavior OA - 32
D.3 Testable Implications OA - 33

Online Appendix

A Proofs and Derivations

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

We start from the recursive value function (9). 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

$$B_1 = G_1^{\text{alt}}(\bar{U}_1)\mathcal{L}, \qquad B_2 = G_2^{\text{alt}}(U_2).$$

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

$$W_1: \quad \frac{\partial \bar{U}_1}{\partial W_1} g_1^{alt}(\bar{U}_1)\mathcal{L}\left[MRPL_1 - W_1 - \rho W_2(1-\tau)B_2 + \rho(1-\tau)B_2 V'\Big((1-\tau)B_2B_1\Big)\right] - B_1 = 0,$$

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.

$$\begin{aligned} W_2: \quad & \frac{\partial \bar{U}_1}{\partial W_2} g_1^{alt}(\bar{U}_1)\mathcal{L}\Big[MRPL_1 - W_1 - \rho W_2(1-\tau)B_2 + \rho(1-\tau)B_2\, V'\big((1-\tau)B_2B_1\big)\Big] \\ & - \rho(1-\tau)B_1B_2 + \rho(1-\tau)B_1 \frac{\partial U_2}{\partial W_2} g_2^{alt}(U_2)\Big[- W_2 + V'\big((1-\tau)B_2B_1\big)\Big] \\ & + \frac{1+\eta}{\eta}\frac{\partial \tilde{e}}{\partial W_2}\left\{ d\left[F_1(B_1, l_2, \tilde{e}, e_2^*)\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial \tilde{e}} + d\left[F_2(l_2, B_1, e_2^*, \tilde{e})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial \tilde{e}} \right\} = 0, \end{aligned}$$

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.

$$\begin{aligned} \tau: \quad & \frac{\partial \bar{U}_1}{\partial \tau} g_1^{alt}(\bar{U}_1)\mathcal{L}\Big[MRPL_1 - W_1 - \rho W_2(1-\tau)B_2 + \rho(1-\tau)B_2\, V'\big((1-\tau)B_2B_1\big)\Big] \\ & + \rho W_2 B_2 B_1 - \rho B_2 B_1 V'\big((1-\tau)B_2B_1\big) \\ & + \frac{1+\eta}{\eta}\frac{\partial \tilde{e}}{\partial \tau}\left\{ d\left[F_1(B_1, l_2, \tilde{e}, e_2^*)\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial \tilde{e}} + d\left[F_2(l_2, B_1, e_2^*, \tilde{e})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial \tilde{e}} \right\} = 0. \end{aligned}$$

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 (5), we have

$$\frac{\partial \bar{U}_1}{\partial W_1} = \alpha.$$

Meanwhile, from (4), (5), and the definition of $\bar{\bar{U}}_2$, we have that

$$\frac{\partial \bar{U}_1}{\partial W_2} = \rho(1-\tau)B_2\alpha.$$

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

$$g_1^{alt}(\bar{U}_1)\mathcal{L}\Big[MRPL_1 - W_1 - \rho W_2(1-\tau)B_2 + \rho(1-\tau)B_2 V'\big((1-\tau)B_2B_1\big)\Big] = \frac{B_1}{\alpha}.$$

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 $\partial \bar{U}_1 / \partial W_1 = \alpha$, we obtain

$$\alpha g_1^{alt}(\bar{U}_1) = \frac{\psi_1}{W_1} G_1^{alt}(\bar{U}_1). \tag{OA.2}$$

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.2.3 *Proof of Proposition 3*

Proof. By Proposition 1, under irrelevant effort we have $W_2 = MRPL_2$. By Proposition 2, monopsony implies $NPV_W < NPV_{MRPL}$, that is,

$$W_1 + \rho(1-\tau)G_2^{alt}(U_2)W_2 < MRPL_1 + \rho(1-\tau)G_2^{alt}(U_2)MRPL_2.$$

Using $W_2 = MRPL_2$, it follows that $W_1 < MRPL_1$.

Consider now a policy imposing $W_1 = MRPL_1$. Given that the model is monopsonistic, equilibrium lies on the labor supply curve. Since effort is irrelevant for production, the policy does not affect output through incentives. The increase in W_1 raises junior utility $\bar{U}_1$, and therefore increases junior employment:

$$L_1 = G_1^{alt}(\bar{U}_1)\mathcal{L} \quad \Rightarrow \quad L_1 \uparrow .$$

Since U_2 is unchanged, $G_2^{alt}(U_2)$ is unchanged, and thus

$$L_2 = (1-\tau)G_2^{alt}(U_2)L_1 \quad \Rightarrow \quad L_2 \uparrow .$$

Because $\partial F_s / \partial L_k \geq 0$, output increases:

$$Q_s \uparrow, \quad s \in \{1,2\}.$$

Finally, since demand is downward sloping ($\eta < -1$), higher output implies lower prices:

$$P_s \downarrow, \quad s \in \{1, 2\}.$$

□

A.2.4 Proof of Proposition 4

Proof. By Proposition 1, under productive effort we have $W_2 > MRPL_2$. Under perfect competition, $NPV_W = NPV_{MRPL}$, so

$$W_1 + \rho(1-\tau)G_2^{alt}(U_2)W_2 = MRPL_1 + \rho(1-\tau)G_2^{alt}(U_2)MRPL_2.$$

Since $W_2 > MRPL_2$, it follows that $W_1 < MRPL_1$.

Now impose $W_1 = MRPL_1$. To preserve $NPV_W = NPV_{MRPL}$, the contract must become less backloaded, implying that W_2 decreases. From the incentive-compatibility condition,

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

and since $\bar{\tilde{U}}_2$ is increasing in W_2, it follows that $\tilde{e}$ is increasing in W_2. Hence,

$$W_2 \downarrow \quad \Rightarrow \quad \tilde{e} \downarrow.$$

Because effort is productive, lower $\tilde{e}$ reduces output for any given employment levels:

$$Q_s = F_s(L_s, L_{-s}, \tilde{e}, e_2^*), \quad \frac{\partial F_s}{\partial \bar{e}_1} > 0,$$

so Q_s decreases holding (L_1, L_2) fixed.

Under perfect competition, labor supply is perfectly elastic, so employment is determined by labor demand. The reduction in effort lowers the marginal revenue product of labor, leading firms to reduce employment:

$$L_1 \downarrow, \quad L_2 \downarrow.$$

This further reduces output:

$$Q_s \downarrow, \quad s \in \{1, 2\}.$$

Finally, with downward-sloping demand, lower output implies higher prices:

$$P_s \uparrow, \quad s \in \{1, 2\}.$$

□

A.3 Proof of Corrolary 1

Proof. Under monopsony, Proposition 2 implies that the equilibrium contract features a lifecycle wage markdown:

$$NPV_W < NPV_{MRPL}.$$

Therefore, raising W_1 toward $MRPL_1$ tends to relax the entry distortion, increasing junior employment and, through workforce dynamics, senior employment as well. This force tends to increase output and reduce prices, as in Proposition 3.

