Monopsony and Backloaded Compensation: Theory and Evidence from Public Accountants

Michael Rubens* Bernardo Silveira†

July 27, 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 among early-career workers and are thus efficient. To reconcile—and empirically distinguish—these two mechanisms, we build a dynamic model incorporating labor market power and endogenous effort. Estimating a team production model on novel data on U.S. public accounting firms, we find evidence of both: markdowns for junior workers and markups for senior ones reflect incentive-providing backloading, while monopsony power induces a 'lifetime' wage markdown of 15%.

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

*Boston College, NBER, and CEPR, rubensmi@bc.edu
†University of Zurich and NBER, bernardo.silveira@econ.uzh.ch
We thank Daniel Haanwinckel, John Asker, Moritz Meyer-ter-Vehn, Charles Angelucci, Thi Mai Ahn Nguyen, Oscar Volpe, and seminar participants at MIT, IIOC, BU, Stockholm U., UCLA, and RIDGE IO for useful feedback.

1 Introduction

There is increasing evidence of monopsony power in labor markets (Card, 2022; Azar and Marinescu, 2024; Caldwell et al., 2025; Kline, 2025). The standard model focuses on how market power distorts employment and wages, treating worker effort as exogenous (Card et al., 2018; Berger et al., 2022). Under this view, the wage 'markdown' below the worker's marginal revenue product reflects the employer's market power, generating deadweight loss and misallocation (Berger et al., 2022; Lamadon et al., 2022). A distinct literature, however, casts markdowns in a different light. In models of backloaded incentive pay—which typically assume competitive labor markets—firms pay junior workers below their marginal product while promising above-market compensation later in their careers (Lazear, 1981; Lazear and Rosen, 1981; Malcomson, 1984). Workers who shirk and are dismissed forfeit this future premium, so the prospect of deferred rents disciplines current effort. In this framework, markdowns are an equilibrium incentive device, and removing them can weaken effort and reduce welfare.

In this paper, we bring these two strands of the literature together by empirically testing whether wage markdowns arise from monopsony power, backloaded incentive pay, or both. To this end, we construct a partial-equilibrium 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, following Malcomson (1984), and monopsonistic competition among differentiated firms, following 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 at the firm. Workers have idiosyncratic preferences over non-wage amenities and draw heterogeneous outside offers each period, generating an upward-sloping residual labor supply curve. Firms set wages for junior and senior workers and choose promotion thresholds based on noisy performance signals.

Solving the model 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 *present-discounted-compensation* elasticity of labor supply faced by the firm when hiring junior workers—that is, the elasticity of labor supply with respect to the net-present value of compensation workers expect to obtain throughout their entire tenure within the firm. This provides a dynamic analogue

to the static labor supply model of Card et al. (2018), in which wage markdowns are a function of the elasticity of labor supply with respect to *current* wages 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. Intuitively, since all workers enter the firm through the junior position, the firm can exercise its monopsony power at the point of entry; senior compensation is thus reserved for retention and incentive provision.

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. A unique feature of our dataset is that we observe both costs and revenues at the input level (for the different labor categories that are billed to the clients), unlike commonly-used administrative data in which costs are input-specific, but revenue is aggregated across inputs. Moreover, we observe quantities and prices at both the cost and revenue side.

Our empirical strategy focuses on estimating wage markdowns by worker seniority level while imposing a minimal set of assumptions on the labor supply and demand side. We specify and estimate a multi-output production model that allows for output differentiation by worker seniority, and for team production. 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 seniority level, 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. Wages of entry-level employees are marked down by 38% below their marginal revenue product, whereas employees begin earning

substantial wage mark*ups* after approximately 15 years of service. This indicates that backloaded incentive pay contracts are in use.[1]

Our second main finding is that the net present value (NPV) of professional staff wages at career entry is on average 15% below the NPV of the marginal revenue product of labor, implying a statistically significant lifetime wage markdown. This result provides direct evidence of monopsony power—in contrast to classical models of backloaded pay under perfect labor market competition, in which the NPV of wages equals the NPV of the MRPL. Using the model, we convert these lifetime markdowns into current- and present-discounted-compensation residual labor supply elasticities. We find that the residual labor supply to current compensation is very low, with an average elasticity of 0.4. That is, workers choose firms primarily based on the prospect of very high future wages rather than the much lower entry-level pay. In contrast, the present-discounted-compensation elasticity of labor supply is substantially higher, with an average elasticity of 5.4—yet, as 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 26% on average. For instance, the wage markdown for entry-level employees would shrink from 0.62 (wages 38% below MRPL) to 0.78 (wages 22% below MRPL). This exercise suggests that 42% of the entry-level worker's wage markdown is due to monopsony power and 58% 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 compress markdown dispersion within firms, such as pay-ratio caps, or unionization, can either increase or decrease welfare depending on the relative magnitudes of the monopsony and incentive-wage mechanisms in driving wage markdowns. Fourth, in labor markets with implicit wage contracts, the relevant metric for monopsony

[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.

power is the present-discounted-compensation elasticity of labor supply, rather than the current-compensation elasticity, as the expected markdown is set as a function of the former. Therefore, using estimated labor supply elasticities to current wages only can lead to substantial overestimation of the true degree of monopsony power.

These implications extend to other industries in which careers follow a tournament structure. Beyond public accounting, prominent examples include law firms, strategy consulting, and other professional services, which together account for 7% of non-farm U.S. employment (U.S. Bureau of Labor Statistics, 2026). Tournament features also characterize many high-wage occupations—including physicians, academics, investment bankers, and software engineers—a segment of the labor market that the prior monopsony literature has largely overlooked in favor of low-wage workers (Manning, 2021).

This paper contributes to four main strands of literature. First, we build on monopsony models of horizontally differentiated employers in the tradition of Card et al. (2018), in which markdowns are the result of preference heterogeneity of workers (Azar et al., 2022; Berger et al., 2022; Lamadon et al., 2022; Goolsbee and Syverson, 2023; Roussille and Scuderi, 2021; Volpe, 2024). In these static models of labor markets, wage markdowns are proportional to current-compensation 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).[2] We contribute to this literature by incorporating backloaded-pay incentive contracts with forward-looking workers and firms into the differentiated-firms monopsony model.[3]

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, 1980a; Harris and Nguyen, 2025). On the theory side, 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 seniority levels, a feature that is crucial for bringing the model to the data. Empirically, whereas most prior empirical examination of wage backloading has relied on proxies of marginal revenue products (Medoff and Abraham, 1980b, 1981;

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

[3]Our paper also relates to prior papers that have incorporated efficiency pay into models of imperfect labor market competition (Bowles, 1985; Manning, 2003; Emanuel and Harrington, 2026). In contrast to these models, in which workers are incentivized by current pay, we consider monopsony power in settings in which firms incentivize forward-looking workers by committing to future wages and promotion rates.

Kotlikoff and Gokhale, 1992; Flabbi and Ichino, 2001; Dohmen, 2004), we exploit data on seniority-specific inputs and output to estimate wage-to-MRPL ratios at various stages of employees' careers, finding evidence for backloaded incentive-wage contracts.[4]

Third, we contribute to the literature on estimating wage markdowns using production models (Brooks et al., 2021; Morlacco, 2019; Yeh et al., 2022; Mertens, 2022; Delabastita and Rubens, 2025; Syverson, 2025). Whereas this literature has focused on single-output setups, we estimate a multi-output production function that accommodates team production and output differentiation by worker seniority. This involves estimating a full matrix of own- and cross-output elasticities across seniority levels, building on recent empirical studies that assess market power through multi-product production models (De Loecker et al., 2016; Dhyne et al., 2022; Orr, 2022; Valmari, 2023).[5] We further depart from prior work on production-function estimation of markups and markdowns—which has focused predominantly on manufacturing industries—by examining professional services firms instead.[6]

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 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 current- and present-discounted-compensation labor supply elasticities. Section 5 concludes.

[4] To the best of our knowledge, previous empirical studies of the evolution of wage markdowns that estimated production functions found no conclusive support for backloading—see, for example, Hellerstein and Neumark (1995), Hellerstein et al. (1999), and Cardoso et al. (2011). Using employer-employee matched data from the United States, Hellerstein and Neumark (2007) do find evidence that compensation increases with age at a faster rate than productivity, but they stop short of showing that the compensation of senior employees actually surpasses their marginal revenue products.

[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 workers' seniority levels. Whereas this has the benefit of providing 'conduct-free' markdown estimates by worker seniority, it comes at the cost of assuming away within-seniority worker heterogeneity.

[6] Another recent paper using production-cost data from professional services firms is Verboven and Yontcheva (2024).

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.[7] 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, 2023c).

Public accounting firms are typically organized as partnerships. 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 a partnership. In addition to this professional staff, firms also employ 'paraprofessionals'—who are not in the partnership track—as well as interns and administrative staff.

Given the up-or-out system, professional staff turnover is high, at an annual 15% rate. The few who make it to the partnership usually have around 15 years of experience before reaching that stage. 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. Lateral mobility is limited, with merely 14% of new equity partners drawn from another firm (Inside Public Accounting, 2023c). Equity partners enjoy high job stability, and usually have the option of remaining in their position until retirement.[8] Most firms impose a mandatory retirement age for equity partners (Inside Public Accounting, 2023c).

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.

Public accounting employees are typically paid on a salaried rather than hourly basis and, for the most part, do not receive overtime pay (The CPA Journal, 2017). The majority

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

[8] Public accounting firms have very recently started to change this custom by demoting or firing equity partners (Kissin, 2026), but not during our sample period.

of firms offer no profit-sharing incentives to their employees (Inside Public Accounting, 2023b). Non-equity partners also receive a salary—but, in addition, usually a bonus (Rosenberg Associates, 2013). Equity partners are, in contrast, owners of the firms. They 'buy into' the firm by purchasing an ownership stake when being promoted to the equity partnership and redeem it upon retirement. Equity partners are compensated according to formulas combining non-discretionary components (based on metrics such as charge hours, billing budgets, and business development) and discretionary ones, typically decided upon by a compensation committee (Inside Public Accounting, 2023c). The average annual compensation per seniority level are in Appendix Figure OA-9.

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, hours worked, and annual compensation (USD). Compensation includes bonuses and profit-sharing agreement, which are most commonly offered to equity partners. We also observe the buy-in amounts required from equity partners. Regarding hours, one caveat 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.[9]

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 as a share of revenue for marketing, training and continued professional education, recruiting, and technology costs. 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

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

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.

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	0.017	0.008
Paraprofessionals (Share of Employment)	0.055	0.052	0.025
Admin staff (Share of Employment)	0.166	0.168	0.020
Professionals 0-2 Years (Share of Employment)	0.156	0.155	0.025
Professionals 3-5 Years (Share of Employment)	0.157	0.156	0.027
Professionals 6-8 Years (Share of Employment)	0.107	0.103	0.023
Professionals 9+ Years (Share of Employment)	0.178	0.179	0.033
Non-Equity Partners (Share of Employment)	0.049	0.049	0.019
Equity Partners (Share of Employment)	0.117	0.115	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 is common in professional services industries. We compute firm-specific promotion probabilities by assuming a steady-state labor force and using data on annual promotions, and average these probabilities across firms.[10] As documented in Appendix Figure OA-11, the industry-wide employment shares of the various job titles are very stable over time, which lends credibility to the steady-state assumption when computing promotion probabilities.

The resulting promotion probabilities for an entry-level employee are in Figure 1. The ex-ante probability of still being at the firm after 2 years is 66%, this reduces to 46% after 5 years, and so on. Merely 10% of starters ultimately become non-equity partners, and 7% become equity partners. There is a sharp drop in the promotion probabilities right before reaching non-equity partnership.

As mentioned earlier, 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 tend to transition into in-house accounting positions outside the public accounting industry (Dalton et al., 2022).

Fact 2: Spreads between compensation and billing revenue increase with seniority

We compute the spread between compensation and billing revenue for employees of seven different seniority levels, denoted by s:[11]

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

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), $\delta_s \equiv W_s/MRPL_s$.[12] As mentioned earlier, we measure total annual compensation, with the only caveat that the difference between the buy-in and buy-out of equity partners is not included. We come back to interpreting equity partner compensation in Section 3.6.1.

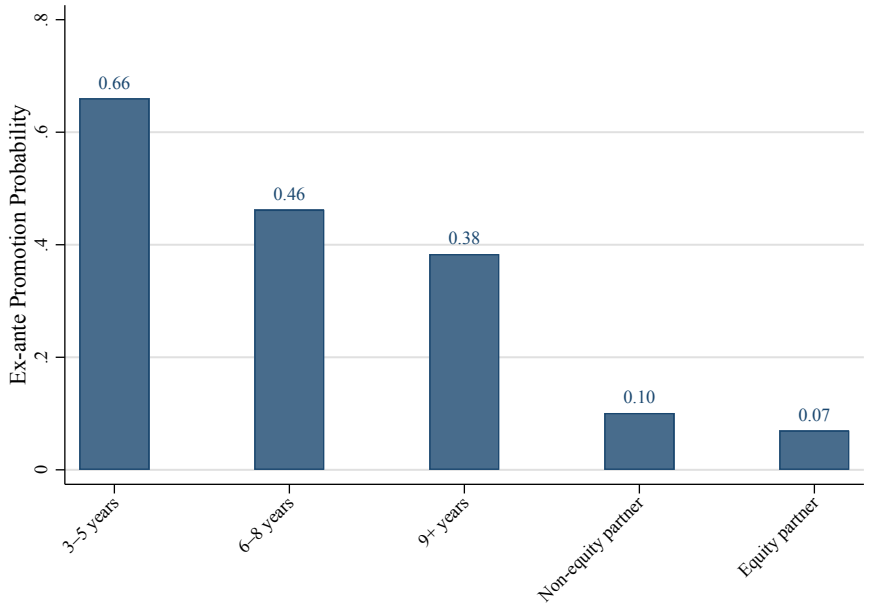
Figure 2 shows the salary-billing spreads for the various employee levels. Upon joining

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

[11] 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.

[12] Under this assumption, production for seniority-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 wage per head, P_s billing rates, L_s employment, and Q_s billable hours.

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.2.

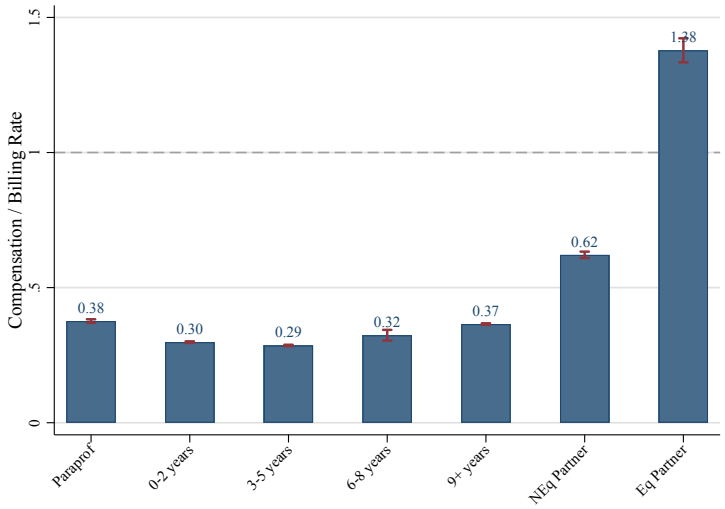
the firm, professional staff in the 'partner-track' are paid on average 30% of their respective billing revenue. As seniority increases, this compensation rate slightly raises, up to 37% for staff with 9+ years of experience. Non-equity partners are paid 62% of what they are billed. In contrast, equity 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, four 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, some of the hours worked by employees, and especially by partners, likely do not generate billing hours directly, but do so indirectly through team production. We provide further evidence of this in the next stylized fact. Fourth, even if these spreads did identify wage markdowns, they could reflect either monopsony power or backloaded compensation.

Fact 3: The Ratio of Billing Hours to Worked Hours Decreases After 5 Years of Seniority

Finally, in Figure 3, we compare the ratio of billed hours to worked hours by seniority for the 'partnership-track' employees. The share of billable hours slightly increases at first

Figure 2: Compensation-Billing Spreads and Seniority



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

until 5 years of seniority (which might be due to initial training), but then sharply decreases over time. Partners produce far fewer billing hours per hour worked than non-partner employees. 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. This is yet another reason why marginal and average revenue products of labor likely differ.

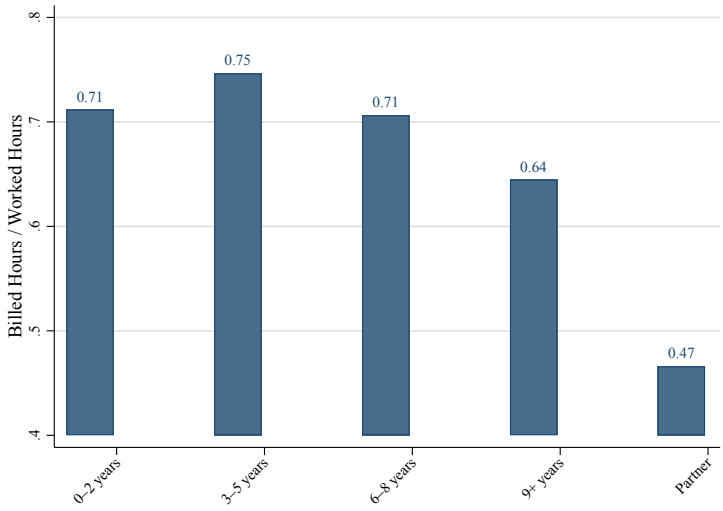
Roadmap

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 and firms, (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.