At the same time, because effort is productive, senior compensation affects junior incentives. By Lemma 1, junior effort is strictly increasing in W_2. Hence, if imposing $W_1 = MRPL_1$ compresses the wage profile by reducing backloading, then junior effort falls. Since effort is productive, this tends to reduce output and labor demand, as in Proposition 4.

Thus, in the presence of both monopsony and productive effort, two opposing forces are at work. The first is the standard monopsony channel, through which raising W_1 alleviates the hiring distortion and tends to raise employment and output. The second is the incentive channel, through which compressing the wage profile weakens effort incentives and tends to reduce employment and output. Without additional assumptions on the relative strength of these channels, the net effects on employment, output, and prices are ambiguous. □

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:

$$L_{if} = s_{fg} L_{ig}, \qquad \boldsymbol{V}_f = s_{fg} \boldsymbol{V}_g, \tag{OA.4}$$

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

$$\begin{aligned} Q_{ig} &= \sum_{f \in g} \left[\exp(\omega_{if} + \varepsilon_{if}) \prod_k (L_{ikf})^{\beta_{ik}} \boldsymbol{V}_{if}^{\beta_i^v} \right] \\ &= \left[\sum_{f \in g} \exp(\omega_{if} + \varepsilon_{if}) \, s_{fg}^{\sum_k \beta_{ik} + \beta_i^v} \right] \prod_k L_{ikg}^{\beta_{ik}} \boldsymbol{V}_{ig}^{\beta_i^v}. \end{aligned} \tag{OA.5}$$

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

$$q_{ig} = \sum_k \beta_{ik} l_{ikg} + \beta_i^v \boldsymbol{v}_{ig} + \omega_{ig}, \tag{OA.6}$$

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

$$\omega_{ig} = \sigma_i \tilde{\omega}_{ig} + \nu_{ig}, \tag{OA.7}$$

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

$$\nu_{ig} = q_{ig} - \sigma_i \tilde{q}_{ig} - \Big(\sum_k \beta_{ik} l_{ikg} - \sigma_i \sum_k \beta_{ik}\tilde{l}_{ikg}\Big) - \Big(\beta_i^v \boldsymbol{v}_{ig} - \sigma_i \beta_i^v \tilde{\boldsymbol{v}}_{ig}\Big). \tag{OA.8}$$

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

$$E(\nu_{ig} \mid \boldsymbol{L}_{ig}, \tilde{\boldsymbol{L}}_{ig}) = 0. \tag{OA.9}$$

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

$$E(\nu_{ig} \mid \tilde{\boldsymbol{V}}_{ig}) = 0. \tag{OA.10}$$

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.6) 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 Productivity spillovers

In the main empirical specification, we did not allow for effort spillovers across worker types. In this Appendix, we extend the model to allow for such spillovers. We decompose $\omega_{s,f}$ into an effort component, $h_s(\bar{e}_{s,f}, \bar{\boldsymbol{e}}_{-s,f})$, and a component that is unrelated to effort, $\tilde{\omega}_{s,f}$.

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

By including effort of other worker types in $h_s(\cdot)$, we allow for cross-type effort spillovers in productivity. However, we still restrict the non-effort component $\tilde{\omega}_{s,f}$ to be free of such spillovers—though we do allow this term to be correlated arbitrarily across worker types.

In our model, effort of junior workers (professional staff excluding partners) is a function of senior (partner) wages $W_{f,s=7}$ and partner promotion probabilities $Pr_{s=7}$. Effort of senior partners is exogenous. Therefore, we write productivity as

$$\omega_{s,f} = \tilde{\omega}_{s,f} + h_s(W_{f,s=7}, Pr_{s=7}, e_{s=7}).$$

We impose a linear specification for the function h_s

$$\omega_{s,f} = \tilde{\omega}_{s,f} + h_s^w w_{f,s=7} + h_s^p \ln(Pr_{s=7}) + h_s(e_{s=7}).$$

We estimate the coefficients (h_s^w, h_s^p) by including log partner wages and log partner promotion probabilities in the right-hand side of the production functions, and include lagged log senior wages and lagged log promotion probabililities in the instruments vector. We treat $\tilde{\omega}_{s,f} + h_s(e_{s=7}) + \epsilon_{s,f}$ as the productivity residual on which we impose the AR(1) assumption, rather than $\omega_{s,f} + \epsilon_{s,f}$.

The resulting estimates are in Table OA-1. The estimates are very similar to the main specification and delivers very similar wage markdowns.

Table OA-1: Production Estimates with Effort Spillovers across Worker Types

	(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	0.941	0.063	0.005	0.041	-0.048	0.061	0.004	0.081	-0.057	0.081	-0.057	0.081	-0.061	0.217
3-5 years	0.045	0.028	1.027	0.022	0.004	0.033	0.015	0.024	0.031	0.048	0.031	0.048	0.011	0.071
6-8 years	0.042	0.043	0.014	0.021	1.044	0.042	-0.071	0.066	0.059	0.052	0.059	0.052	0.056	0.144
9 years	-0.022	0.039	-0.024	0.016	-0.021	0.038	0.977	0.033	-0.019	0.045	-0.019	0.045	0.006	0.103
Non-eq. Partners	0.012	0.020	0.008	0.012	0.003	0.020	-0.015	0.018	0.983	0.024	-0.017	0.024	-0.028	0.054
Eq. Partners	0.102	0.074	0.084	0.038	0.181	0.068	0.077	0.076	0.217	0.090	1.217	0.090	0.267	0.260
Paraprof.	-0.145	0.063	-0.074	0.025	-0.105	0.059	-0.023	0.056	-0.128	0.071	-0.128	0.071	0.659	0.197
Interns	0.011	0.033	0.008	0.020	-0.042	0.029	0.002	0.035	-0.022	0.039	-0.022	0.039	0.062	0.090
Materials	0.008	0.019	-0.011	0.013	0.012	0.026	0.030	0.035	-0.030	0.026	-0.030	0.026	0.049	0.067
Admin. staff	-0.003	0.023	-0.043	0.015	-0.027	0.025	0.016	0.018	-0.044	0.031	-0.044	0.031	0.009	0.077
Ser. corr.	0.814	0.064	0.713	0.095	0.821	0.070	0.936	0.054	0.809	0.062	0.809	0.062	0.852	0.064
Obs.	174		174		174		174		174		174		174	
174														

B.3 Promotion Probabilities

B.3.1 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.3.2 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.3.3 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:

$$\frac{N_{\mathrm{i}}}{N_{\mathrm{i}}^{\mathrm{new}}} = \frac{1 - (1 - \tau_{\mathrm{i}})^{T_i}}{\tau_{\mathrm{i}}} \tag{OA.11}$$

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.4 Alternative Specifications and Robustness

B.4.1 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.[31]

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

[31] https://www.glassdoor.com/Salaries/accounting-intern-salarySRCH_KO0,17.htm

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.[32].

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-2. The production 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.

Table OA-2: Production Estimates with Fixed Costs in Production

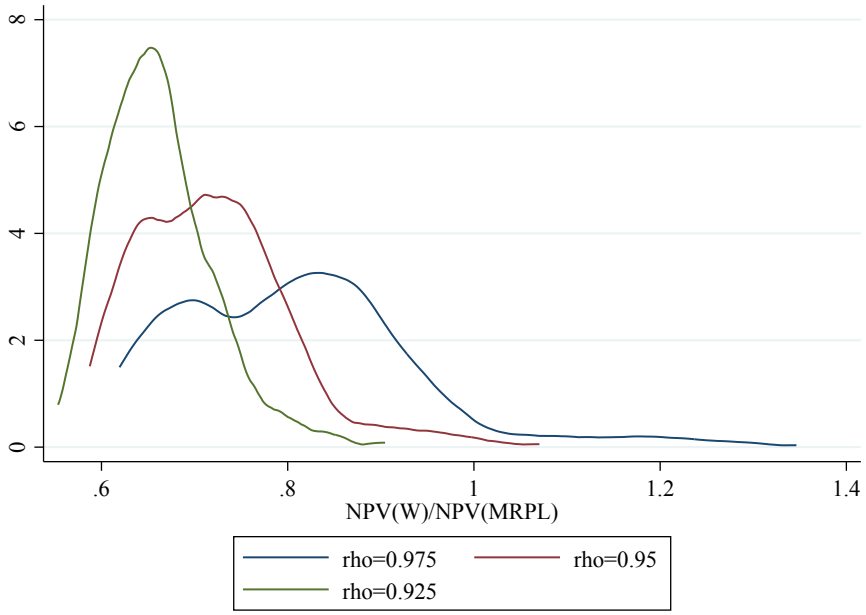
	(I) 0-2 years		(II) 3-5 years		(III) 6-8 years		(IV) 9+ years		(V) Non-eq. Part.		(VI) Eq. Part.		(VII) Paraprof.	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
0-2 years	0.969	0.045	-0.026	0.045	-0.024	0.042	-0.050	0.049	-0.069	0.078	-0.069	0.078	0.089	0.149
3-5 years	0.031	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		165		165		165	
165														

B.4.2 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-1. 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.

[32]https://www.cpapracticeadvisor.com/2014/12/16/how-much-admin-expense-should-accounting-firms-incur/17560/

Figure OA-1: Robustness: Discount Factors



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

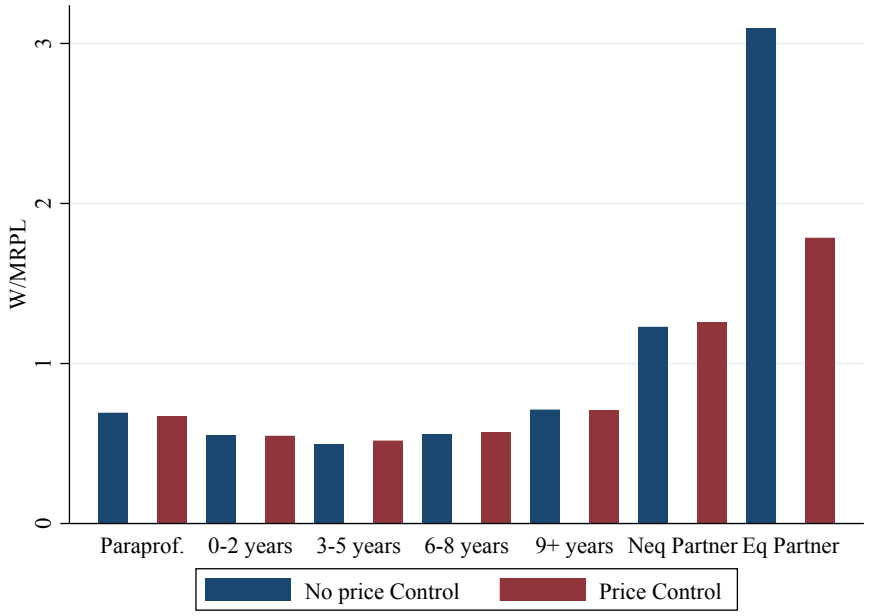
B.4.3 Product Differentiation

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).

$$q_{if} = \sum_{k \in I} \left(\beta_{ik} l_{ikf}\right) + \beta_i^v \boldsymbol{v}_{if} + \beta^p \ln(p_{if}) + \omega_{if} + \varepsilon_{if}. \tag{OA.12}$$

The resulting wage markdowns are very similar to those estimated without price control, as shown in Figure OA-2.

Figure OA-2: Including Price Controls



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

B.4.4 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.13), and a translog specification, Equation (OA.14)

$$r_f = \sum_{k \in I} \left(\beta_k l_{kf} \right) + \beta^o \boldsymbol{o}_f + \omega_f + \varepsilon_f. \tag{OA.13}$$

$$\begin{aligned} r_f = \beta_0 &+ \sum_{k \in I} \beta_k \, l_{kf} + \beta^o \, o_f \\ &+ \tfrac{1}{2} \sum_{k \in I} \sum_{j \in I} \gamma_{kj} \, l_{kf} l_{jf} \\ &+ \omega_f + \varepsilon_f. \end{aligned} \tag{OA.14}$$