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 out-

Figure 3: Billed vs. Worked Hours, by Seniority



puts, 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 D presents a more general version with N levels. Workers live and work for two periods: they enter the labor market as juniors and 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 usually the case in 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 (seniority level $s = 1$) hired in period t becomes a senior ($s = 2$) in $t + 1$ and retires at 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 a per-period wage $W_{s(i,t),f,t}$ (with a random

coefficient α_f) and decreases in effort cost.[13,14] Thus, if firm f pays compensation $W_{s,f,t}$ to its level-s workers in period t, the flow utility of worker i is

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

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 level-$s(i,t)$ workers equally, whereas $\epsilon_{i,f,t}$ is idiosyncratic to worker i. For notational simplicity, we assume that $\xi_{s(i,t),f}$ does not vary over time within seniority levels in the presentation of the theoretical model, but we relax this in our empirical application.[15] Section 3.1.4 presents further distributional assumptions that we impose on $\xi_{s(i,t),f}$ and $\epsilon_{i,f,t}$.[16] We allow for random wage coefficients in the workers' flow utility, α_f.

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

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

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

3.1.2 *Production*

Each firm employs both juniors and seniors whose headcounts and average effort are denoted as $L_{s,f,t}$ and $\bar{e}_{s,f,t}$. The firm produces two outputs $Q_{s,f,t}$,

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

where $L_{-s,f,t}$ and $\bar{e}_{-s,f,t}$ denote the headcount and effort of the other seniority level. Effort is modeled as a productivity shifter and 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

[13] Given our random wage coefficient α_f, we allow for flexible variation in labor supply elasticities across firms. This parallels Volpe (2024), although we include firm-specific rather than worker-specific random coefficients.

[14] Whereas prior models of labor supply often include wages in logs, doing so heavily complicates our dynamic model. However, the inclusion of random coefficients mitigates some of this concern, as we allow higher-paying firms to have different wage coefficients than low-paying firms.

[15] When allowing for time-varying amenities $\xi_{s(i,t),f,t}$, these amenities become state variables. While this adds notation to the model, additional state-variables can be incorporated as shown in Appendix E.

[16] 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.

write firm f's output of level s in period t as $F_s(\mathcal{I}_{s,f,t})$.

That junior and senior employment enter the production function as different inputs may reflect systematic accumulation of human capital—*general* or *firm-specific*—over the course of workers' careers; see Section 3.6.5 for a discussion. In our empirical application, $Q_{s,f,t}$ corresponds to the total number of billing hours produced by workers of seniority level $s \in \{1, 2\}$.

3.1.3 *Compensation Policy*

At the time of hiring junior workers, firm f commits to a compensation–promotion policy, $\{W_{1,f,t}, W_{2,f,t+1}, \tau_{f,t}\}$, with per-period wages for juniors and seniors (i.e., promoted juniors) $W_{1,f,t}$ and $W_{2,f,t+1}$, and a promotion cutoff $\tau_{f,t}$, defined as the share of juniors dismissed. Junior workers for which $\theta_{i,f,t} \geq \theta^P_{f,t}(\tau_{f,t})$ 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.[17]

We assume that compensation of employees is set by the firm, rather than negotiated.[18] 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 3.6.4 and Appendix F, 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.

3.1.4 *Workers' Outside Options*

When making their employment decisions regarding firm f in period t, junior and senior workers 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.[19] Differences between the outside options drawn by the

[17] Although firms cannot commit to future wages in the legal sense, commitment is realistic in this setting given that equity partners are residual claimants of profits at most firms. Even when this is not the case, reputation concerns can induce firms to still uphold their wage commitments, as indicated by the empirical literature on relational contracts (Macchiavello and Morjaria, 2015; Brugues, 2020).

[18] 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 U.S. private-sector employment.

[19] 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.

same worker in two consecutive periods (first as a junior, $\ddot{U}^{alt}_{1,i,f,t}$, and then as a senior, $\ddot{U}^{alt}_{2,i,f,t+1}$) may reflect the accumulation of *general* human capital over the course of her career. Similarly, firm-specific outside option distributions can capture heterogeneity in the human capital accumulation opportunities offered by different employers (see Section 3.6.5).

To simplify the notation in the remainder of this section, for $s \in \{1, 2\}$, define $U^{alt}_{s,i,f,t} \equiv \ddot{U}^{alt}_{s,i,f,t} - \epsilon_{i,f,t}$. This is the outside option net of the idiosyncratic, non-pecuniary term. To determine whether worker i works at firm f, we can compare $U^{alt}_{s,i,f,t}$ to the firm- and seniority-specific terms of worker i's utility, as explained in Section 3.2.2. We assume that the draws of $U^{alt}_{s,i,f,t}$ are independently distributed across s, i, f, and t, and that, conditional on s and f, they are identically distributed across i and t—that is, individuals of the same seniority level considering working at the same firm draw their outside options from the same distribution, in every period. We relax the stationarity assumption in our empirical analysis in Section 4.[20] These assumptions still allow the distribution of $U^{alt}_{s,i,f,t}$ to vary arbitrarily with f and s, reflecting cross-firm heterogeneity in amenities and systematic differences between the outside options of junior and senior workers. Denote the common distribution of $U^{alt}_{s,i,f,t}$ across i and 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 $U^{alt}_{2,i,f,t+1}$ in advance. A worker hired by firm f in period t and subsequently promoted observes $U^{alt}_{2,i,f,t+1}$ at the beginning of period $t + 1$, before deciding whether to remain with the firm.

3.1.5 Product Market

Firms are monopolistic competitors in the product market. The demand for accounting services of level s is loglinear with elasticity $\eta < -1$. We assume that firms price each seniority level separately since the IPA data shows that two thirds of firms follow this pricing practice.[21] 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

$$Q_{s,f,t} = \overline{Q}_t \left(\frac{P_{s,f,t}}{\overline{P}_t} \right)^{\eta}. \tag{3}$$

[20] Allowing the distribution of $U^{alt}_{s,i,f,t}$ to vary over time would require us to account for an additional state variable; see footnote 15 and Appendix E for further discussion.

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

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

$$R_s(\mathcal{I}_{s,f,t}) \equiv P_s\left(F_s(\mathcal{I}_{s,f,t})\right) F_s(\mathcal{I}_{s,f,t}) = d_t\left[F_s(\mathcal{I}_{s,f,t})\right]^{\frac{1+\eta}{\eta}} \quad s \in \{1,2\}, \tag{4}$$

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 empirical exercise in Section 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 $e^*_{2,f,t}$.[22] Let $\bar{U}_{2,f,t}$ denote the senior workers' value of working at firm f in period t, net of the idiosyncratic term $\epsilon_{i,f,t}$. We thus have

$$\bar{U}_{2,f,t} = \alpha_f W_{2,f,t} - C(e^*_{2,f,t}) + \xi_{2,f}. \tag{5}$$

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

$$\begin{aligned}\bar{U}_{1,f,t} = \max_e\ & \alpha_f W_{1,f,t} - C(e) + \xi_{1,f} \\ & + \rho\left\{\left[1-\Phi\left(\theta^P_{f,t}(\tau_{f,t}) - e\right)\right]\bar{\bar{U}}_{2,f,t+1} + \Phi\left(\theta^P_{f,t}(\tau_{f,t}) - e\right)\bar{U}^{alt}_{2,f,t}\right\},\end{aligned} \tag{6}$$

where

$$\bar{\bar{U}}_{2,f,t+1} \equiv \mathbb{E}\left[\max\left\{\bar{U}_{2,f,t+1}, U^{alt}_{2,i,f,t+1}\right\}\right] = G^{alt}_{2,f}(\bar{U}_{2,f,t+1})\bar{U}_{2,f,t+1} + \int_{\bar{U}_{2,f,t+1}}^{\infty} u\, dG^{alt}_{2,f}(u),$$

and $\bar{U}^{alt}_{2,f,t+1}$ is the expected value of outside options for seniors at firm f in period $t+1$, $\mathbb{E}[U^{alt}_{2,i,f,t+1}]$. 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-

[22] 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 $\{W_{1,f,t}, W_{2,f,t}, \tau_{f,t}\}$.

pecuniary terms. A worker exerting effort e is promoted with probability $1-\Phi(\theta^P_{f,t}(\tau_{f,t})-e)$ and dismissed with the complementary probability. Accordingly, the term in braces in (6) represents the worker's *ex-ante* continuation value—before the realization of effort signals and the determination of promotion outcomes—as a probability-weighted average of $\bar{\bar{U}}_{2,f,t+1}$ and $\bar{U}^{alt}_{2,f,t+1}$.

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

$$C'(\tilde{e}_{f,t}) = \rho\phi(\kappa^P(\tau_{f,t}))[\bar{\bar{U}}_{2,f,t+1} - \bar{U}^{alt}_{2,f,t+1}], \tag{7}$$

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

The following lemma follows directly from (7) 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 $W_{2,f,t+1}$.*

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:

$$L_{1,f,t} = G^{alt}_{f,1}(\bar{U}_{1,f,t})\,\mathcal{L}. \tag{8}$$

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

$$L_{2,f,t+1} = (1-\tau_{f,t})G^{alt}_{2,f}(\bar{U}_{2,f,t+1})\,L_{1,f,t}. \tag{9}$$

3.2.3 *Firm's Problem*

Given incumbent seniors $L_{2,f,t}$, firm f chooses $\{W_{1,f,t}, W_{2,f,t+1}, \tau_{f,t}\}$ to maximize the discounted value:[23,24]

$$V(L_{2,f,t}) = \max_{W_{1,f,t}, W_{2,f,t+1}, \tau_{f,t}} \Big\{ R_1(L_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t}) + R_2(L_{2,f,t}, L_{1,f,t}, e^*_{2,f,t}, \tilde{e}_{f,t}) \\ - W_{1,f,t} L_{1,f,t} - \rho W_{2,f,t+1}(1-\tau_{f,t}) G^{alt}_{2,f}(\bar{U}_{2,f,t+1}) L_{1,f,t} + \rho V\left(L_{2,f,t+1}\right) \Big\}, \quad (10)$$

subject to the incentive-compatibility constraint (7), and with $L_{1,f,t}$ and $L_{2,f,t+1}$ given by (8) and (9), respectively. Using our product demand equation, 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 (10) gives the marginal value of seniors:

$$V'(L_{2,f,t}) = \frac{1+\eta}{\eta} d \Bigg\{ \left[F_1(L_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_{2,f,t}} \\ + \left[F_2(L_{2,f,t}, L_{1,f,t}, e^*_{2,f,t}, \tilde{e}_{f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_{2,f,t}} \Bigg\}. \quad (11)$$

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 headcounts $L_{2,f,1}$; and sequences of compensation–promotion policies, $\{W_{1,f,t}, W_{2,f,t+1}, \tau_{f,t}\}_{t \in \mathbb{N}^+}$; junior and senior employee headcounts $\{L_{1,f,t}, L_{2,f,t}\}_{t \in \mathbb{N}^+}$; and junior effort levels $\{\tilde{e}_{f,t}\}_{t \in \mathbb{N}^+}$ for each firm $f \in \{1, \ldots, F\}$, such that:

- Given $L_{2,f,t}$, the policy $\{W_{1,f,t}, W_{2,f,t+1}, \tau_{f,t}\}$ solves the firm's problem (10),
- Given $\{W_{1,f,t}, W_{2,f,t+1}, \tau_{f,t}\}$, the effort level $\tilde{e}_{f,t}$ solves the junior worker's incentive-

[23] 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).

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

compatibility constraint (7),

- Given $\{W_{1,f,t}, W_{2,f,t+1}, \tau_{f,t}\}$, the labor supply headcounts $\{L_{1,f,t}, L_{2,f,t+1}\}$ satisfy the transition rules (8) and (9),

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.[25] 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.

Before proceeding, we introduce notation that will be useful in deriving the testable implications. First, let the marginal revenue products of junior and senior labor in firm f and period t be

$$\begin{aligned}\text{MRPL}_{1,f,t} \equiv \frac{1+\eta}{\eta} d \Bigg\{ & \left[F_1(L_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_{1,f,t}} \\ & + \left[F_2(L_{2,f,t}, L_{1,f,t}, e^*_{2,f,t}, \tilde{e}_{f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_{1,f,t}} \Bigg\}, \end{aligned} \tag{12}$$

$$\begin{aligned}\text{MRPL}_{2,f,t} \equiv \frac{1+\eta}{\eta} d \Bigg\{ & \left[F_1(L_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_{2,f,t}} \\ & + \left[F_2(L_{2,f,t}, L_{1,f,t}, e^*_{2,f,t}, \tilde{e}_{f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_{2,f,t}} \Bigg\}. \end{aligned} \tag{13}$$

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

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

products for a new hire in period t:

$$\text{NPV}^{W}_{f,t} = W_{1,f,t} + \rho(1-\tau_{f,t})G^{alt}_{2,f}(\bar{U}_{2,f,t+1})\, W_{2,f,t+1}, \tag{14}$$

$$\text{NPV}^{\text{MRPL}}_{f,t} = \text{MRPL}_{1,f,t} + \rho(1-\tau_{f,t})G^{alt}_{2,f}(\bar{U}_{2,f,t+1})\, \text{MRPL}_{2,f,t+1}. \tag{15}$$

3.4.1 Test for Efficiency-Wage Backloading

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, $\partial F_1/\partial \bar{e}_{1,f,t} > 0$, to that in which junior effort is irrelevant for production, $\partial F_1/\partial \bar{e}_{1,f,t} = 0$. 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 $\tau_{f,t} \in (0,1)$ for firm f in period t. Then, the relationship between senior workers' compensation and their marginal revenue product of labor differs across effort regimes as follows:*

- ***Irrelevant Effort:*** $W_{2,f,t+1} = MRPL_{2,f,t+1}$.
- ***Productive Effort:*** $W_{2,f,t+1} > MRPL_{2,f,t+1}$.

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 $\tau_{f,t} \in (0,1)$ 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.

The proposition delivers a sharp result. In the Irrelevant Effort scenario, senior workers face no instantaneous wage markdown and therefore receive $W_{2,f,t+1} = MRPL_{2,f,t+1}$, 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.[26] 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.[27]

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

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

For intuition on the result, consider the firm's optimal choice of $W_{1,f,t}$ and $W_{2,f,t+1}$ (for the first-order conditions to the firm's problem, see Section A.1). In setting $W_{1,f,t}$, 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 $W_{2,f,t+1}$, 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, $W_{2,f,t+1}$ affects retention: a higher senior wage induces more senior workers to stay, generating marginal profits equal to $MRPL_{2,f,t+1} - W_{2,f,t+1}$.

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 $W_{1,f,t}$ and $W_{2,f,t+1}$. As a result, once $W_{1,f,t}$ 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 $W_{2,f,t+1} = MRPL_{2,f,t+1}$.

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 $W_{2,f,t+1} > MRPL_{2,f,t+1}$.

3.4.2 Test for Monopsony Power

Our second test concerns the firm's exercise of monopsony power. We begin by defining the current- and present-discounted-compensation elasticities of junior labor supply.

The *current-compensation* elasticity of residual labor supply by juniors for firm f in period t is

$$\psi_{1,f,t} \equiv \frac{\partial \ln L_{1,f,t}}{\partial \ln W_{1,f,t}} = \frac{g^{alt}_{1,f}(\bar{U}_{1,f,t})\, W_{1,f,t}}{G^{alt}_{1,f}(\bar{U}_{1,f,t})}\alpha_f. \tag{16}$$

Similarly, define the *present-discounted-compensation* elasticity of residual labor supply as

$$\psi_{PD,f,t} \equiv \frac{\partial \ln L_{1,f,t}}{\partial \ln \mathrm{NPV}^{W}_{f,t}}. \tag{17}$$

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

$$\psi_{PD,f,t} = \frac{\partial \ln L_{1,f,t}}{\partial \ln W_{1,f,t}} \frac{\partial \ln W_{1,f,t}}{\partial \ln \text{NPV}^W_{f,t}} = \psi_{1,f,t} \frac{\text{NPV}^W_{f,t}}{W_{1,f,t}}, \tag{18}$$

where the second equality uses $\partial \text{NPV}^W_{f,t} / \partial W_{1,f,t} = 1$ and the definition of $\psi_{1,f,t}$. Hence, the present-discounted-compensation elasticity $\psi_{PD,f,t}$ scales the current-compensation elasticity $\psi_{1,f,t}$ by (the inverse of) the share of the current wage in the present value of total expected earnings within the firm.

Proposition 2. *The firm's monopsony power in any period is characterized by the firm-specific present-discounted-compensation residual elasticity of labor supply. This elasticity determines the lifecycle wage markdown:*

$$\frac{NPV^W_{f,t}}{NPV^{MRPL}_{f,t}} = \frac{\psi_{PD,f,t}}{1 + \psi_{PD,f,t}}.$$

Appendix A.2.2 contains the proof. The result follows from the first-order condition for $W_{1,f,t}$ in the solution to the firm's problem. Proposition 2 shows that the firm faces an intertemporal wage–productivity wedge arising from the present-discounted-compensation elasticity of junior labor supply, $\psi_{PD,f,t}$. 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, e.g. Card et al. (2018). When residual labor supply becomes perfectly elastic with respect to present-discounted compensation, $\psi_{PD,f,t} \to \infty$, the firm pays the competitive lifetime wage, $\text{NPV}^W_{f,t} = \text{NPV}^{\text{MRPL}}_{f,t}$, and the model becomes an extension of Malcomson (1984) with nonlinear production. Under a finite labor supply elasticity $\psi_{PD,f,t}$, the ratio $\text{NPV}^W_{f,t} / \text{NPV}^{\text{MRPL}}_{f,t}$ captures the extent to which lifetime wages are marked down relative to the lifetime marginal revenue product of labor. Proposition 2 implies that the present-discounted-compensation residual labor supply elasticity for juniors, $\psi_{PD,f,t}$, can be recovered from $\text{NPV}^W_{f,t}$ and $\text{NPV}^{\text{MRPL}}_{f,t}$.

3.5 Effects of Eliminating Wage Markdown Variation

We use our model to characterize the effects of a policy that eliminates the junior wage markdown—the gap between the MRPL and wages for workers of level $s = 1$—on output

and prices across different environments. For this set of results, we focus on steady-state comparisons, abstracting from transitional dynamics. Accordingly, all model objects appear without time subscripts. The proofs of the propositions and the corollary below are in Appendix A.2.2.

Proposition 3 (Irrelevant Effort and Monopsony)**.** *Consider the steady-state equilibrium of the model, and assume that: (i) junior effort is irrelevant for production, so* $\partial F_s / \partial \bar{e}_{1,f} = 0$ *for all* $s \in \{1, 2\}$*; (ii) the firm has monopsony power, so* $\psi_{PD,f} < \infty$ *and* $NPV_f^W < NPV_f^{MRPL}$*; and (iii)* $\tau_f \in (0, 1)$*.*

Then, the equilibrium under a policy imposing $W_{1,f} = MRPL_{1,f}$ *features higher employment at both levels, higher output, and lower prices.*

Proposition 4 (Productive Effort and Perfect Competition)**.** *Consider the steady-state equilibrium of the model, and assume that: (i) junior effort is productive, so* $\partial F_s / \partial \bar{e}_{1,f} > 0$ *for at least one* $s \in \{1, 2\}$*; (ii) labor markets are perfectly competitive at both junior and senior margins, so* $NPV_f^W = NPV_f^{MRPL}$*; and (iii)* $\tau_f \in (0, 1)$*.*

Then, the equilibrium under a policy imposing $W_{1,f} = MRPL_{1,f}$ *features lower employment at both levels, lower output, and higher 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)**.** *Consider the steady-state equilibrium of the model, and assume that: (i) junior effort is productive, so* $\partial F_s / \partial \bar{e}_{1,f} > 0$ *for at least one* $s \in \{1, 2\}$*; (ii) the firm has monopsony power, so* $\psi_{PD,f} < \infty$ *and* $NPV_f^W < NPV_f^{MRPL}$*; and (iii)* $\tau_f \in (0, 1)$*.*

Then, the effects of imposing $W_{1,f} = MRPL_{1,f}$ *are, in general, ambiguous: the effects on employment, output, and prices can be either positive or negative.*

Real-life policies are unlikely to explicitly impose firms to pay workers their MRPL. However, the results above help to understand the potential implications of 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, minimum wages, 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 at firm f has a present-discounted value of income equal to

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

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

$$b^{in}_{f,t+1} = \pi^{s}_{f,t+1} + \rho\, b^{out}_{f,t+2}. \tag{19}$$

Therefore, the partner's present-discounted income simplifies to $W^{lump}_{f,t+1}$. 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 $W_{2,f,t}$ in the model.[28]

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 Levels

In Appendix D, 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

[28] 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 $W_{1,f,t} + W_{2,f,t} = MRPL_{1,f,t} + MRPL_{2,f,t}$. However, with discounting it becomes possible that $W_{1,f,t} + \rho W_{2,f,t} < MRPL_{1,f,t} + \rho MRPL_{2,f,t}$.

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 present-discounted-compensation 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 Propositions (1) and (2) to test for efficiency wages and monopsony power in our empirical application, which features more than two seniority levels.

3.6.3 Uncertainty and random productivity shocks

In Appendix E, we extend our baseline model to incorporate random productivity shocks into the firms' production functions. This extension brings the theoretical framework in line with our empirical analysis in Section 4, where our production function estimates rely on shocks of this type. Thus, the theoretical results that form the basis for our empirical tests—Propositions 1 and 2— equally apply in the presence of productivity shocks.

3.6.4 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 F, 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 F effectively enlarges the firm's contract space by allowing distinct compensation structures for different 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.

3.6.5 *Human Capital Accumulation*

Our framework flexibly accommodates human capital accumulation by workers. First, we impose no restriction on the extent to which $G^{alt}_{1,f}$ and $G^{alt}_{2,f}$, the distributions of outside offers for junior and senior workers, differ from each other. Systematic differences between these distributions may reflect changes in *general* human capital over workers' careers, which translate into distinct outside opportunities. That $G^{alt}_{1,f}$ and $G^{alt}_{2,f}$ also vary at the firm level allows for heterogeneity in the outside opportunities that different employers generate—an important dimension of firm quality that likely shapes junior workers' initial firm choice.

Second, junior and senior workers enter the production function as different inputs. Differences in their productivity may reflect the accumulation of human capital over time. Given that our model imposes no constraint on the joint distribution of outside offers and within-firm productivity of junior and senior workers, we remain agnostic about whether productivity differences across seniority levels arise from changes in general or *firm-specific* human capital.

Our empirical implementation in Section 4 retains this flexibility—remaining agnostic about the sources of human capital accumulation. While we estimate the production function—and thus productivity differences across seniority levels—we impose no restrictions on the distributions of outside offers $G^{alt}_{1,f}$ and $G^{alt}_{2,f}$. Our analysis is therefore robust to a broad class of human capital accumulation processes.

3.6.6 *Incentives vs. Sorting*

Although we allow for the services of different employee seniority levels to be differentiated, we impose that employees of a given seniority s are homogeneous within firms, up to the effort term $e_{i,t}$. If employees of the same seniority were persistently heterogeneous for reasons other than effort, promotion policies would serve not just as an incentive mechanism, but also as a sorting device that reflects employers' learning about workers' abilities. Indeed, there is a vast theoretical literature attributing promotions and compensation growth to information about employee productivity—whether innate or acquired through on-the-job human capital accumulation—that firms obtain over time (Harris and Holmstrom, 1982; Waldman, 1984; Prendergast, 1993; Gibbons and Waldman, 1999; Kahn and Lange, 2014). This type of mechanism has been used in particular to rationalize 'up-or-out' promotion policies common in public accounting firms, law partnerships, and many other professional services industries (Kahn and Huberman, 1988; Waldman, 1990). However, such models typically cannot generate strictly positive wage *markups* for senior employees—the key feature of tournament models that we exploit in our test for incentive

contracting. As shown in Section 4, we find strong evidence for these markups in our data. Thus, while we believe that firms' learning about worker ability plays some role in promotion decisions, we view the incentive-provision mechanism on which we focus as an essential component of career dynamics—at least in industries related to the one we study.

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 employees' careers while maintaining a minimal set of assumptions on labor supply.

4.1 Empirical Model

We implement empirical specifications for the model presented in Section 3. For expositional reasons, we suppress time notation throughout the empirical model.

In contrast to the two-level model in Section 3, we now let firms $f \in \mathcal{F}$ employ $|S| = 7$ 'billable' employee levels of seniorities $s \in \{1, ..., |S|\}$. We follow the level categorization provided in the dataset: professional staff with 0–2, 3–5, 6–8, and 9+ years of experience, paraprofessionals, and both non-equity and equity partners.

4.1.1 Product Demand

We implement the loglinear demand curve for level-s accountants from Equation (3) at the firm level with a common demand elasticity $\eta < -1$, adding an unobserved demand shifter $\zeta_{s,f}$:

$$Q_{s,f} = \overline{Q}_s \left(\frac{P_{s,f}}{\overline{P}_s} \right)^{\eta} \zeta_{s,f}. \tag{20}$$

Given the earlier-made assumption of monopolistic competition and CES demand, firms set billing rates at constant markups above accountant marginal costs, which are common across all seniority levels (Feenstra and Ma, 2007; Orr, 2022):

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

Rather than estimating η ourselves, we draw from Gerakos and Syverson (2015), who

estimate residual demand elasticities for auditing services. We use their average own-price elasticity for non-Big-4 auditing companies, $\eta = -1.827$ (implying a Lerner index of 55%), which we impose across all seniority levels.

4.1.2 Production of Accounting Services

Service production is characterized by Equation (21). Billable hours produced by each employee level s depend on labor of all seniority levels via a Cobb-Douglas production function. Since the dataset records both labor headcounts $L_{s,f}$ and hours worked $H_{s,f}$, we use their product as the labor input in order to correctly measure labor quantities. As we show in Appendix Figure OA-10, there is very little variation in recorded hours across firms, and, as mentioned earlier, employees are not on hourly contracts. It is thus possible that recorded hours, especially for non-partners, may not accurately reflect actual hours worked. To the extent that individual workers differ in their actual hours worked, this enters the latent effort variable $e_{i,f,t}$. Therefore, one can think of effort as a composite of worker productivity per recorded hour and latent hours worked beyond what is recorded.

In addition, production involves a vector of $|O|$ other inputs, denoted by $\boldsymbol{O}_f$, that do not directly generate billing revenue. These consist of interns, administrative staff, and three types of non-labor inputs: expenditure on training, recruiting, and technology.

$$Q_{s,f} = \left(\prod_{k \in S} (L_{k,f} H_{k,f})^{\beta^l_{s,k}} \right) \boldsymbol{O}_f{}^{\boldsymbol{\beta}^o_s} \Omega_{s,f}(\bar{e}_{s,f}) \exp(v_{s,f}). \tag{21}$$

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 $|O| \times |S|$.

Seniority-specific productivity levels are denoted by $\Omega_{s,f}$. In our empirical specification, we let productivity of level-s workers be a function only of the average effort of workers at the same seniority level.[29] Finally, measurement error in output is denoted by $\exp(v_{s,f})$.

Unfortunately, we do not observe fixed assets, such as office space rent or insurance expenses. In a robustness analysis in Appendix C.1.1, 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.

[29] Allowing for cross-seniority effort spillovers would require including output levels of all other seniority levels in the production function, substantially increasing the number of parameters to be estimated.

Discussion. Our model features multiple outputs and team production. This formulation offers at least two advantages. First, it allows us to distinguish output across worker seniority levels, whereas a standard firm-level production model would require aggregating billing hours across seniority levels.[30] Second, our model allows labor inputs to be non-rival: we do not require mapping each service to specific inputs. For instance, each unit of partner labor and effort can increase the billing revenue of more junior accountants through business development—acting as a public good within the firm.

Caveats. There are two important caveats to our production model. First, by letting labor worked by other employees—rather than their output—enter the production function of each seniority level, we assume away economies of scope and productivity spillovers across goods, in contrast to Dhyne et al. (2022).[31] 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 sample size. However, we do allow level-s productivity to be correlated arbitrarily across seniority levels.

Second, although we allow output to be differentiated across seniority levels 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 C.1.3, 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 level-s labor, $\text{MRPL}_{s,f} \equiv \partial R_f / \partial L_{s,f}$, can be expressed as:

$$\text{MRPL}_{s,f} = \frac{1}{L_{s,f}} \sum_k \beta^l_{k,s} R_{k,f} \left(\frac{1+\eta}{\eta} \right), \tag{22}$$

with total firm revenue being denoted as $R_{f,t} = \sum_s R_{s,f,t}$.

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

Using the MRPL expression from Equation (22), we write the 'instantaneous mark-

[30] We compare our estimates with a more conventional firm-level revenue production function that aggregates output across employees in Appendix C.2.

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

down' $\delta_{s,f}$ at every seniority level as:

$$\delta_{s,f} = \frac{W_{s,f} L_{s,f}}{\sum_k \beta^l_{k,s} R_{k,f} \left(\frac{\eta+1}{\eta}\right)}. \tag{23}$$

Hence, to recover these instantaneous markdowns, we use the observed seniority-specific annual billing revenues $R_{s,f}$ and annual earnings $W_{s,f}L_{s,f}$, the estimated output elasticity matrix $\boldsymbol{\beta}^l$, and the calibrated product demand elasticity η. Given the instantaneous markdown estimates and observed wages, we can directly recover the implied MRPL.

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 seniority level:

$$q_{s,f} = \sum_{k \in S} \left(\beta^l_{s,k}(l_{k,f} + h_{k,f})\right) + \beta^o_s \boldsymbol{o}_{s,f} + \omega_{s,f} + v_{s,f}. \tag{24}$$

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 $v_{s,f}$, with seniority-specific serial correlation σ_s and a constant c_s.[32]

$$\omega_{s,f} + v_{s,f} = c_s + \sigma_s(\hat{\omega}_{s,f} + \hat{v}_{s,f}) + \nu_{s,f}. \tag{25}$$

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

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

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

[32] In our main specification, we impose the AR(1) assumption on the entire productivity residual, which includes effort. In Appendix C.1.2, we instead separate out effort-related productivity from non-effort productivity, and only impose the AR(1) assumption on the latter. In Appendix C.1.4, we impose an AR(2) assumption instead, given that effort might not be AR(1) if based on persistent contract terms. Both of these alternative specifications lead to very similar results.

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

We assume that revenue-generating labor inputs of any seniority level s are dynamic inputs, because hiring these staff members likely induces adjustment costs, and because firms commit to promotion rates ex-ante in our model.[33] This means that firms chooses the amounts of these seniority levels to hire before observing the transient productivity shocks $\nu_{s,f}$. We also include the non-labor inputs (technology, training, and recruiting costs) into the predetermined vector, as these look mostly like investments in the firm's intangible capital stock. In contrast, we assume that firms can flexibly adjust both their interns and administrative staff, meaning that these inputs are flexibly chosen after productivity shocks are observed by the firms.

Consistent with these timing assumptions, we impose the following moment conditions to estimate the parameters $(\sigma_s, c_s, \boldsymbol{\beta}_s)$:

$$\begin{aligned} E(\nu_{s,f}|\boldsymbol{X}_f^{\text{pre}}, \hat{\boldsymbol{X}}_f^{\text{pre}}) &= 0, \\ E(\nu_{s,f}|\hat{\boldsymbol{X}}_f^{\text{flex}}) &= 0. \end{aligned} \tag{26}$$

Intuitively, identification comes from variation in predetermined inputs across firms and over time. Conditional on lagged productivity, firms that differ in staffing levels by seniority tier, technology expenditures, training expenditures, and recruiting investments should exhibit different levels of billable hours according to the corresponding output elasticities. Because these inputs are chosen before the realization of the innovation $\nu_{s,f}$, the moment conditions exploit the orthogonality between contemporaneous productivity shocks and predetermined inputs to identify the coefficients $\boldsymbol{\beta}_s$. Commitment to promotion rates is helpful for identification: given that seniors are mainly internally promoted, current productivity shocks only affect senior employment after many years, but rapidly transmit to junior hiring. This variation in productivity pass-through timing helps identify, for instance, the elasticity of senior billing hours with respect to junior employment.

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 level s. Similarly, labor is measured as the product of employment headcounts $L_{s,f}$ and annual hours worked per employee $H_{s,f}$. Because we observe neither worked nor billed hours for non-equity partners, we assume

[33] When treating junior accountants with 0-2 years of experience as a flexible input instead, very similar markdown estimates are obtained.

their working and billing hours per worker are the same as those of equity partners. Thus, variation in the contribution of these two worker seniority levels stems only from their headcounts. Relatedly, administrative staff and interns are measured in employee counts, as we do not observe hours for these employees. Recurring training, recruiting, and technology expenses are reported in the data as shares of revenue. We compute the dollar value of these non-labor expenses by multiplying these shares 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 (24) at the bin-year level under two additional assumptions. First, we require that input prices, residual input supply elasticities, and output elasticities are identical across firms within each bin-year. This is a major reason why we need to impose the Cobb-Douglas specification, rather than using for more flexible functional forms for the production functions. Second, the firm-level timing assumptions need to extend to the bin level, which requires the additional assumption that exit and entry are orthogonal to the transient productivity shocks $v_{s,f}$.