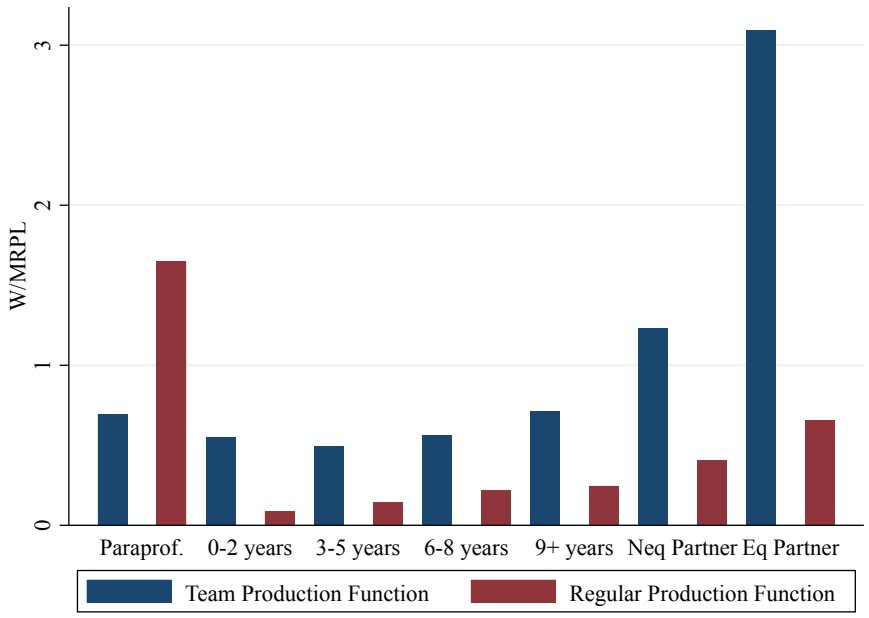
We estimate these revenue production functions with the same moment conditions and TFP transition process as assumed in the main text.

The production function estimates are in Table OA-3, and the resulting markdowns in Figure OA-3. 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.

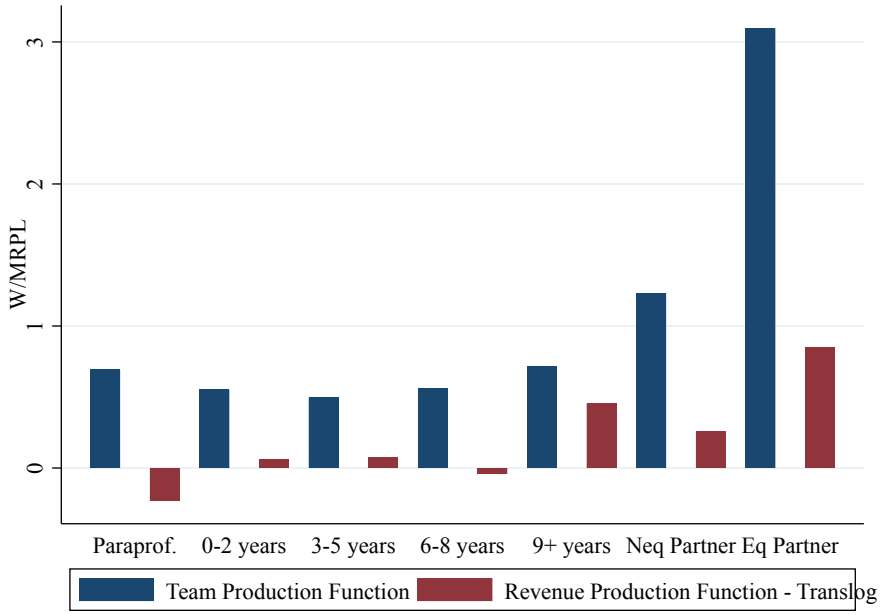
We think that the failure of the firm-level production model to generate realistic marginal revenue products stems from two sources. First, the input-specific team production functions do not aggregate to a firm-level Cobb-Douglas or translog specification, even if the product market would be in perfect 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-3: 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-3: Firm-level Production Function

	(I) Cobb-Douglas		(II) Translog	
	Est.	S.E.	Est.	S.E.
0-2 years	0.270	0.088	-1.506	2.241
3-5 years	0.202	0.034	1.361	1.199
6-8 years	0.121	0.056	-1.573	0.772
9 years	0.229	0.038	0.452	0.730
Non-eq. Partners	0.066	0.018	0.539	0.594
Eq. Partners	0.238	0.088	2.632	1.816
Paraprof.	0.006	0.054	-0.228	0.390
Interns	-0.032	0.029	-0.037	0.026
Materials	0.029	0.016	0.011	0.010
Admin. staff	-0.065	0.024	-0.007	0.094
Ser. corr.	0.885	0.052	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	

B.5 Bundle Pricing

B.5.1 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:

$$Q_f = [\sum_{i=1}^{I} a_i Q_{if}^{\frac{\sigma-1}{\sigma}}]^{\frac{\sigma}{\sigma-1}} \tag{OA.15}$$

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

$$Q_f = \overline{Q}\left(\frac{P_f}{\overline{P}}\right)^{\eta} \zeta_f. \tag{OA.16}$$

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

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

B.5.2 Recovering Product Weights

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

$$\frac{Q_i}{Q_j} = \left(\frac{a_i}{a_j}\right)^{\sigma}\left(\frac{P_i}{P_j}\right)^{-\sigma}. \tag{OA.17}$$

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

$$\kappa_i \equiv \left(\frac{Q_i}{Q_7}\right)^{1/\sigma}\frac{P_i}{P_7}, \qquad i = 1, \ldots, 6. \tag{OA.18}$$

The relative demand conditions then become

$$a_i = \kappa_i a_7, \qquad i = 1, \ldots, 6. \tag{OA.19}$$

We impose the normalization

$$\sum_{i=1}^{7} a_i = 1. \tag{OA.20}$$

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

$$a_i = \frac{\kappa_i}{1 + \sum_{k=1}^{6} \kappa_k}. \tag{OA.21}$$

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

$$a_7 = \frac{1}{1 + \sum_{i=1}^{6} \kappa_i}, \tag{OA.22}$$

B.5.3 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-4. 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.

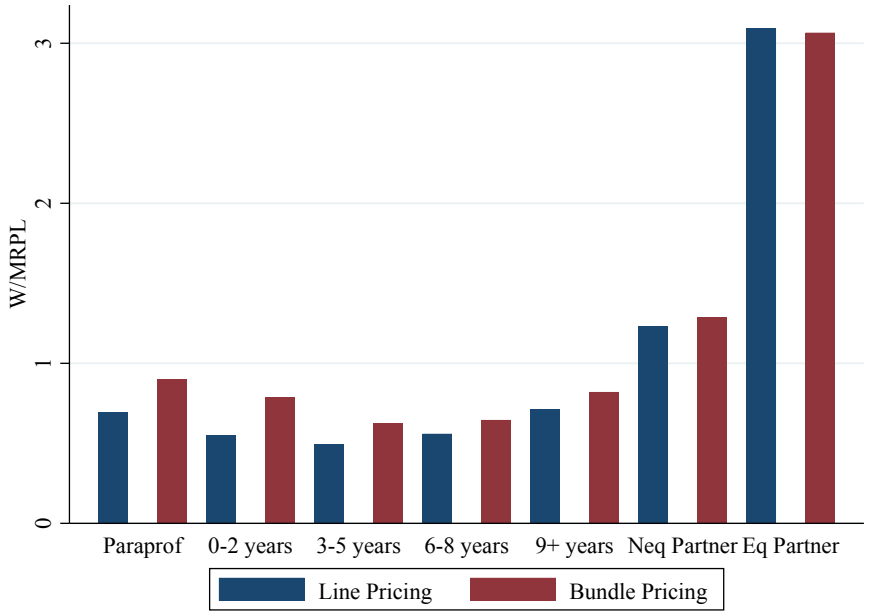
B.6 Observed vs. Model-implied Seniority Levels