We estimate the production functions by specifying the moment conditions (26) at the bin level. For the flexible inputs, we include a single lag in the instruments vector. For the predetermined inputs, we include current values and a single lag in the instruments vector. We 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 all employee seniority levels. The off-diagonal elements measure spillovers in production. The lower-triangular elements are the effects of senior hours on junior output. Row six, referring to the equity partners, stands out. This row has mostly positive and significant coefficients, meaning that hours worked by equity partners increase billing revenues of all other employees. This highlights the crucial role of equity partners in providing general management and mentorship, and in business development. These spillovers from equity partners are quantitatively important: as shown in Appendix C.3, ignoring them would lead to substantial underestimation of the equity partners' marginal revenue products.

The upper-triangular elements measure externalities of junior employees on more senior colleagues. Here, the picture is more mixed, with some significant positive and some significant negative coefficients. For instance, employees with 6-8 years of experience seem to impose a significant negative externality on their immediate superiors (9+ years),

possibly reflecting the cost of training or supervising more junior employees.

Finally, paraprofessionals seem to impose negative spillovers on all other employee seniority levels, which could be due to task duplication or congestion effects. Interns, administrative staff, and non-labor inputs 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	1.000	0.024	0.020	0.015	0.027	0.022	-0.011	0.022	0.008	0.031	0.008	0.031	0.029	0.064
3-5 years	0.040	0.019	1.026	0.016	-0.014	0.022	0.016	0.021	0.025	0.031	0.025	0.031	-0.016	0.095
6-8 years	-0.002	0.020	-0.012	0.015	0.997	0.020	-0.050	0.019	0.022	0.034	0.022	0.034	0.066	0.077
9 years	-0.010	0.022	-0.018	0.014	-0.014	0.026	0.962	0.028	-0.012	0.035	-0.012	0.035	-0.033	0.052
Non-eq. Partners	0.021	0.011	0.009	0.007	0.007	0.012	-0.011	0.013	0.974	0.015	-0.026	0.015	0.005	0.039
Eq. Partners	-0.022	0.023	0.037	0.017	0.076	0.026	0.095	0.029	0.051	0.050	1.051	0.050	0.118	0.081
Paraprof.	-0.039	0.011	-0.030	0.007	-0.020	0.012	-0.032	0.010	-0.006	0.016	-0.006	0.016	1.010	0.039
Interns	0.024	0.016	0.016	0.011	-0.027	0.018	-0.005	0.020	-0.023	0.020	-0.023	0.020	-0.147	0.066
Training	-0.011	0.009	0.003	0.005	0.003	0.008	-0.004	0.008	0.014	0.011	0.014	0.011	-0.011	0.025
Recruiting	0.006	0.004	0.006	0.002	0.007	0.004	0.007	0.003	-0.002	0.006	-0.002	0.006	0.032	0.010
Technology	-0.002	0.009	-0.003	0.006	0.002	0.011	0.021	0.012	-0.025	0.012	-0.025	0.012	-0.025	0.029
Admin. staff	-0.001	0.016	-0.035	0.009	-0.030	0.018	0.017	0.015	-0.019	0.019	-0.019	0.019	-0.027	0.106
Ser. corr.	0.944	0.056	0.937	0.044	0.885	0.053	0.912	0.041	0.934	0.040	0.934	0.040	0.437	0.156
Obs.	174		174		174		174		174		174		174	

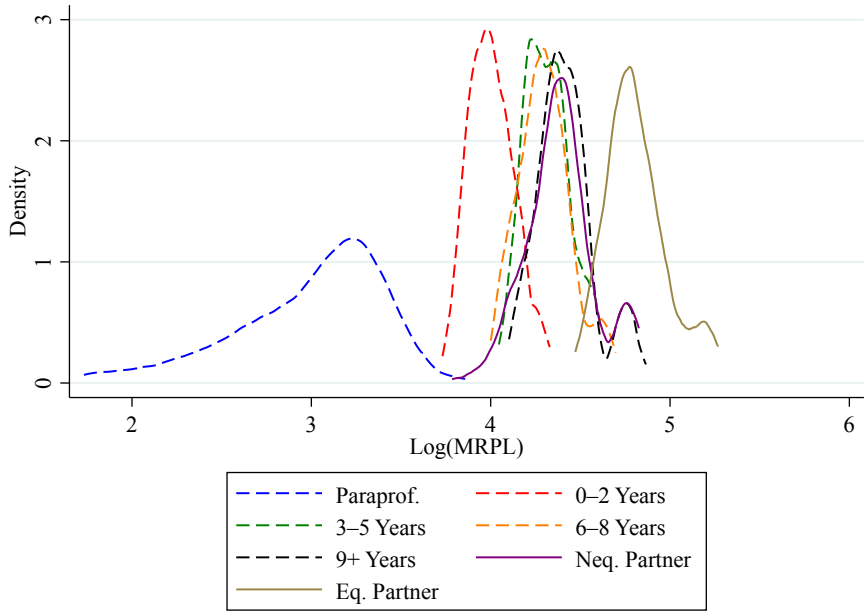
Notes: This table shows the estimated production functions for the seven billing revenue-generating employee titles in the dataset (the columns). The own- and cross-output elasticities for all inputs are reported in the rows. The number of observations shrinks to 174 because interns are only observed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

4.2.4 Results: Marginal Revenue Products of Labor

We recover instantaneous wage markdowns from Equation (23), using the estimated output elasticities together with observed billing rates and compensation.

Figure 4 plots the distribution of log MRPL across seniority levels. We find evidence of an 'MRPL ladder': paraprofessionals have much lower MRPL compared with the professional staff. Professional staff members with 0-2 years of experience have a substantially lower MRPL compared to more experienced employees. Staff with 3-5 and 6-8 years of experience have similar MRPL distributions, which is below the MRPL distributions for 9+ years of experience and non-equity partners. Equity partners have by far the highest MRPL, and some of that is because of their team production spillovers on the other employees, as shown in Appendix C.3.

Figure 4: MRPL Distributions



Notes: This figure shows the distributions of log(MRPL) for paraprofessionals and professional staff by seniority level, across firm-bins and years. All professional staff MRPL distributions are trimmed at the 1st and 99th percentiles. The paraprofessional MRPL distribution is trimmed at the 10th percentile because of a long left tail.

4.3 Testing for Incentive Contracts and Monopsony Power

4.3.1 Empirical Tests

We use the theoretical results from Section 3 to empirically test for the usage of backloaded incentive contracts and for the exertion of monopsony power.

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

$$\begin{cases} H_0 : W_{s^*,f} = MRPL_{s^*,f}, \\ H_1 : W_{s^*,f} > MRPL_{s^*,f} \end{cases} \quad \text{for } s^* > 1. \tag{27}$$

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

In the baseline theoretical model, we had two seniority levels, meaning that we had to test whether wages surpass the MRPL of level-2 workers, $s^* = 2$. When interpreting the results through the lens of the model with $N > 2$ seniority levels, the complication arises that the seniority levels measured in the dataset are not necessarily the seniority levels actually considered by firms in their promotion schemes—which would correspond

to the levels in the model. We therefore simply test whether wages surpass the marginal revenue product for any seniority level $s > 1$, noting that backloaded incentive contracts imply above-MRPL compensation from some seniority level onward. In Appendix B.3, we show more formally how our N-level model results can be used to map observed seniority levels into model-implied ones.

Second, following Proposition 2, we test empirically for monopsony power. Under the null hypothesis of no monopsony power, we consider:

$$\begin{cases} H_0 : \text{NPV}_f^{\text{W}} = \text{NPV}_f^{\text{MRPL}}, \\ H_1 : \text{NPV}_f^{\text{W}} < \text{NPV}_f^{\text{MRPL}}. \end{cases} \tag{28}$$

Rejecting H_0 indicates that firms exert monopsony power in the design of their wage schemes.

4.3.2 Backloaded Compensation Contracts

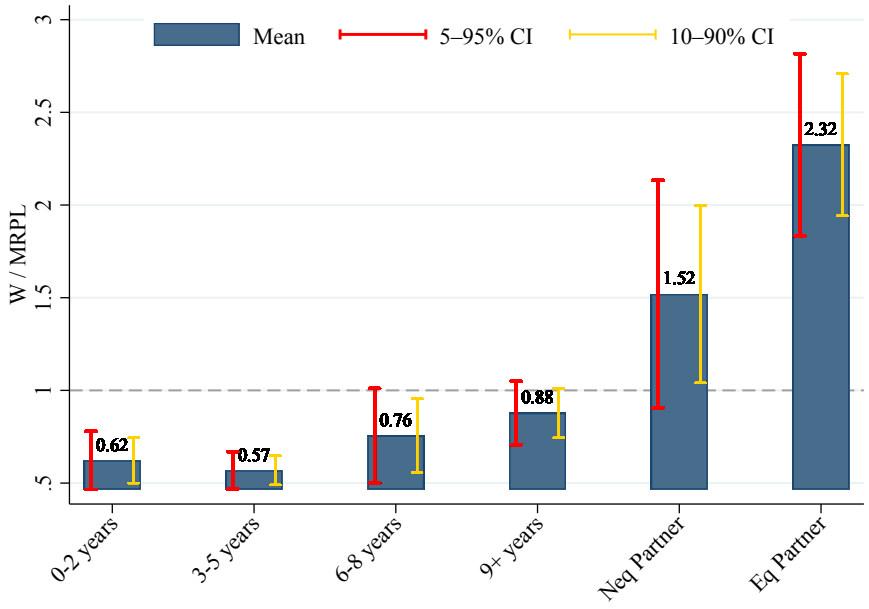
The estimated average instantaneous wage markdowns/markups for professional staff are summarized in Figure 5. They are significantly different from one at conventional levels for all worker seniority types. Employees with less than 5 years of experience are paid only 57-62% of their MRPL, which increases to 76% after 6 years of experience and 88% after 9 years. Wages are significantly marked down for all professional staff levels.

As soon as accountants are promoted to non-equity partners—who, in practice, still count as employees rather than firm owners—they begin earning wage markups, with wages being 52% *above* their marginal revenue product. Equity partners, in turn, earn 132% more than their marginal revenue product.

Both wage markups are statistically significant. Given our Proposition 1, we therefore reject the null hypothesis of irrelevant effort: the 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.

A possible concern regarding our equity partner results is that the wage markups used in our test are conceptually based on lump-sum payments, rather than total equity partner compensation (see Section 3.6.1). The compensation reported in our data, however, includes profit payouts as well, since equity partners are residual claimants of profits. In Appendix B.5, we address this by re-estimating equity partner wage markups after correcting their compensation for profit payouts, using observed partner buy-in data. This analysis delivers wage markups for equity partners that are only marginally below those reported in our main analysis. Note that no adjustment is necessary for the compensation

Figure 5: Instantaneous Wage Markdowns and Seniority



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

of non-equity partners, since they are not residual claimants of the firm.

Although our model in principle allows us to compute markdowns also for other seniority levels, such as paraprofessionals, doing so results in very noisy estimates, given that the MRPL estimates are much more dispersed for these employees compared to professional staff. We discuss these estimates in Appendix B.4. Since the lifecycle markdowns and contract-level labor supply elasticities only require estimates of the professional staff and partner markdowns, we focus on these estimates in the main text.

4.3.3 Monopsony Power

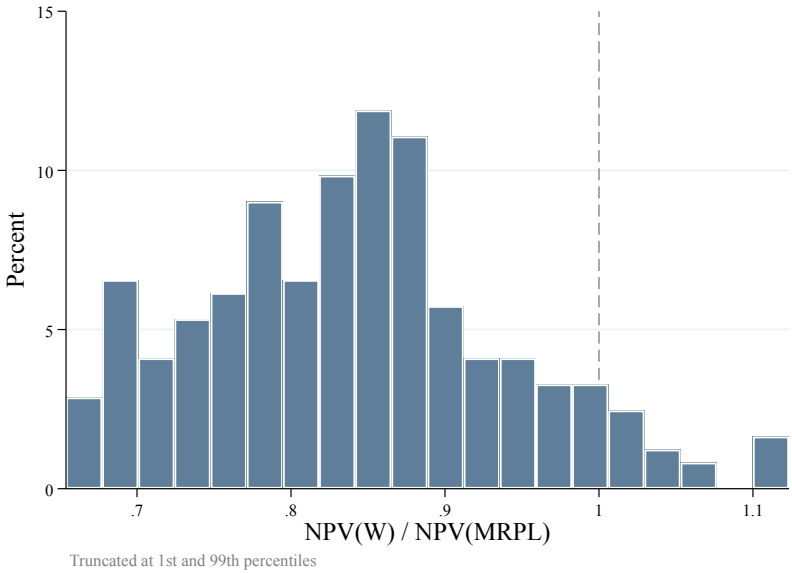
We compute the net present values (NPVs) of both wages and marginal revenue products from the start of a worker's career until retirement at T. Let $\lambda_{f,t}$ denote the ex-ante probability of still being employed at the firm in year t. We define the NPV of wages as

$$\text{NPV}_{W_f} = \sum_{t=1}^{T} \rho^{t-1} \lambda_{f,t} W_{ft}, \tag{29}$$

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

We estimate the survival probabilities $\lambda_{f,t}$ using worker stocks and assuming a steady-state labor market, but allow for $\lambda_{f,t}$ to differ by firm, as explained in Appendix B.2. We

Figure 6: Distribution of Expected Wage Markdowns



Notes: Distribution of the markdown of the NPV of earnings over the NPV of the marginal revenue product of labor, across firm-bins and years.

impose an annual discount factor $\rho = 0.95$, which corresponds to a real interest rate of around 5%.[34]

Figure 6 presents the distribution of the resulting estimated NPV markdowns across firm bins and years. Most of the distribution lies below one, indicating that the NPV of wages falls short of the NPV of the marginal revenue product of labor. The average markdown is 0.85, implying that, in expectation, the NPV of the wage is 15% below the NPV of the MRPL, with a 5–95% confidence interval of [0.76, 0.93]. In accordance with Proposition 2, these 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 reflect monopsony power.

4.4 Labor Supply Elasticities

We turn our estimated expected wage markdowns into present-discounted-compensation labor supply elasticities using Proposition 2, and compute current-compensation labor supply elasticities from present-discounted-compensation elasticities, current wages, and the NPV of wages using Equation (18). Key moments from both elasticity distributions across firm bins are reported in Table 3. Labor supply is very inelastic at the current-compensation level—with both the average and median elasticity across firm bins being

[34] In Appendix C.4, we examine robustness of the results to different discount factors.

0.4. 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).

At the present-discounted-compensation level, however, labor supply is more elastic, with an elasticity of 5.4 on average and 4.6 at the median. These are comparable in magnitude to the elasticities reported in meta-analyses in Card (2022) and Azar and Marinescu (2024), and of a similar magnitude as recent elasticity estimates that allow for random coefficients in labor supply (Volpe, 2024).

Table 3: Present-Discounted- vs. Current-Compensation Labor Supply Elasticities

	Elasticities of Labor Supply			
	Current		Present Disc.	
	Est.	S.E.	Est.	S.E.
Average	0.413	0.141	5.355	2.241
1st Quartile	0.271	0.137	2.790	2.539
Median	0.398	0.121	4.563	1.453
3rd Quartile	0.586	0.213	7.399	2.624

Notes: Selected moments of the present-discounted- and current-compensation 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 Labor Supply Elasticities

There is substantial heterogeneity in the present-discounted-compensation labor supply elasticity across firm bins and years. 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 and more elastic at medium-sized firms, compared to large firms. This finding rejects oligopsonistic conduct, under which markdowns should monotonically increase with size. Under our assumed model of conduct, monopsonistic competition, we assume instead that firms are all atomistic but differentiated. Therefore, the size effects suggest that very small firms are the most horizontally differentiated as perceived by workers, followed by very large firms—with mid-sized firms being the least differentiated. In terms of the time-series variation of labor supply elasticity, the present-discounted-compensation elasticity

appears to have increased substantially starting in 2022, reaching levels more than 60% above those observed in 2015.

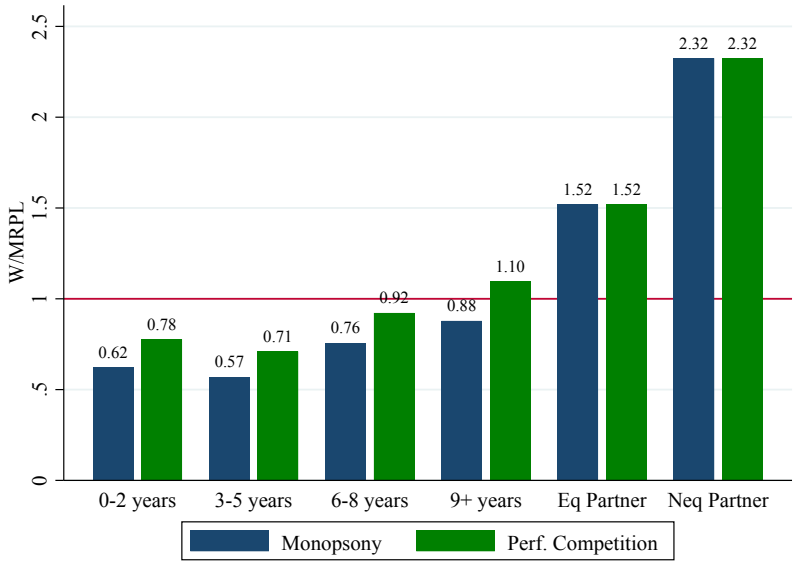
In column (II), we add intermediate input expenditures and the share of CPA-certified personnel to the right-hand side, with the caveat that these are endogenous variables. None of these yield significant coefficients.

Table 4: Covariates of Present-Discounted-Compensation Supply Elasticities

	(I)		(II)	
	log(Labor Supply Elast.)			
	Est.	S.E.	Est.	S.E.
Medium-size firm	0.373	0.126	0.396	0.148
Small-size firm	-0.495	0.093	-0.438	0.124
Year = 2016	0.007	0.124	0.021	0.120
Year = 2017	0.027	0.134	0.065	0.132
Year = 2018	0.148	0.153	0.192	0.153
Year = 2019	0.108	0.158	0.164	0.163
Year = 2020	0.212	0.180	0.312	0.186
Year = 2021	0.026	0.148	0.204	0.188
Year = 2022	0.357	0.181	0.539	0.210
Year = 2023	0.673	0.242	0.900	0.326
Year = 2024	0.614	0.183	0.871	0.238
log(Marketing exp.)			0.316	0.210
log(Training exp.)			0.082	0.168
log(Recruiting exp.)			-0.020	0.092
log(Technology exp.)			-0.049	0.243
CPA share			0.920	0.651
R-squared	.306		.320	
Observations	225		225	

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. Level of observation is firm-bin by year.

Figure 7: Markdown Decomposition



4.5 Markdown Decomposition

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 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. Let T^{sen} denote the time of promotion to partnership.[35] Then, χ_f satisfies:

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

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 26% to close the gap. Isolating the backloading component substantially compresses the markdowns: rather than ranging from 0.62 (rookies) to 0.88 (associates with 9+ years of tenure), they would range from 0.78 to 1.10 , as Figure 7 shows.

This allows us to decompose junior markdowns into their monopsony and incentive contract components. For instance, 43% of the markdown for workers with 0–2 years of

[35] We define the 'senior' rank as non-equity and equity partners (Appendix B.3), with promotion assumed to occur at 15 years of seniority (Appendix B.2.2)

experience is attributable to monopsony power, with the remaining 57% being due to wage backloading.

One important caveat is that by holding senior wages and promotion probabilities fixed, we do not allow for these endogenous variables to reoptimize when firms hold different degrees of market power over their junior employees. We therefore view this decomposition as a back-of-the-envelope calculation rather than a full counterfactual exercise, which would require solving 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 backloaded compensation contracts to incentivize worker effort. Using a tournament model with multi-output team production and monopsonistic competition, we show that wage markdowns can originate both from monopsony power and from dynamic implicit contracts—two mechanisms with different welfare implications—and derive empirical tests for each. First, we find that a sufficient test for backloaded incentive pay is that the wage of senior employees exceeds their marginal revenue product of labor. Second, we characterize the degree of monopsony power 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 present-discounted-compensation residual labor supply elasticity. Our estimates for the U.S. public accounting industry indicate that residual labor supply is very inelastic with respect to current compensation, and more elastic—with an elasticity around 5.4—with respect to the present-discounted compensation. 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 reflect 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 account for both their effects on worker effort and their implications for 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.

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*.

Bowles, S. (1985). The production process in a competitive economy: Walrasian, neo-Hobbesian, and Marxian models. *The American economic review 75*(1), 16–36.

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.

Brugues, F. (2020). Take the goods and run: Contracting frictions and market power in supply chains. *Work. Pap., Brown Univ., Providence, RI*.

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.

Cardoso, A. R., P. Guimarães, and J. Varejão (2011). Are older workers worthy of their pay? An empirical investigation of age-productivity and age-wage nexuses. *De Economist 159*(2), 95–111.

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.

Dalton, D. W., E. J. Mertens, and T. J. Rupert (2022). Career Paths and Compensation for Accounting Graduates. *Accounting Horizons 36*(1), 77–98.

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 (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.

Dohmen, T. J. (2004). Performance, seniority, and wages: formal salary systems and individual earnings profiles. *Labour Economics 11*(6), 741–763.

Emanuel, N. and E. Harrington (2026). The Payoffs of Higher Pay: Labor Supply and Productivity Responses to a Voluntary Firm Minimum Wage. *FRB of New York Staff Report* (1182).

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

Flabbi, L. and A. Ichino (2001). Productivity, seniority and wages: new evidence from personnel data. *Labour Economics 8*(3), 359–387.

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

Gibbons, R. and M. Waldman (1999). A theory of wage and promotion dynamics inside firms. *The Quarterly Journal of Economics 114*(4), 1321–1358.

Goolsbee, A. and C. Syverson (2023). Monopsony power in higher education: A tale of two tracks. *Journal of Labor Economics 41*(S1), S257–S290.

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

Harris, A. and T. M. A. Nguyen (2025). Long-term relationships in the us truckload freight industry. *American Economic Journal: Microeconomics 17*(1), 308–353.

Harris, M. and B. Holmstrom (1982). A theory of wage dynamics. *The Review of Economic Studies 49*(3), 315–333.

Hellerstein, J. K. and D. Neumark (1995). Are earnings profiles steeper than productivity profiles? Evidence from Israeli firm-level data. *Journal of human resources*, 89–112.

Hellerstein, J. K. and D. Neumark (2007). Production function and wage equation estimation with heterogeneous labor: Evidence from a new matched employer-employee data set. In *Hard-to-measure goods and services: Essays in honor of Zvi Griliches*, pp. 31–71. University of Chicago Press.

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 Human Resources Report.

Inside Public Accounting (2023c). 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.

Kahn, C. and G. Huberman (1988). Two-sided uncertainty and" up-or-out" contracts. *Journal of Labor Economics 6*(4), 423–444.

Kahn, L. B. and F. Lange (2014). Employer learning, productivity, and the earnings distribution: Evidence from performance measures. *The Review of Economic Studies 81*(4), 1575–1613.

Kissin, E. (2026, April). KPMG and EY Demote Partners in End of Job-for-Life Model. *Australian Financial Review*. Accessed: 2026-04-28.

Kline, P. M. (2025). Labor market monopsony: Fundamentals and frontiers. *National Bureau of Economic Research*.

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.

Macchiavello, R. and A. Morjaria (2015). The value of relationships: evidence from a supply shock to Kenyan rose exports. *American Economic Review 105*(9), 2911–2945.

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.

Manning, A. (2021). Monopsony in Labor Markets: A Review. *ILR Review 74*(1), 3–26.

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

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

Medoff, J. L. and K. G. Abraham (1981). Are those paid more really more productive? The case of experience. *Journal of Human resources*, 186–216.

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.

Prendergast, C. (1993). The role of promotion in inducing specific human capital acquisition. *The Quarterly Journal of Economics 108*(2), 523–534.

Rosenberg Associates (2013, June). CPA Firms' Mad Dash for Non-Equity Partners. *Rosenberg Associates*. Accessed: 2026-05-18.

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

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.

Syverson, C. (2025). Markups and markdowns. *Annual review of economics 17*(1), 57–76.

The CPA Journal (2017, December). ICYMI: The Overtime Pay Issue in Public Accounting. *The CPA Journal*. Accessed: 2026-05-18.

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

U.S. Bureau of Labor Statistics (2026). Professional, Scientific, and Technical Services: NAICS 54. Accessed: 2026-05-27.

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

Verboven, F. and B. Yontcheva (2024). Private monopoly and restricted entry—evidence from the notary profession. *Journal of Political Economy 132*(11), 3658–3707.

Volpe, O. (2024). Job Preferences, Labor Market Power, and Inequality. Technical report, Discussion paper, Working Paper.

Waldman, M. (1984). Job assignments, signalling, and efficiency. *The Rand journal of economics 15*(2), 255–267.

Waldman, M. (1990, April). Up-or-Out Contracts: A Signaling Perspective. *Journal of Labor Economics 8*(2), 230–250.

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

Monopsony and Backloaded Compensation: Theory and Evidence from Public Accountants

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: Additional Material . . . . . **OA - 8**
B.1 Aggregation to the Bin-Level OA - 8
B.2 Computing Promotion Probabilities OA - 10
B.3 Observed vs. Model-implied Seniority Levels OA - 11
B.4 Paraprofessional Wage Markdown OA - 12
B.5 Equity Partner Wage Markup Using Lump-Sum Payments OA - 12

C Robustness Checks . . . . . **OA - 14**
C.1 Alternative Production Function Specifications OA - 14
C.2 Single-Product Production Function OA - 17
C.3 Marginal Revenue Products: Comparison of Production Models OA - 19
C.4 Alternative Discount Factors OA - 19
C.5 Bundle-Pricing vs. Line-Pricing OA - 20

D Extending the Model to N Employee Levels . . . . . **OA - 24**
D.1 Environment and Timing OA - 24
D.2 Production and Prices OA - 24
D.3 Compensation–Promotion Policy OA - 25
D.4 Outside Options OA - 25
D.5 Workers' Values and Incentives OA - 25
D.6 Employment Dynamics OA - 26
D.7 Firm's Problem OA - 26

D.8 Steady State OA - 27
D.9 Incentive Contract Implications OA - 28
D.10 Entry Elasticity and the Exact Intertemporal Wedge OA - 32

E Incorporating Random Productivity Shocks into the Model . . . . . OA - 34

F Extending the Model to Allow for Lateral Hiring . . . . . OA - 37
F.1 Primitives OA - 37
F.2 Worker and Firm Behavior OA - 37
F.3 Testable Implications OA - 38

G Additional Figures . . . . . OA - 39

Online Appendix

A Proofs and Derivations

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

We start from the recursive value function (10). We derive the first-order conditions associated with $W_{1,f,t}$, $W_{2,f,t+1}$, and $\tau_{f,t}$ in steady state. To simplify notation and improve readability, we define

$$B_{1,f,t} = G^{\text{alt}}_{1,f}(\bar{U}_{1,f,t})\mathcal{L}, \qquad B_{2,f,t+1} = G^{\text{alt}}_{2,f}(\bar{U}_{2,f,t+1}).$$

Using these definitions, the first-order conditions for $W_{1,f,t}$, $W_{2,f,t+1}$, and $\tau_{f,t}$ can be written as:

$$W_{1,f,t}: \quad \frac{\partial \bar{U}_{1,f,t}}{\partial W_{1,f,t}} g^{alt}_{1,f}(\bar{U}_{1,f,t})\mathcal{L}\Bigg[MRPL_{1,f,t} - W_{1,f,t} - \rho W_{2,f,t+1}(1-\tau_{f,t})B_{2,f,t+1} + \rho(1-\tau_{f,t})B_{2,f,t+1}V'\Big((1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t}\Big)\Bigg] - B_{1,f,t} = 0, \quad \text{(OA.1)}$$

In (OA.56), the expression within square brackets captures how the firm's value changes as the number of juniors increases, in response to a raise in $W_{1,f,t}$. The term $B_{1,f,t}$ outside the brackets represents the marginal increase in the firm's wage bill resulting from the higher junior compensation.

$$\begin{aligned} W_{2,f,t}: \quad & \frac{\partial \bar{U}_{1,f,t}}{\partial W_{2,f,t+1}} g^{alt}_{1,f}(\bar{U}_{1,f,t})\mathcal{L}\Bigg[MRPL_{1,f,t} - W_{1,f,t} - \rho W_{2,f,t+1}(1-\tau_{f,t})B_{2,f,t+1} \\ & + \rho(1-\tau_{f,t})B_{2,f,t+1}\, V'\big((1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t}\big)\Bigg] \\ & - \rho(1-\tau_{f,t})B_{1,f,t}B_{2,f,t+1} + \rho(1-\tau_{f,t})B_{1,f,t}\frac{\partial \bar{U}_{2,f,t+1}}{\partial W_{2,f,t}} g^{alt}_{2,f}(\bar{U}_{2,f,t+1})\Bigg[- W_{2,f,t+1} \\ & + V'\big((1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t}\big)\Bigg] \\ & + \frac{1+\eta}{\eta}\frac{\partial \tilde{e}_{f,t}}{\partial W_{2,f,t+1}}\Bigg\{ d\left[F_1(B_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial \tilde{e}_{f,t}} \\ & + d\left[F_2(L_{2,f,t}, B_{1,f,t}, e^*_{2,f,t}, \tilde{e}_{f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial \tilde{e}_{f,t}}\Bigg\} = 0, \end{aligned}$$

As in the previous case, the term inside the square brackets in the first and second lines reflects the change in the number of juniors attracted by a higher $W_{2,f,t+1}$. In the third line, the first term

represents the marginal (discounted) increase in the wage bill due to the higher $W_{2,f,t+1}$, while the second term (continuing into the fourth line) captures the additional number of senior employees the firm retains. Finally, the term in the fifth and sixth lines corresponds to the greater effort exerted by junior employees in response to the change in $W_{2,f,t+1}$.

$$\begin{aligned}\tau_{f,t}: \ \frac{\partial \bar{U}_{1,f,t}}{\partial \tau_{f,t}} g_{1,f}^{alt}(\bar{U}_{1,f,t})\mathcal{L}\Bigg[& MRPL_{1,f,t} - W_{1,f,t} - \rho W_{2,f,t+1}(1-\tau_{f,t})B_{2,f,t+1} \\ & + \rho(1-\tau_{f,t})B_{2,f,t+1}\, V'\big((1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t}\big)\Bigg] \\ & + \rho W_{2,f,t+1}B_{2,f,t+1}B_{1,f,t} - \rho B_{2,f,t+1}B_{1,f,t}V'\big((1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t}\big) \\ & + \frac{1+\eta}{\eta}\frac{\partial \tilde{e}_{f,t}}{\partial \tau_{f,t}}\Bigg\{ d\left[F_1(B_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial \tilde{e}_{f,t}} \\ & + d\left[F_2(L_{2,f,t}, B_{1,f,t}, e^*_{2_{f,t}}, \tilde{e}_{f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial \tilde{e}_{f,t}}\Bigg\} = 0.\end{aligned}$$

Once again, the term in square brackets in lines one and two captures how the number of juniors hired by the firm in period t responds to a change in $\tau_{f,t}$. The term in the third line reflects the impact on the retention of senior employees, while the terms in the fourth and fifth lines account for the change in equilibrium effort induced by the variation in $\tau_{f,t}$. Note that this last effect can be positive or negative, depending on the shape of $\phi(\cdot)$ and the value of $\tau_{f,t}$. For instance, if $\phi(\cdot)$ is symmetric around zero, the equilibrium effort increases with $\tau_{f,t}$ when $\tau_{f,t} < 0$ and decreases when $\tau_{f,t} > 0$. It is also worth recalling that, from the envelope theorem, $V'(L_{2,f,t+1})$ is given by (11).

A.2 Proofs of Propositions in the Main Text

A.2.1 Proof of Proposition 1

From (6), we have

$$\frac{\partial \bar{U}_{1,f,t}}{\partial W_{1,f,t}} = \alpha_f.$$

Meanwhile, from (5), (6), and the definition of $\bar{\bar{U}}_{2,f,t+1}$, we have that

$$\frac{\partial \bar{U}_{1,f,t}}{\partial W_{2,f,t+1}} = \rho(1-\tau_{f,t})B_{2,f,t+1}\alpha_f.$$

From the firm's first-order condition with respect to $W_{1,f,t}$,

$$g_{1,f}^{alt}(\bar{U}_{1,f,t})\mathcal{L}\Bigg[MRPL_{1,f,t} - W_{1,f,t} - \rho W_{2,f,t+1}(1-\tau_{f,t})B_{2,f,t+1} + \rho(1-\tau_{f,t})B_{2,f,t+1}V'\big((1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t}\big)\Bigg] = \frac{B_{1,f,t}}{\alpha_f}.$$

Substituting the left-hand side for the right-hand side on the firm's first-order condition with respect to $W_{2,f,t+1}$,

$$\begin{aligned}
&\rho(1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t} - \rho(1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t} \\
&\quad + \rho(1-\tau_{f,t})B_{1,f,t}\alpha g_{2,f}^{alt}(\bar{U}_{2,f,t+1})\left[-W_{2,f,t+1} + MRPL_{2,f,t+1}\right] \\
&\quad + \frac{1+\eta}{\eta}\frac{\partial \tilde{e}_{f,t}}{\partial W_{2,f,t+1}}\Bigg\{d\left[F_1(B_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e_{2,f,t}^*)\right]^{\frac{1}{\eta}}\frac{\partial F_1}{\partial \tilde{e}_{f,t}} \\
&\qquad + d\left[F_2(L_{2,f,t}, B_{1,f,t}, e_{2,f,t}^*, \tilde{e}_{f,t})\right]^{\frac{1}{\eta}}\frac{\partial F_2}{\partial \tilde{e}_{f,t}}\Bigg\} = 0. \qquad \text{(OA.2)}
\end{aligned}$$

The terms in the first line of (OA.2) cancel each other out. Then, we can conclude the following:

- In the Irrelevant Effort scenario, the terms in the third and fourth lines of (OA.2) equal zero. Under the assumption that $\tau_{f,t} \in (0,1)$, the prefactor in the second line is positive. Therefore, we have $W_{2,f,t+1} = MRPL_{2,f,t+1}$.
- In the Productive Effort scenario, the term in the third and fourth lines of (OA.2) are strictly positive, from Lemma 1. We must then have $W_{2,f,t+1} > MRPL_{2,f,t+1}$.