Denote the observed seniority levels as $\tilde{s}$ and the 'true' seniority levels at which level firms set wages as s. Our result in Appendix shows that from period $s = 2$ onward, the NPV of the MRPL exceeds the NPV of wages. Therefore, we can determine the mapping from $\tilde{s}$ to s by evaluating the wedge between the NPV of MRPL and wage at any observed seniority level $\tilde{s}$.

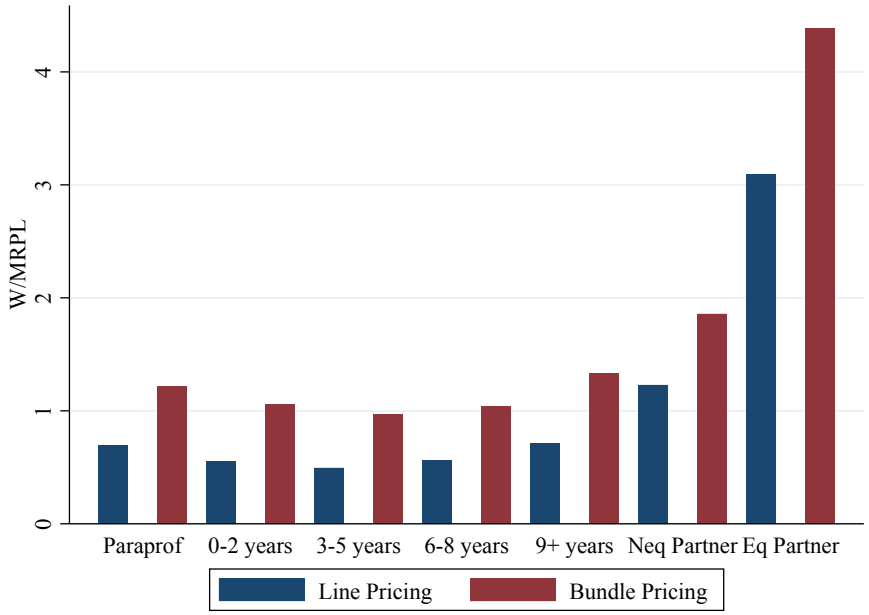
The results of this exercise are in Figure OA-5. We find that the expected wage starts exceeding the expected MRPL around 9 years of seniority. This implies that the promotion to seniority level $s = 2$ does not happen at the observed seniority levels $\tilde{s} \in (0-2\text{ y}, 3-5\text{ y}, 6-8\text{ y})$, which therefore constitute the 'junior' seniority level $s = 1$ in the model. The promotion to the 'senior' grades $s > 1$ should happen either at $\tilde{s} \in (9+\text{ y}, \text{non-equity partner}, \text{equity partner})$. Given that the largest drop in promotion rates occurs just before the non-equity partnership, and that wages start exceeding MRPL at that seniority level, it is natural to think of all non-partner professional staff as the junior grade $s = 1$ and of non-equity and equity partners as the senior grades $s = 2$, $s = 3$.

Figure OA-4: Wage Markdowns: Bundle Pricing vs. Line Pricing

(a) $\sigma = 0.5$

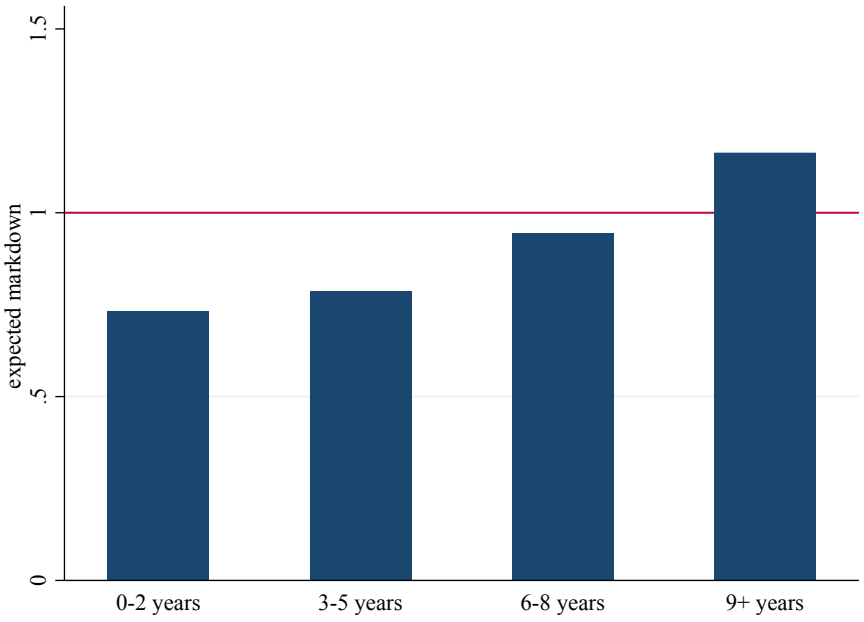


(b) $\sigma = 2$



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

Figure OA-5: Expected Wage Markdowns by Seniority



Notes: We compute the NPV of wages and MRPL at every observed seniority level $\tilde{s}$, up to the "9+ years" seniority band.

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. Similarly, we present the model as if the firm was a monopsonist in the labor market, rather than a monopsonistic competitor. Extending the analysis to monopsonistic competition in the labor market and monopolistic competition in the product market, 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

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

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

$$\bar{U}^{alt}_{s} \equiv \mathbb{E}[U^{alt}_{i,s}], \qquad \bar{\bar{U}}_{s,t} \equiv \mathbb{E}\Big[\max\{U_{s,t}, U^{alt}_{i,s}\}\Big] = G^{alt}_{s}(U_{s,t})U_{s,t} + \int_{U_{s,t}}^{\infty} u\, dG^{alt}_{s}(u). \tag{OA.23}$$

C.5 Workers' Values and Incentives

Let $\bar{U}_{s,t}$ 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

$$\bar{U}_{N-1,t} \ =\ \alpha_{N-1}W_{N-1,t} - C(e^{*}_{N-1}), \tag{OA.24}$$

where e^{*}_{N-1} is determined exogenously.[33] For any $s = 0, 1, \ldots, N-2$, a level-s worker chooses effort e_s to maximize

$$\bar{U}_{s,t} \ =\ \max_{e_s}\ \alpha_s W_{s,t} - C(e_s) + \rho\left\{\left[1 - \Phi(\theta^{P}_{s,t}(\tau_{s,t}) - e)\right]\bar{\bar{U}}_{s+1,t} + \Phi(\theta^{P}_{s,t}(\tau_{s,t}) - e)\bar{U}^{alt}_{s+1}\right\}. \tag{OA.25}$$

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

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

[33] See Footnote 15 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

$$C'(\tilde{e}_{s,t}) \;=\; \rho\,\phi(\kappa^P(\tau_{s,t}))\Big[\bar{\tilde{U}}_{s+1,t} - \bar{U}^{alt}_{s+1}\Big], \qquad s = 0, 1, \ldots, N-2, \tag{OA.26}$$

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

C.6 Employment Dynamics

The number of juniors willing to work for the firm is

$$L_{0,t} \;=\; G^{alt}_0(\bar{U}_{0,t})\,\mathcal{L}. \tag{OA.27}$$

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

$$L_{s+1,t+1} \;=\; (1-\tau_{s,t-s})\,G^{alt}_{s+1}(\bar{U}_{s+1,t-s})\,L_{s,t}. \tag{OA.28}$$

C.7 Firm's Problem

We specify the firm's problem recursively, which requires defining the state space. From (OA.27) and (OA.28), 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.[34] Finally, define $l \equiv \{\vec{l}_k\}_{k=0}^{N-1}$ and $\epsilon \equiv \{\vec{\epsilon}_k\}_{k=0}^{N-2}$, for conciseness.[35] 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

$$V(l,\epsilon) \;=\; \max_{\{W_s,\tau_s\}_{s=0}^{N-1}} \left\{ \sum_{s=0}^{N-1} p_s\,F_s(L_0, \vec{l}_0, e_s) \;-\; W_0 L_0 \;-\; \sum_{s=1}^{N-1} \rho^s\,W_s\,L^{\text{cohort}}_s \;+\; \rho\,V(l^+,\epsilon^+) \right\}, \tag{OA.29}$$

[34] 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.

[35] 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.27), the future employment path for the cohort hired this period is

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

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.26) if $s = 0$.[36] The stock of incumbent employed workers evolve as follows:

$$l_{s,k}^+ = \begin{cases} l_{s,k+1} & \text{if} \quad s \geq k+2, \\ L_s^{\text{cohort}} & \text{if} \quad s = k+1. \end{cases} \tag{OA.30}$$

Meanwhile, $\epsilon_{s,k}^+$ is equal to $\epsilon_{s,k+1}$ for $s \geq k+2$ and is given by (OA.26) 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.28) specialize to

$$\bar{\bar{U}}_s = G_s^{alt}(\bar{U}_s)\bar{U}_s + \int_{U_s}^{\infty} u\, dG_s^{alt}(u), \tag{OA.31}$$

$$\bar{U}_{N-1} = \alpha_{N-1} W_{N-1} - C(e_{N-1}^*), \tag{OA.32}$$

$$\bar{U}_s = \alpha_s W_s - C(\tilde{e}_s) + \rho\left[(1-\tau_s)\bar{\bar{U}}_{s+1} + \tau_s \bar{U}_{s+1}^{alt}\right], \quad s = 0, \ldots, N-2, \tag{OA.33}$$

$$C'(\tilde{e}_s) = \rho\, \phi(\kappa^P(\tau_s))\, [\bar{\bar{U}}_{s+1} - \bar{U}_{s+1}^{alt}], \quad s = 0, \ldots, N-2, \tag{OA.34}$$

$$L_0 = G_0^{alt}(\bar{U}_0)\mathcal{L}, \qquad L_{s+1} = (1-\tau_s) G_{s+1}^{alt}(\bar{U}_{s+1})\, L_s, \quad s = 0, \ldots, N-2. \tag{OA.35}$$

[36] 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.29) with respect to W_s is

$$\begin{aligned}
&\frac{\partial \bar{U}_0}{\partial W_s} g_0^{alt}(\bar{U}_0)\mathcal{L}\left\{MRPL_0 - W_0 - \sum_{k=1}^{N-1} \rho^k \prod_{m=0}^{k-1} \left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_k - W_k)\right\} \\
&\quad + \sum_{j=1}^{s} \rho^j \prod_{m=1}^{j} \left[(1-\tau_{m-1})G_{m-1}^{alt}(\bar{U}_{m-1})\right] \mathcal{L}\frac{\partial \bar{U}_j}{\partial W_s} g_j^{alt}(\bar{U}_j)\left\{MRPL_j - W_j\right. \\
&\quad \left. + \sum_{k=j+1}^{N-1} \rho^{k-j} \prod_{m=j}^{k-1} \left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_k - W_k)\right\} \\
&\quad - \rho^s \prod_{j=0}^{s-1} \left[(1-\tau_j)G_{j+1}^{alt}(\bar{U}_{j+1})\right] G_0^{alt}(\bar{U}_0)\mathcal{L} \\
&\quad + \sum_{j=0}^{s-1} \frac{\partial F_j}{\partial \tilde{e}_j}\frac{\tilde{e}_j}{\partial W_s} = 0. \qquad \text{(OA.36)}
\end{aligned}$$

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.29) is

$$\frac{\partial \bar{U}_0}{\partial W_0} g_0^{alt}(\bar{U}_0)\mathcal{L}\left\{MRPL_0 - W_0 + \sum_{s=1}^{N-1} \rho^s \prod_{m=0}^{s-1} \left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_s - W_s)\right\} = L_0. \qquad \text{(OA.37)}$$

Meanwhile,

$$\frac{\partial \bar{U}_s}{\partial W_s} = \alpha, \quad \text{for } s \geq 1 \qquad \text{(OA.38)}$$

$$\text{and} \quad \frac{\partial \bar{U}_k}{\partial W_s} = \rho^{(s-k)} \prod_{j=k}^{s-1} \left[(1-\tau_j)G_{j+1}^{alt}(\bar{U}_{j+1})\right]\alpha, \quad \text{for } s > k \geq 1. \qquad \text{(OA.39)}$$

From (OA.35), (OA.37) and (OA.38), we obtain

$$g_0^{alt}(\bar{U}_0)\mathcal{L}\left\{MRPL_0 - W_0 - \sum_{k=1}^{N-1} \rho^k \prod_{m=0}^{k-1} \left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_k - W_k)\right\} = \frac{G_0^{alt}(\bar{U}_0)\mathcal{L}}{\alpha}.$$