A.2.2 Proof of Proposition 2

From (16) and $\partial \bar{U}_{1,f,t}/\partial W_{1,f,t} = \alpha_f$, we obtain

$$\alpha_f g_{1,f}^{alt}(\bar{U}_{1,f,t}) = \frac{\psi_1}{W_{1,f,t}} G_{1,f}^{alt}(\bar{U}_{1,f,t}). \qquad \text{(OA.3)}$$

Moreover, from (11), we have

$$V'\Big((1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t}\Big) = \text{MRPL}_{2,f,t+1}. \qquad \text{(OA.4)}$$

Substituting the expressions in (OA.3) and (OA.4) into the first-order condition for $W_{1,f,t}$, and rearranging, yields

$$\begin{aligned}&\text{MRPL}_{1,f,t} + \rho(1-\tau_{f,t})G^{alt}_{2,f}(\bar{U}_{2,f,t+1})\,\text{MRPL}_{2,f,t+1}\\ &= W_{1,f,t} + \rho(1-\tau_{f,t})G^{alt}_{2,f}(\bar{U}_{2,f,t+1})\,W_{2,f,t+1} + \frac{W_{1,f,t}}{\psi_{1,f,t}}.\end{aligned}$$

Finally, combining (14), (15), and (18), we can rewrite this condition as

$$\frac{\text{NPV}^{\text{W}}_{f,t}}{\text{NPV}^{\text{MRPL}}_{f,t}} = \frac{\psi_{PD,f,t}}{1+\psi_{PD,f,t}},$$

which corresponds to the statement of the proposition.

A.2.3 Proof of Proposition 3

Proof. By Proposition 1, under irrelevant effort we have $W_{2,f} = MRPL_{2,f}$. By Proposition 2, monopsony implies $NPV^{W}_{f} < NPV^{MRPL}_{f}$, that is,

$$W_{1,f} + \rho(1-\tau_f)G^{alt}_2(\bar{U}_{2,f})W_{2,f} < MRPL_{1,f} + \rho(1-\tau_f)G^{alt}_2(\bar{U}_{2,f})MRPL_{2,f}.$$

Using $W_{2,f} = MRPL_{2,f}$, it follows that $W_{1,f} < MRPL_{1,f}$.

Consider now a policy imposing $W_{1,f} = MRPL_{1,f}$. 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 $W_{1,f}$ raises junior utility $\bar{U}_{1,f}$, and therefore increases junior employment:

$$L_{1,f} = G^{alt}_{1,f}(\bar{U}_{1,f})\mathcal{L} \quad \Rightarrow \quad L_{1,f} \uparrow .$$

Since $U_{2,f}$ is unchanged, $G^{alt}_{2,f}(U_{2,f})$ is unchanged, and thus

$$L_{2,f} = (1-\tau_f)G^{alt}_{2,f}(\bar{U}_{2,f})L_{1,f} \quad \Rightarrow \quad L_{2,f} \uparrow .$$

Because $\partial F_s/\partial L_{k,f} \geq 0$, output increases:

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

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

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

□

A.2.4 Proof of Proposition 4

Proof. By Proposition 1, under productive effort we have $W_{2,f} > MRPL_{2,f}$. Under perfect competition, $NPV_f^W = NPV_f^{MRPL}$, so

$$W_{1,f} + \rho(1-\tau_f)G_{2,f}^{alt}(\bar{U}_{2,f})W_{2,f} = MRPL_{1,f} + \rho(1-\tau_f)G_{2,f}^{alt}(\bar{U}_{2,f})MRPL_{2,f}.$$

Since $W_{2,f} > MRPL_{2,f}$, it follows that $W_{1,f} < MRPL_{1,f}$.

Now impose $W_{1,f} = MRPL_{1,f}$. To preserve $NPV_f^W = NPV_f^{MRPL}$, the contract must become less backloaded, implying that $W_{2,f}$ decreases. From the incentive-compatibility condition,

$$C'(\tilde{e}_f) = \rho\phi(\kappa^P(\tau_f))\left[\bar{\bar{U}}_{2,f} - \bar{U}_{2,f}^{alt}\right],$$

and since $\bar{\bar{U}}_{2,f}$ is increasing in $W_{2,f}$, it follows that $\tilde{e}_f$ is increasing in $W_{2,f}$. Hence,

$$W_{2,f} \downarrow \quad \Rightarrow \quad \tilde{e}_f \downarrow .$$

Because effort is productive, lower $\tilde{e}_f$ reduces output for any given employment levels, so $Q_{s,f}$ decreases holding $(L_{1,f}, L_{2,f})$ 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:

$$L_{1,f} \downarrow, \quad L_{2,f} \downarrow .$$

This further reduces output:

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

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

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

□

A.3 Proof of Corrolary 1

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

$$NPV_f^W < NPV_f^{MRPL}.$$

Therefore, raising $W_{1,f}$ toward $MRPL_{1,f}$ 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 $W_{2,f}$. Hence, if imposing $W_{1,f} = MRPL_{1,f}$ 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 $W_{1,f}$ 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: Additional Material

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:

$$L_{sf}H_{sf} = \varsigma_{fg}L_{sg}H_{sg}, \qquad O_f = \varsigma_{fg}O_g, \tag{OA.5}$$

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

$$\begin{aligned} Q_{sg} &= \sum_{f\in g}\left[\exp(\omega_{sf}+\varepsilon_{sf})\prod_k (L_{skf}H_{skf})^{\beta_{sk}}\boldsymbol{O}_{sf}^{\beta_s^o}\right] \\ &= \left[\sum_{f\in g}\exp(\omega_{sf}+\varepsilon_{sf})\,\varsigma_{fg}^{\sum_k\beta_{sk}+\beta_s^o}\right]\prod_k (L_{skg}H_{skg})^{\beta_{sk}}\boldsymbol{O}_{sg}^{\beta_s^o}. \end{aligned} \tag{OA.6}$$

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

$$q_{sg} = \sum_k \beta_{sk}(l_{skg}+h_{skg}) + \beta_s^o \boldsymbol{o}_{sg} + \omega_{sg}, \tag{OA.7}$$

where the group productivity term is $\omega_{sg} = \ln\big(\sum_{f\in g}\exp(\omega_{sf}+\varepsilon_{sf})\varsigma_{fg}^{\sum_k\beta_{sk}^l+\sum_m\beta_{sm}^o}\big)$. Hence, the output elasticities (β_{sk}, β_s^o) are preserved under aggregation.

We impose the following bin-level law of motion for the aggregate productivity term:

$$\omega_{sg} = \sigma_s\tilde{\omega}_{sg} + \nu_{sg}, \tag{OA.8}$$

for a bin-level innovation ν_{sg} that aggregates the firm-level shocks ν_{sf}.

This condition should be interpreted as a maintained assumption on the bin-level residual,

rather than as an exact implication of firm-level AR(1) productivity. In particular, the bin-level productivity term is given by

$$\omega_{sg} = \ln\left(\sum_{f \in g} \exp(\omega_{sf} + \varepsilon_{sf})\, \varsigma_{fg}^{\Gamma_s}\right), \qquad \Gamma_s \equiv \sum_k \beta_{sk}^l + \sum_m \beta_{sm}^o, \tag{OA.9}$$

which is a log-sum aggregate of firm-level productivities.

Replacing firm-level variables with bin-level aggregates, the residual used in estimation is

$$\nu_{sg} = q_{sg} - \sigma_s \tilde{q}_{sg} - \Big(\sum_k \beta_{sk}(l_{skg} + h_{skg}) - \sigma_s \sum_k \beta_{sk}(\tilde{l}_{skg} + \tilde{h}_{skg})\Big) - \Big(\beta_s^o \boldsymbol{o}_{sg} - \sigma_s \beta_s^o \tilde{\boldsymbol{o}}_{sg}\Big) - (1 - \sigma_s) c_s \tag{OA.10}$$

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

$$E(\nu_{sg} \mid \boldsymbol{L}_{sg}\boldsymbol{H}_{sg}, \tilde{\boldsymbol{L}}_{sg}\tilde{\boldsymbol{H}}_{sg}) = 0. \tag{OA.11}$$

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

$$E(\nu_{sg} \mid \tilde{\boldsymbol{O}}_{sg}) = 0. \tag{OA.12}$$

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.7) 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 Computing Promotion Probabilities

B.2.1 Career Profile for the Representative Worker

We assume that the representative worker in our dataset joins the public accounting firm at the age of 23, having taken a 4-year college degree and passing CPA licensing, which usually happens through a one-year master's degree. The seniority grades listed in the reports are "0-2 years of experience", "3-5 years of experience", "6-8 years of experience","9+ years of experience", "non-equity partner", and "partner". We assume that the rank "9+ years of experience" is held for six years before being promoted to non-equity partner, which therefore happens at 15 years of seniority. We assume that the non-equity partnership is being held for 3 years before possible promotion to equity partner, which implies promotion to non-equity partner at 38 and to equity partner at 41. We choose 6 and 3 years to match the IPA reports, which report average experience to reach the 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.[36]

Denote the annual probability of leaving the firm, either voluntarily or involuntarily, as ϑ_s, for each seniority s. 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) ϑ^{neqp}. Similarly, equity partners stay at the firm until retirement, with a different annual probability of separating prior the the retirement age ϑ^{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 puts the entire career at 42 years.

B.2.2 Estimating Promotion Probabilities

Let $s = 1, 2, .., 6$ denote the 'partner-track' grades in increasing seniority, with $s = 1$ being the entry-level grade and $s = 6$ being equity partnership. We assume that non-partner employees never leave within a grade band s. Denote the number of employees in each grade by N_s and the length in years of grade s by T_s. Assuming a steady-state labor force, the ex-ante cumulative probability of being promoted to grade s when starting the career at firm f is denoted by $\lambda_{s,f}$. For all grades except non-equity and equity partners, we compute these probabilities using averages of

[36] In order to obtain an ex-ante promotion probability to the 9+ years rank that is not higher than the ex-ante probability of being in the 5-8 seniority level, workers need to spend at least 5 years in the 9+ seniority rank before being promoted to the partnership.

the employee counts firm-by-firm, taken across years.

$$\lambda_{s,f} = \frac{N_{s,f}/T_s}{N_{1,f}/T_1}, \quad s \in \{1, \ldots, 4\}.$$

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:

$$\lambda_{s,f} = \frac{N_{s,f}^{new}}{N_{1,f}/T_1}, \quad s \in \{5, 6\}.$$

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

B.2.3 Partner Turnover

Having reached partnership, we assume that individuals have annual separation probabilities ϑ_5 and ϑ_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:

$$N_s = N_s^{\text{new}} \sum_{t=0}^{T_s} (1 - \vartheta_s)^t \quad s \in \{5, 6\}$$

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

$$\frac{N_s}{N_s^{\text{new}}} = \frac{1 - (1 - \vartheta_s)^{T_i}}{\vartheta_s} \tag{OA.13}$$

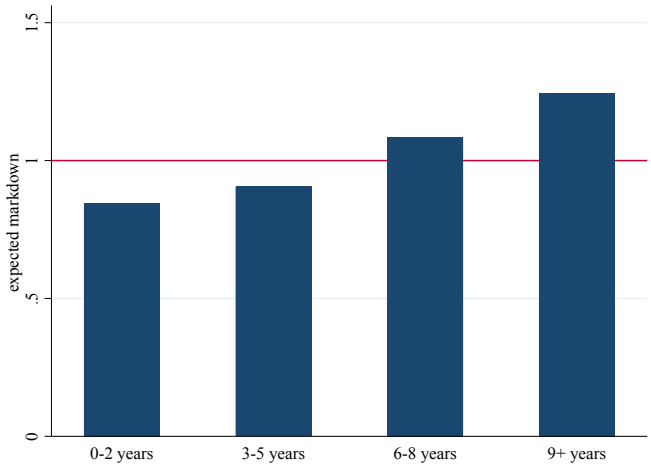
Since the latter equation lacks a closed-form solution for ϑ_s, we solve it numerically. We find $\vartheta_5 = .170$ and $\vartheta_6 = .041$. This means that non-equity partners have an annual probability of leaving their position of 17%, which includes leaving the firm and being promoted to equity partner. Meanwhile, equity partners have an annual probability of separating from the firm of 4.1%.

B.3 Observed vs. Model-implied Seniority Levels

Denote the observed seniority levels by $\tilde{s}$ and the 'true' seniority levels at which firms set wages by s. Our result in Appendix D 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 NPV wedge between the MRPL and wages at any observed seniority level $\tilde{s}$.

The results of this exercise are in Figure OA-1. We find that the expected wage starts exceeding the expected MRPL at around 6–8 years of seniority. This implies that the promotion to seniority

Figure OA-1: 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.

level $s = 2$ does not happen at the observed seniority levels $\tilde{s} \in \{0\text{–}2\,\text{y}, 3\text{–}5\,\text{y}\}$, which thus correspond to the 'junior' seniority level $s = 1$ in the model. The promotion to the 'senior' grades $s > 1$ should happen at some $\tilde{s} \in \{6\text{–}8\,\text{y}, 9+\,\text{y}, \text{non-equity partner}, \text{equity partner}\}$. Given that the largest drop in promotion rates occurs just before the non-equity partnership, and that wages start exceeding the MRPL at that seniority level, it is natural to interpret all non-partner professional staff as the 'junior' grade $s = 1$ and non-equity and equity partners as the 'senior' grade $s = 2$ (or, possibly, non-equity and equity partners could correspond to $s = 2$ and $s = 3$, respectively).

B.4 Paraprofessional Wage Markdown

Given that paraprofessionals are directly billed, we can estimate their marginal revenue product and wage markdown similarly to the professional staff. When we do so, we obtain a wage markup of 109%. However, this wedge is, in contrast to the professional staff, very imprecisely estimated. Given that paraprofessionals are not part of the partnership track, and therefore do not impact the computation of lifecycle wage markdowns and labor supply elasticities, we omit them from our main analysis.

B.5 Equity Partner Wage Markup Using Lump-Sum Payments

Given that equity partners are residual claimants of profits, it is key to distinguish their lump-sum payments from their profit payouts, as discussed in Section 3.6.1. For the test of backloaded incentive pay, the lump-sum payment needs to be compared to the equity partner MRPL, whereas the dataset arguably includes total compensation.

Denoting measured total equity partner compensation as $W^{obs}_{f,t}$, we allow it to include both

Figure OA-2: Paraprofessionals Markdown

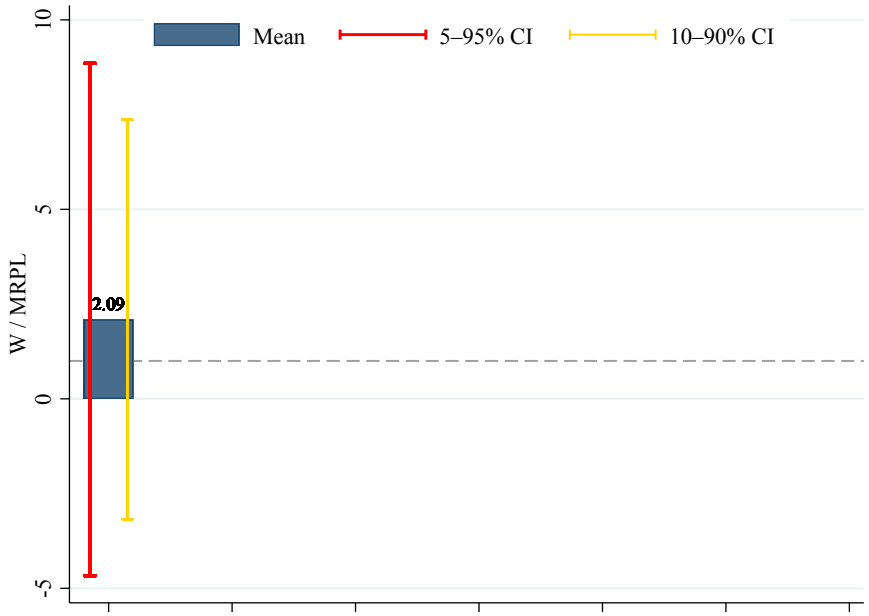
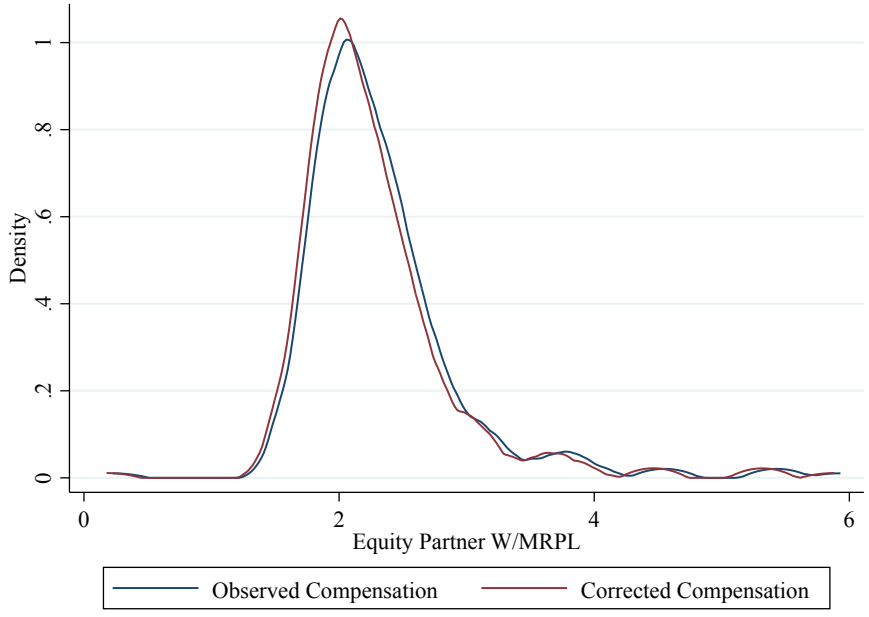


Figure OA-3: Equity Partner Markup with Lump-Sum Pay Correction



Notes: Figure OA-2 reports the average instantaneous wage markdown for paraprofessionals; confidence intervals are block-bootstrapped using 250 iterations. Figure OA-3 plots the equity partner wage markup distributions using observed compensation data and using imputed lump-sum payments only.

lump-sum payments and the profit share per equity partner, $\pi^s_{f,t}$:

$$W^{obs}_{f,t} = W^{lump}_{f,t} + \pi^s_{f,t}.$$

We assume steady-state profits, which allows us to drop time subscripts from profits and implies that buy-ins equate buy-outs at $b^{in}_f = b^{out}_f = \pi^s_f/(1-\rho)$. Substituting into Equation (19) and rearranging, we obtain:

$$W^{lump}_{f,t} = W^{obs}_{f,t} - (1-\rho)b^{in}_f. \tag{OA.14}$$

The dataset reports average buy-ins per firm-bin-year, which we use to compute lump-sum payments using Equation (OA.14). The resulting wage markup distribution using the imputed lump-sum payments is depicted as the red line in Figure OA-3, and is very close to the original wage markups that were computed using the raw compensation data. The average wage markup for equity partners drops from 2.32 to 2.27 when correcting for profit payouts. Therefore, this result is robust to correcting for profit payouts.

C Robustness Checks

C.1 Alternative Production Function Specifications

C.1.1 Adding Fixed Overhead Costs

Although the data does not report fixed costs and overhead, such as office rent and insurance costs, we can impute them by imposing a zero-profit condition. Such a 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 \$100,000 and \$62,000, respectively. The administrative salary benchmark is based on administrator wages observed in the IPA reports, and the internship salary is based on Glassdoor.[37]

We then back out fixed costs as

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

We find a ratio of fixed costs to revenue of 26.5% on average, which is slightly higher than the 20–25% rate reported in industry reports.[38]

As a robustness exercise, we add the logarithm of fixed costs as an input 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, reported in

[37] https://www.glassdoor.com/Salaries/accounting-intern-salary-SRCH_KO0,17.htm

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

Table OA-1, are very similar to those from the baseline model in which fixed costs are excluded.

Figure OA-4 shows that, compared to our baseline specification, the production function specification with fixed costs leads to very similar markdown estimates for the various seniority levels.

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

	(I) 0-2 years		(II) 3-5 years		(III) 6-8 years		(IV) 9+ years		(V) Non-eq. Part.		(VI) Eq. Part.		(VII) Paraprof.	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
0-2 years	0.997	0.019	0.028	0.014	0.020	0.020	0.005	0.019	0.013	0.027	0.013	0.027	-0.010	0.052
3-5 years	0.016	0.025	0.993	0.017	-0.033	0.022	-0.010	0.019	0.024	0.030	0.024	0.030	-0.028	0.055
6-8 years	0.001	0.022	0.009	0.013	1.003	0.018	-0.030	0.018	0.039	0.028	0.039	0.028	0.061	0.056
9 years	-0.021	0.024	-0.039	0.014	-0.001	0.022	0.968	0.019	-0.032	0.034	-0.032	0.034	-0.003	0.064
Non-eq. Partners	0.015	0.013	-0.001	0.007	0.010	0.012	-0.007	0.010	0.963	0.015	-0.037	0.015	-0.009	0.029
Eq. Partners	0.031	0.023	0.033	0.017	0.078	0.025	0.066	0.023	0.035	0.046	1.035	0.046	0.039	0.062
Paraprof.	-0.034	0.011	-0.029	0.007	-0.023	0.011	-0.033	0.009	-0.021	0.014	-0.021	0.014	0.975	0.026
Interns	-0.013	0.017	0.005	0.010	-0.019	0.013	-0.001	0.011	-0.027	0.016	-0.027	0.016	-0.017	0.034
Training	-0.000	0.009	0.003	0.005	0.006	0.008	-0.007	0.008	0.023	0.012	0.023	0.012	-0.005	0.019
Recruiting	-0.008	0.010	0.007	0.005	-0.002	0.009	0.012	0.008	-0.023	0.013	-0.023	0.013	0.022	0.024
Technology	0.000	0.011	-0.007	0.007	-0.009	0.009	0.005	0.010	-0.014	0.012	-0.014	0.012	-0.032	0.027
Admin. staff	0.000	0.023	-0.024	0.011	-0.020	0.016	0.010	0.013	-0.003	0.017	-0.003	0.017	0.032	0.041
Capital	0.005	0.013	0.013	0.007	-0.003	0.011	0.009	0.009	0.007	0.017	0.007	0.017	0.021	0.029
Ser. corr.	0.568	0.132	0.730	0.110	0.863	0.062	0.896	0.053	0.923	0.040	0.923	0.040	1.000	0.061
Obs.	165		165		165		165		165		165		165	

C.1.2 Adding Wage and Promotion Controls in the Production Function

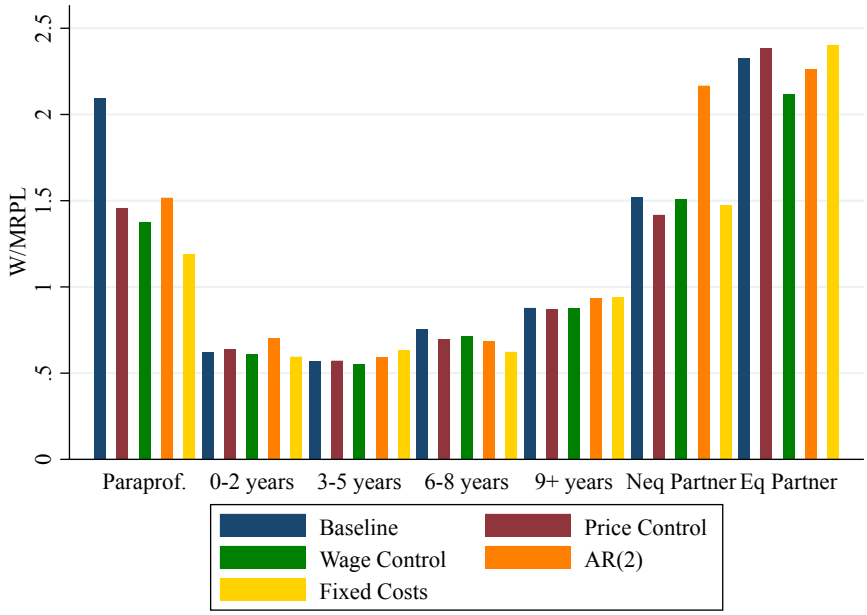
In the main model, we imposed the AR(1) transition model on the entire TFP residual for each worker seniority s, including the part of TFP that is due to worker effort. An alternative is to assume log-additivity between the 'effort-unrelated' and 'effort-related' productivity terms—which we denote, respectively, by $\tilde{\omega}_{s,f}$ and $b_s(e_{s,f})$—and impose the AR(1) assumption on the former. As for the effort-related productivity term $b_s(e_{s,f})$, we assume it is a deterministic linear function of the log of equity partner compensation, $W_{s=part,f}$, and the log probability of being promoted to equity partner from level s, $\lambda_{part,s,f}$—note that our model implies that junior effort is indeed a function of these two variables. Specifically, let

$$b_s(e_{s,f}) = c_s + \beta_s^w \ln(W_{s=part,f}) + \beta_s^\lambda \ln(\lambda_{s=part,f}).$$

The productivity shock expression, which can still be used to estimate $(\sigma_s, \boldsymbol{\beta}_s)$, becomes:

$$\nu_{s,f} = q_{s,f} - \sigma_s \hat{q}_{s,f} - \Big(\sum_k \beta_{s,k}^l (l_{k,f} + h_{k,f}) - \sigma_s \sum_k \beta_{s,k}^l (\hat{l}_{k,f} + \hat{h}_{k,f}) \Big) - (\beta_s^o \boldsymbol{o}_f - \sigma_s \beta_s^o \hat{\boldsymbol{o}}_{s,f})$$
$$-\beta_s^w \big(\ln(\hat{W}_{s=part,f}) - \sigma_s \ln(W_{s=part,f}) \big) - \beta_s^\lambda \big(\ln(\lambda_{s=part,f}) - \sigma_s \ln(\hat{\lambda}_{s=part,f}) \big) - (1-\sigma_s) c_s.$$

Figure OA-4: Alternative Production Function Specifications



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

We implement this version of the model, including equity partner compensation and transition rates as predetermined inputs, as they are determined ex ante.

Figure OA-4 shows that markdown estimates are very similar between the main model and the extension in which wages and promotion rates are included in the production function.

C.1.3 Allowing for Product Differentiation across Firms

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

$$q_{s,f} = \sum_{k \in S} \left(\beta^{l}_{s,k}(l_{k,f} + h_{k,f})\right) + \beta^{o}_{s}\boldsymbol{o}_{s,f} + \beta^{p}\ln(p_{sf}) + \omega_{s,f} + v_{s,f}. \tag{OA.15}$$

We assume that firms can flexibly adjust prices even after observing the productivity shocks. Therefore, we include lagged prices in the instruments vector, in addition to the other instruments from the baseline specification.

The resulting wage markdowns are again very similar to those estimated without price control, as shown in Figure OA-4.

C.1.4 AR(2) Productivity Transition

As a robustness check, we impose an AR(2) model on the productivity transition, rather than the AR(1) model in the main text. Denoting two lags of any variable X as $\tilde{X}$ and the AR(2) coefficient as $\tilde{\sigma}_s$, we get:

$$\omega_{s,f} = c_s + \sigma_s(\hat{\omega}_{s,f} + \hat{v}_{s,f}) + \tilde{\sigma}_s(\tilde{\omega}_{s,f} + \tilde{v}_{s,f}) + \nu_{s,f}. \tag{OA.16}$$

As a result, the moment conditions can be expressed as a function of current, once-lagged, and twice-lagged values of every input variable, rather than just the first two. We re-estimate the production function with the moment conditions under the AR(2) process, including up to two lags as instruments rather than just one. The resulting markdowns, shown in Figure OA-4, are very similar to those from the baseline AR(1) specification.

C.2 Single-Product 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 single-product production model and compare its implied markdowns to our (multi-product) baseline team production model. Given that it is not obvious how 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 the following Cobb-Douglas specification:

$$r_f = \sum_{k \in \mathcal{S}} \left(\beta_k (l_{kf} + h_{kf})\right) + \beta^o \boldsymbol{o}_f + \omega_f + \nu_f. \tag{OA.17}$$

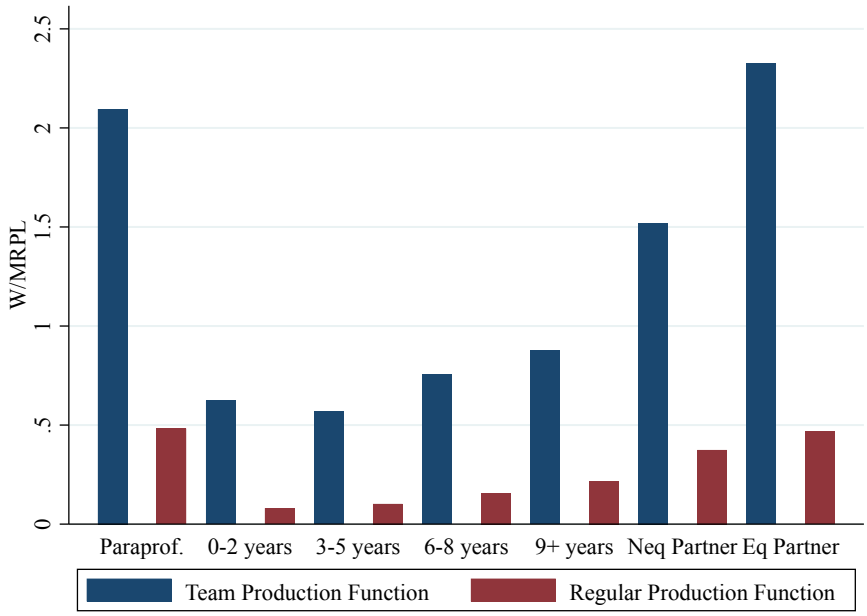
For the estimation, we use the same moment conditions and TFP transition process as assumed in the main text.

The production function estimates are in Table OA-2, and the resulting markdowns in Figure OA-5. The single-product model yields unrealistically large markdowns throughout—possibly because input-specific team production functions do not aggregate to a firm-level Cobb-Douglas (or translog) specification, even when the product market is perfectly competitive.

Table OA-2: Single-Product Production Function

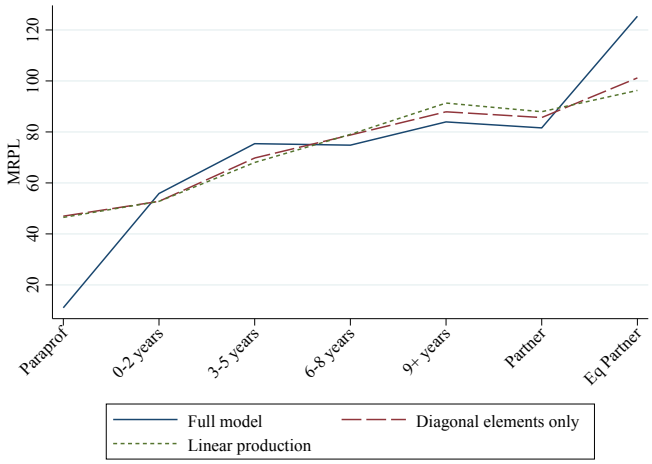
	(I)	
	Est.	S.E.
0-2 years	0.228	0.033
3-5 years	0.221	0.028
6-8 years	0.130	0.034
9 years	0.202	0.037
Non-eq. Partners	0.056	0.016
Eq. Partners	0.259	0.039
Paraprof.	0.015	0.015
Interns	-0.056	0.021
Training	-0.006	0.015
Recruiting	0.013	0.005
Technology	0.012	0.016
Admin. staff	-0.054	0.019
Ser. corr.	0.971	0.032
Obs.	174	

Figure OA-5: Single-Product Production Function: Markdowns



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

Figure OA-6: MRPL comparison across models



Notes: We estimate the MRPL at every seniority level under three models: the full model (nonlinear and team production), a restricted model that only takes into account the diagonal elements of the production model (nonlinear indepedent production) and the restricted model that imposes linear production (MRPL = R/L).

C.3 Marginal Revenue Products: Comparison of Production Models

To assess how important the off-diagonal elements in the output elasticities matrix are, and the importance of allowing for nonlinear production, we re-estimate the MRPL under two restricted versions of our model. First, we only take into account the diagonal elements for the output elasticities, setting all off-diagonal elements to zero. This implies that production is independent for each seniority level. Second, in addition, we set the own-diagonal elements to one, implying linear and independent production.

The resulting MRPL ladder is in Figure OA-6. The linear and nonlinear models with independent production (green dotted line and red dashed line) are closely aligned, which is unsurprising given that the own-output elasticities are all estimated to be close to one.