Together with (OA.39), this allows us to rewrite the first line of (OA.36) as

$$\rho^s \prod_{j=0}^{s-1} \left[(1-\tau_j) G_{j+1}^{alt}(\bar{U}_{j+1})\right] G_0^{alt}(\bar{U}_0)\mathcal{L},$$

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.36) cancel each other out, and we can rewrite the FOC as

$$\begin{aligned}\sum_{j=1}^{s} \rho^j \prod_{m=1}^{j} \left[(1-\tau_{m-1}) G_{m-1}^{alt}(\bar{U}_{m-1})\right] \mathcal{L} \frac{\partial \bar{U}_j}{\partial W_s} g_j^{alt}(\bar{U}_j) \Bigg\{ MRPL_j - W_j \\ + \sum_{k=j+1}^{N-1} \rho^{k-j} \prod_{m=j}^{k-1} \left[(1-\tau_m) G_{m+1}^{alt}(\bar{U}_{m+1})\right] (MRPL_k - W_k) \Bigg\} \\ + \sum_{j=0}^{s-1} \frac{\partial F_j}{\partial \tilde{e}_j} \frac{\partial \tilde{e}_j}{\partial W_s} = 0. \end{aligned} \tag{OA.40}$$

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.40) vanishes. Evaluating the condition for $s = 1$, we obtain

$$\rho(1-\tau_0) L_0 \frac{\partial \bar{U}_1}{\partial W_1} g_1^{alt}(\bar{U}_1) \Bigg\{ MRPL_1 - W_1 + \sum_{k=2}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G_{m+1}^{alt}(\bar{U}_{m+1})\right] (MRPL_k - W_k) \Bigg\} = 0.$$

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

$$MRPL_1 - W_1 + \sum_{k=2}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G_{m+1}^{alt}(\bar{U}_{m+1})\right] (MRPL_k - W_k) = 0. \tag{OA.41}$$

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

Next, evaluate (OA.40) for $s = 2$, which yields

$$\rho(1-\tau_0)L_0\frac{\partial \bar{U}_1}{\partial W_2}g_1^{alt}(\bar{U}_1)\Bigg\{MRPL_1 - W_1 + \sum_{k=2}^{N-1}\rho^{k-1}\prod_{m=1}^{k-1}\left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_k - W_k)\Bigg\}$$
$$+\rho^2(1-\tau_0)(1-\tau_1)L_0G_1(U_1)\frac{\partial \bar{U}_2}{\partial W_2}g_2^{alt}(\bar{U}_2)\Bigg\{MRPL_2 - W_2$$
$$+\sum_{k=3}^{N-1}\rho^{k-2}\prod_{m=2}^{k-1}\left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_k - W_k)\Bigg\} = 0.$$

By (OA.41), the first term is zero. Since the remaining prefactor is strictly positive, we obtain

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

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.40) at $s = 1$ gives

$$\rho(1-\tau_0)L_0\frac{\partial \bar{U}_1}{\partial W_1}g_1^{alt}(\bar{U}_1)\Bigg\{MRPL_1 - W_1 + \sum_{k=2}^{N-1}\rho^{k-1}\prod_{m=1}^{k-1}\left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_k - W_k)\Bigg\} < 0,$$

from which it follows that

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

That is, from the perspective of a type-1 worker, the present-discounted value of compensation until retirement exceeds the exceeds discounted marginal revenue product.

Under quadratic effort costs, the structure of incentives simplifies substantially. This additional structure allows us to propagate the previous inequality across levels.

Assumption OA-1 (Quadratic Effort Costs)**.** $C(e) = ce^2$.

Combining Assumption OA-1 with (OA.26) yields

$$\tilde{e}_s = \frac{\rho\,\phi(\kappa^P(\tau_s))\left[\bar{\bar{U}}_{s+1} - \bar{U}_s^{alt}\right]}{2c}, \qquad s = 0, 1, \ldots, N-2. \tag{OA.42}$$

Thus, equilibrium effort at level-s is linear in continuation utility. As a consequence, using (OA.31) and (OA.33), we have that

$$\frac{\partial \tilde{e}_s}{\partial W_{s'+j}} = \rho^j \prod_{m=1}^{j} \left[(1-\tau_{s'+m-1}) G^{alt}_{s'+m}(\bar{U}_{s'+m})\right] \frac{\partial \tilde{e}_s}{\partial W_{s'}}, \tag{OA.43}$$

for $s' > s$ and $j \geq 1$.

Meanwhile,

$$\frac{\partial \bar{U}_s}{\partial W_{s'+j}} = \rho^j \prod_{m=1}^{j} \left[(1-\tau_{s'+m-1}) G^{alt}_{s'+m}(\bar{U}_{s'+m})\right] \frac{\bar{U}_s}{\partial W_{s'}}, \tag{OA.44}$$

for $s' > s$ and $j \geq 1$.

An implication of (OA.43) and (OA.44) is that effort and utility respond proportionally to changes in future compensation, reflecting the linear relationship between effort and continuation utility in (OA.42). We now use this proportionality to propagate the previous inequality across levels. Assume

$$\frac{F_s}{\tilde{e}_s} > 0.$$

That is, effort matters for production at all levels. Now evaluate (OA.40) at $s = 2$ to obtain

$$\begin{aligned}
\rho(1-\tau_0) L_0 \frac{\partial \bar{U}_1}{\partial W_2} g^{alt}_1(\bar{U}_1) \Bigg\{ MRPL_1 - W_1 + \sum_{k=2}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\right] (MRPL_k - W_k) \Bigg\} \\
+ \rho^2 (1-\tau_0)(1-\tau_1) L_0 G_1(\bar{U}_1) \frac{\partial \bar{U}_2}{\partial W_2} g^{alt}_2(\bar{U}_2) \Bigg\{ MRPL_2 - W_2 \\
+ \sum_{k=3}^{N-1} \rho^{k-2} \prod_{m=2}^{k-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\right] (MRPL_k - W_k) \Bigg\} \\
+ \frac{\partial F_0}{\partial \tilde{e}_0} \frac{\partial \tilde{e}_0}{\partial W_2} + \frac{\partial F_1}{\partial \tilde{e}_1} \frac{\partial \tilde{e}_1}{\partial W_2} = 0.
\end{aligned}$$

Using (OA.40) evaluated at $s = 1$, together with (OA.43) and (OA.44), the first line and the first term in the third line cancel. Because $\partial F_s / \partial \tilde{e}_s > 0$, the remaining term in the third line is strictly positive, implying

$$MRPL_2 + \sum_{k=3}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\right] MRPL_k < W_2 + \sum_{k=2}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\right] W_k.$$

Repeating this argument for $s = 3, \ldots, N-1$ yields

$$MRPL_s + \sum_{k=s+1}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\right] MRPL_k$$
$$< W_s + \sum_{k=s+1}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\right] W_k.$$

Hence, at every level of the hierarchy, discounted wages exceed discounted marginal revenue product. In particular, at the top level,

$$W_{N-1} > MRPL_{N-1}.$$

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

$$\psi_0 \equiv \frac{g^{alt}_0(\bar{U}_0)\, W_0}{G^{alt}_0(\bar{U}_0)} \cdot \frac{\partial \bar{U}_0}{\partial W_0}, \qquad \frac{\partial \bar{U}_0}{\partial W_0} = \alpha_0, \tag{OA.45}$$

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

$$\frac{\partial V(l,\epsilon)}{\partial l_{s,k}} = \rho^k MRPL_s, \quad k \in \{0, \ldots, N-2\}, \quad s \in \{k+1, \ldots, N-1\}. \tag{OA.46}$$

Substituting (OA.35) and (OA.45) on (OA.37), we obtain

$$MRPL_0 + \sum_{s=1}^{N-1} \rho^s \prod_{m=0}^{s-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\mathcal{L}\right] MRPL_s$$
$$= W_0 + \sum_{s=1}^{N-1} \rho^s W_s \prod_{m=0}^{s-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\mathcal{L}\right] + \frac{W_0}{\psi_0}. \tag{OA.47}$$

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

$$\text{NPV}_W \equiv W_0 + \sum_{s=1}^{N-1} \rho^s W_s \prod_{m=0}^{s-1} \left[(1-\tau_m) G_{m+1}^{alt}(\bar{U}_{m+1})\mathcal{L}\right], \tag{OA.48}$$

$$\text{NPV}_{\text{MRPL}} \equiv MRPL_0 + \sum_{s=1}^{N-1} \rho^s \prod_{m=0}^{s-1} \left[(1-\tau_m) G_{m+1}^{alt}(\bar{U}_{m+1})\mathcal{L}\right] MRPL_s. \tag{OA.49}$$

Then (OA.47) implies the exact wedge

$$\text{NPV}_{\text{MRPL}} \;=\; \text{NPV}_W \;+\; \frac{W_0}{\psi_0}. \tag{OA.50}$$

By defining the long-run elasticity of labor supply as

$$\psi_{LR} = \psi_0 \frac{NPV_W}{W_0} \tag{OA.51}$$

and substituting on (OA.50), 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.

To facilitate the exposition, we present the model as if the firm was a monopsonist in the labor market, rather than a monopsonistic competitor. Extending the analysis to the case of monopsonistic competition, as we do in Section 3, is straightforward.

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 is a mass $\mathcal{H}$ of senior workers available for lateral hiring, in addition to the $\mathcal{L}$ 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. 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_2^{alt}, identical to the distribution faced by workers initially hired as juniors upon reaching the senior stage (irrespective of promotion outcomes). We maintain the assumption that outside-option draws are independent across workers and periods.

In the product market, the firm is a monopolistic competitor and faces a residual demand as specified in Section 3, with marginal revenue function given by (3).

D.2 Worker and Firm Behavior

Junior workers' effort choices, given the compensation–promotion policy offered by the firm, continue to be characterized by (6). For simplicity, we assume that all senior workers—regardless of whether they were promoted internally or hired laterally—exert the same effort, e_2^*. The resulting flow utilities in period t are

$$U_{2,t-1} = \alpha W_{2,t-1} - C(e_2^*) \quad \text{and} \quad U_{lat,t} = \alpha W_{lat,t} - C(e_2^*),$$

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 (7). Substituting for promotions and lateral hiring, the total number of senior workers employed in period t is

$$L_{2,t} = (1-\tau_{t-1})G_2^{alt}(U_{2,t-1})\,L_{1,t-1} + G_2^{alt}(U_{lat,t})\,\mathcal{H}.$$

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.

The firm's problem can now be written as

$$\begin{aligned} V(l_{2,t}) = \max_{W_{1,t},W_{2,t},\tau_t,W_{lat,t}} \Big\{ & R_1\left(L_{1,t}, l_{2,t}+lat_t, \tilde{e}_t, e_2^*\right) + R_2\left(l_{2,t}+lat_t, L_{1,t}, e_2^*, \tilde{e}_t\right) \\ & - W_{1,t}L_{1,t} - \rho W_{2,t}(1-\tau_t)G_2^{alt}(U_{2,t})L_{1,t} - W_{lat,t}lat_t + \rho V\left[(1-\tau_t)G_2^{alt}(U_{2,t})L_{1,t}\right]\Big\} \end{aligned} \quad \text{(OA.52)}$$

with $\tilde{e}_t$ and $L_{1,t}$ given by (6) and (7), respectively, and

$$lat_t = G_2^{alt}(U_{lat,t})\,\mathcal{H}.$$

Crucially, from the envelope condition, the marginal value of seniors is

$$V'(l_{2,t}) = \frac{1+\eta}{\eta} d\left\{\left[F_1(L_{1,t}, l_{2,t}+lat_t, \tilde{e}_t, e_2^*)\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_{2,t}} + \left[F_2(L_{1,t}, l_{2,t}+lat_t, e_2^*, \tilde{e}_t)\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_{2,t}}\right\} \quad \text{(OA.53)}$$

analogous to the baseline model in Section 3.

D.3 Testable Implications

The steady-state conditions mirror those in Section 3, with the additional requirement that lateral hiring and the utility of laterally hired workers are constant over time,

$$lat_t = lat, \quad U_{lat,t} = U_{lat}.$$

We then have

$$L_2 = (1-\tau)G_2^{alt}(U_2)\,L_1 + lat.$$

That is, the stock of senior workers reflects both internal promotion and lateral hiring inflows.

Define

$$\text{MRPL}_1 \equiv \frac{1+\eta}{\eta} d \left\{ \left[F_1(L_1, L_2, \tilde{e}, e_2^*) \right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_1} + \left[F_2(L_2, L_1, e_2^*, \tilde{e}) \right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_1} \right\}$$
$$\text{MRPL}_2 \equiv \frac{1+\eta}{\eta} d \left\{ \left[F_1(L_1, L_2, \tilde{e}, e_2^*) \right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_2} + \left[F_2(L_2, L_1, e_2^*, \tilde{e}) \right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_2} \right\}.$$

One can verify that the first-order conditions associated with (OA.52) coincide with those in Section A.1. Propositions 1 and 2 therefore continue to hold in the presence of lateral hiring.

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.