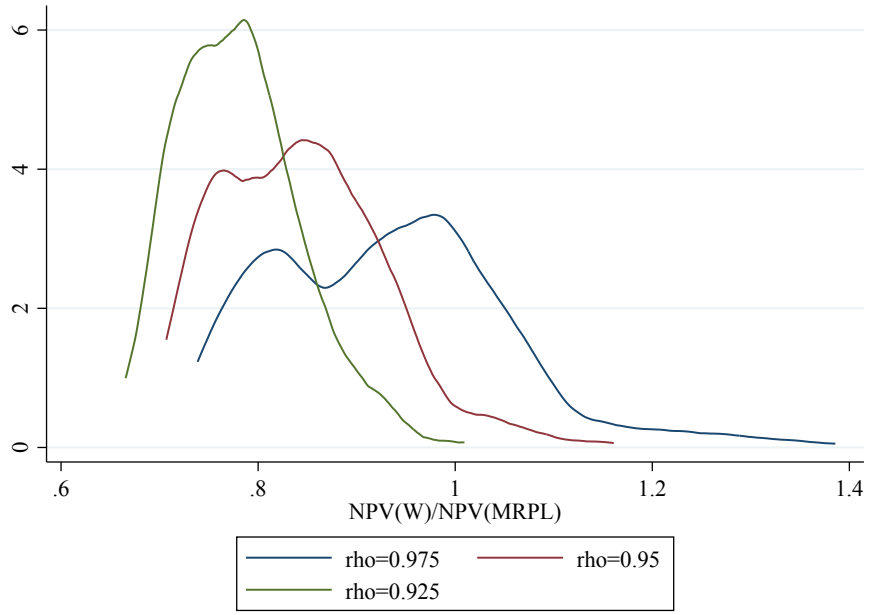
However, allowing for team production (the full model, represented by the solid blue line) markedly changes the MRPL ladder. Most affected are the estimated MRPL of equity partners, which is underestimated under independent production given their positive cross-output elasticities on everyone else, and the MRPL of paraprofessionals, which is overestimated under independent production given their negative cross-output elasticities. Hence, the estimated MRPL ladder becomes steeper under our full model compared to simpler specifications that rule out team production.

C.4 Alternative 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 distributions of the expected markdown ($\text{NPV}^{W}/\text{NPV}^{MRPL}$) are in Figure OA-7. A

higher discount factor pushes the distribution closer to one, as employees are more patient and, therefore, value future wage markups more. The expected markdown is 0.78 if $\rho = 0.925$ and 0.94 if $\rho = 0.975$, compared to 0.85 in the main specification. Although different discount factors result in markedly different expected markdowns, all these markdowns are still below one. The discount factor required to obtain $NPV(W) = NPV(MRPL)$ is $\rho = 0.99$, which is unrealistically high.

Figure OA-7: Robustness: Discount Factors



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

C.5 Bundle-Pricing vs. Line-Pricing

C.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 on a bundled price for the entire accounting service provided by the firm. This was motivated by the IPA reports, which indicate that the vast majority of surveyed firms follow such line-pricing practices. In this appendix, we derive the markdown and MRPL expressions if firms would price bundles instead.

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

$$Q_f = \left[\sum_{s=1}^{S} a_s Q_{sf}^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}} \tag{OA.18}$$

The implied bundle price is:

$$P_f = \left[\sum_{s=1}^{S} a_s^{\sigma} P_{sf}^{1-\sigma} \right]^{\frac{1}{1-\sigma}}$$

Demand for the bundle is

$$Q_f = \overline{Q} \left(\frac{P_f}{\overline{P}} \right)^{\eta} \zeta_f. \tag{OA.19}$$

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

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

C.5.2 Recovering Product Weights

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

$$\frac{Q_s}{Q_j} = \left(\frac{a_s}{a_j} \right)^{\sigma} \left(\frac{P_s}{P_j} \right)^{-\sigma}. \tag{OA.20}$$

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

$$\kappa_s \equiv \left(\frac{Q_s}{Q_7} \right)^{1/\sigma} \frac{P_s}{P_7}, \qquad s = 1, \ldots, 6. \tag{OA.21}$$

The relative demand conditions then become

$$a_s = \kappa_s a_7, \qquad s = 1, \ldots, 6. \tag{OA.22}$$

We impose the normalization

$$\sum_{i=1}^{7} a_s = 1. \tag{OA.23}$$

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

$$a_s = \frac{\kappa_s}{1 + \sum_{k=1}^{6} \kappa_k}, \tag{OA.24}$$

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

$$a_7 = \frac{1}{1 + \sum_{s=1}^{6} \kappa_s}. \tag{OA.25}$$

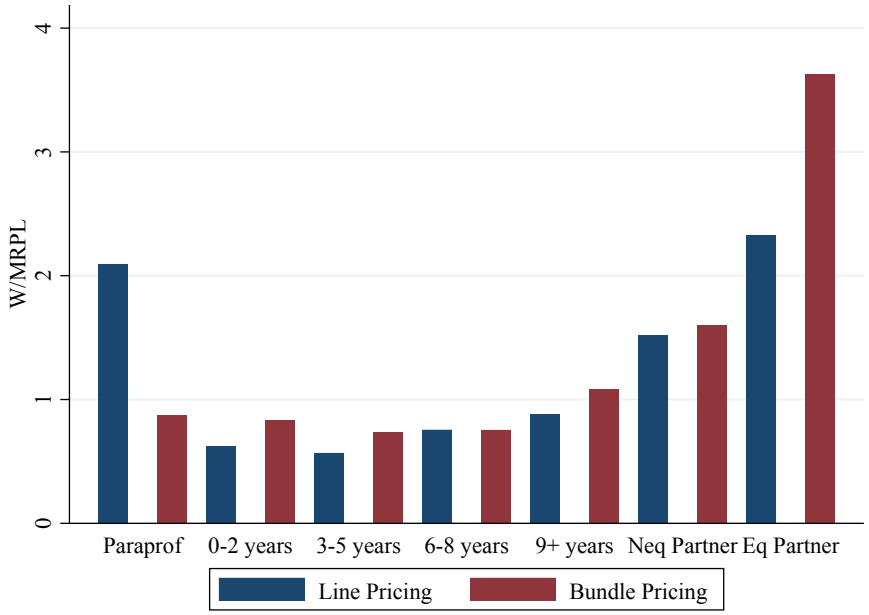
C.5.3 Markdown Estimation

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

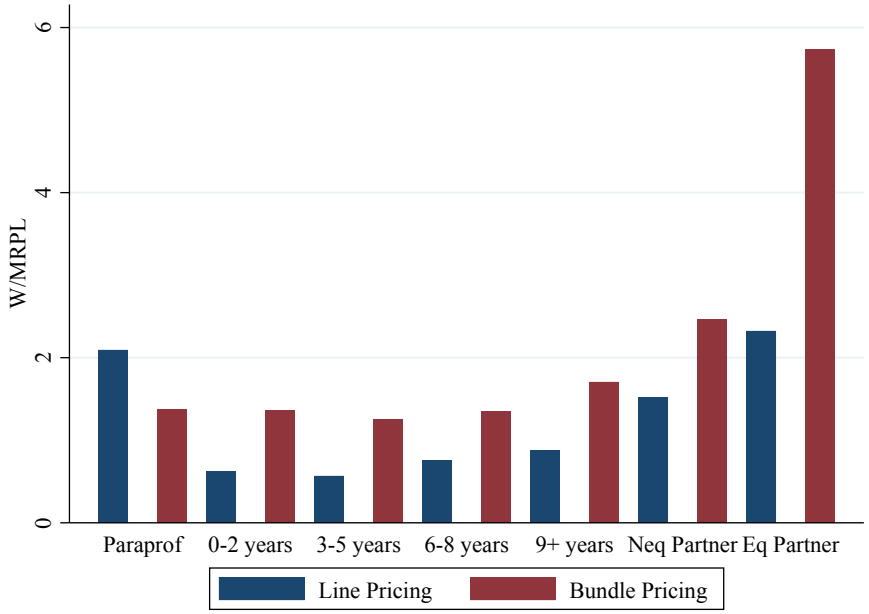
The results are in Figure OA-8. We find similar markdowns for line pricing and bundle pricing when the accountants are gross complements ($\sigma = 0.5$), although equity partner wage markups are slightly smaller under the bundle specification. However, when assuming that the accountants are gross substitutes ($\sigma = 2$), all employees are estimated to earn wage markups—an impossibility, as this would imply loss-making firms.

Figure OA-8: Wage Markdowns: Bundle Pricing vs. Line Pricing

(a) Gross complements ($\sigma = 0.5$)



(b) Gross substitutes ($\sigma = 2$)



Notes: Average wage markdowns by seniority level under bundle pricing, for $\sigma = 0.5$ and $\sigma = 2.0$.

D Extending the Model to N Employee Levels

We extend the rank-order tournament framework presented in Section 3 to an environment with N employee levels: a junior level $s = 0$; $N - 2$ 'intermediate' levels $s = 1, \ldots, N - 2$; and a 'terminal' senior level $s = N - 1$. Workers are potentially productive for all 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 3.1.5, 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 would be conceptually direct, albeit at the cost of heavier notation.

D.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 $\phi(\cdot)$.

D.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.

D.3 Compensation–Promotion Policy

At hiring (period t), the firm commits 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 $s+1$.[39] 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.

D.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

$$\bar{U}^{alt}_{s} \equiv \mathbb{E}[U^{alt}_{i,s}], \qquad \bar{\bar{U}}_{s,t} \equiv \mathbb{E}\Big[\max\{U_{s,t}, U^{alt}_{i,s}\}\Big] = G^{alt}_{s}(U_{s,t})U_{s,t} + \int_{U_{s,t}}^{\infty} u\, dG^{alt}_{s}(u). \tag{OA.26}$$

D.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

$$\bar{U}_{N-1,t} \ =\ \alpha_{N-1}W_{N-1,t} - C(e^{*}_{N-1}), \tag{OA.27}$$

where e^{*}_{N-1} is determined exogenously.[40] For any $s = 0, 1, \ldots, N-2$, a level-s worker chooses effort e_s to maximize

$$\bar{U}_{s,t} \ =\ \max_{e_s}\ \alpha_s W_{s,t} - C(e_s) + \rho\left\{\left[1 - \Phi(\theta^{P}_{s,t}(\tau_{s,t}) - e)\right]\bar{\bar{U}}_{s+1,t} + \Phi(\theta^{P}_{s,t}(\tau_{s,t}) - e)\bar{U}^{alt}_{s+1}\right\}. \tag{OA.28}$$

[39] Note that here the notational convention for the compensation-promotion policy differs from that in the main text, in that the index t refers to the period in which a worker is hired, rather than to the current period.

[40] See Footnote 22 for a discussion of possible values of e^{*}_{N-1}

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

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

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

$$C'(\tilde{e}_{s,t}) \;=\; \rho\,\phi(\kappa^P(\tau_{s,t}))\Big[\bar{\tilde{U}}_{s+1,t} - \bar{U}^{alt}_{s+1}\Big], \qquad s = 0, 1, \ldots, N-2, \tag{OA.29}$$

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

D.6 Employment Dynamics

The number of juniors willing to work for the firm is

$$L_{0,t} \;=\; G^{alt}_0(\bar{U}_{0,t})\,\mathcal{L}. \tag{OA.30}$$

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

$$L_{s+1,t+1} \;=\; (1 - \tau_{s,t-s})\,G^{alt}_{s+1}(\bar{U}_{s+1,t-s})\,L_{s,t}. \tag{OA.31}$$

D.7 Firm's Problem

We specify the firm's problem recursively, which requires defining the state space. From (OA.30) and (OA.31), at any point in time, the stock of intermediate and 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 intermediate and 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.[41] Finally, define $l \equiv \{\vec{l}_k\}_{k=0}^{N-1}$ and $\epsilon \equiv \{\vec{\epsilon}_k\}_{k=0}^{N-2}$, for conciseness.[42] We can then write the

[41] 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.

[42] 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 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 workers in the current period and in the next one.

firm's problem as choosing $\{W_s, \tau_s\}_{s=0}^{N-1}$ (with τ_{N-1} unused) to maximize the discounted value

$$V(l,\epsilon) \;=\; \max_{\{W_s,\tau_s\}_{s=0}^{N-1}} \left\{ \sum_{s=0}^{N-1} p_s\, F_s(L_0, \vec{l}_0, e_s) \;-\; W_0 L_0 \;-\; \sum_{s=1}^{N-1} \rho^s\, W_s\, L_s^{\text{cohort}} \;+\; \rho\, V(l^+, \epsilon^+) \right\}, \quad \text{(OA.32)}$$

where L_0 is given by (OA.30), 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.29) if $s = 0$.[43] The stock of incumbent employed workers evolve as follows:

$$l_{s,k}^+ = \begin{cases} l_{s,k+1} & \text{if} \quad s \geq k+2, \\ L_s^{\text{cohort}} & \text{if} \quad s = k+1. \end{cases} \quad \text{(OA.33)}$$

Meanwhile, $\epsilon_{s,k}^+$ is equal to $\epsilon_{s,k+1}$ for $s \geq k+2$ and is given by (OA.29) if $s = k+1$. We then define $l^+ \equiv \{\vec{l}_k^+\}_{k=0}^{N-1}$ and $\epsilon^+ \equiv \{\vec{\epsilon}_k^+\}_{k=0}^{N-2}$.

D.8 Steady State

To simplify the exposition, we restrict our attention to stationary equilibria of the model, in which $(W_{s,t}, \tau_{s,t}, L_{s,t}, \tilde{e}_{s,t}) = (W_s, \tau_s, L_s, \tilde{e}_s)$ for all s. Equations (OA.26)–(OA.31) specialize to

$$\bar{\bar{U}}_s = G_s^{alt}(\bar{U}_s)\bar{U}_s + \int_{U_s}^{\infty} u\, dG_s^{alt}(u), \quad \text{(OA.34)}$$

$$\bar{U}_{N-1} = \alpha_{N-1} W_{N-1} - C(e_{N-1}^*), \quad \text{(OA.35)}$$

$$\bar{U}_s = \alpha_s W_s - C(\tilde{e}_s) + \rho\left[(1-\tau_s)\bar{\bar{U}}_{s+1} + \tau_s \bar{U}_{s+1}^{alt}\right], \quad s = 0, \ldots, N-2, \quad \text{(OA.36)}$$

$$C'(\tilde{e}_s) = \rho\, \phi(\kappa^P(\tau_s))\, [\bar{\bar{U}}_{s+1} - \bar{U}_{s+1}^{alt}], \quad s = 0, \ldots, N-2, \quad \text{(OA.37)}$$

$$L_0 = G_0^{alt}(\bar{U}_0)\mathcal{L}, \qquad L_{s+1} = (1-\tau_s) G_{s+1}^{alt}(\bar{U}_{s+1})\, L_s, \quad s = 0, \ldots, N-2. \quad \text{(OA.38)}$$

[43] 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.

D.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.32) with respect to W_s is

$$\begin{aligned}
&\frac{\partial \bar{U}_0}{\partial W_s} g_0^{alt}(\bar{U}_0)\mathcal{L}\left\{MRPL_0 - W_0 - \sum_{k=1}^{N-1} \rho^k \prod_{m=0}^{k-1} \left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_k - W_k)\right\} \\
&\quad + \sum_{j=1}^{s} \rho^j \prod_{m=1}^{j} \left[(1-\tau_{m-1})G_{m-1}^{alt}(\bar{U}_{m-1})\right] \mathcal{L}\frac{\partial \bar{U}_j}{\partial W_s} g_j^{alt}(\bar{U}_j)\left\{MRPL_j - W_j\right. \\
&\quad \left. + \sum_{k=j+1}^{N-1} \rho^{k-j} \prod_{m=j}^{k-1} \left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_k - W_k)\right\} \\
&\quad - \rho^s \prod_{j=0}^{s-1} \left[(1-\tau_j)G_{j+1}^{alt}(\bar{U}_{j+1})\right] G_0^{alt}(\bar{U}_0)\mathcal{L} \\
&\quad + \sum_{j=0}^{s-1} \frac{\partial F_j}{\partial \tilde{e}_j}\frac{\tilde{e}_j}{\partial W_s} = 0. \qquad \text{(OA.39)}
\end{aligned}$$

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.32) is

$$\frac{\partial \bar{U}_0}{\partial W_0} g_0^{alt}(\bar{U}_0)\mathcal{L}\left\{MRPL_0 - W_0 + \sum_{s=1}^{N-1} \rho^s \prod_{m=0}^{s-1} \left[(1-\tau_m)G_{m+1}^{alt}(\bar{U}_{m+1})\right](MRPL_s - W_s)\right\} = L_0. \qquad \text{(OA.40)}$$

Meanwhile,

$$\frac{\partial \bar{U}_s}{\partial W_s} = \alpha, \quad \text{for } s \geq 1 \qquad \text{(OA.41)}$$

$$\text{and} \quad \frac{\partial \bar{U}_k}{\partial W_s} = \rho^{(s-k)} \prod_{j=k}^{s-1} \left[(1-\tau_j)G_{j+1}^{alt}(\bar{U}_{j+1})\right]\alpha, \quad \text{for } s > k \geq 1. \qquad \text{(OA.42)}$$

From (OA.38), (OA.40) and (OA.41), 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.42), this allows us to rewrite the first line of (OA.39) 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.39) cancel each other out, and we can rewrite the FOC as

$$\begin{aligned}\sum_{j=1}^{s} \rho^j \prod_{m=1}^{j} \left[(1-\tau_{m-1}) G_{m-1}^{alt}(\bar{U}_{m-1})\right] \mathcal{L} \frac{\partial \bar{U}_j}{\partial W_s} g_j^{alt}(\bar{U}_j) \Bigg\{ MRPL_j - W_j \\ + \sum_{k=j+1}^{N-1} \rho^{k-j} \prod_{m=j}^{k-1} \left[(1-\tau_m) G_{m+1}^{alt}(\bar{U}_{m+1})\right] (MRPL_k - W_k) \Bigg\} \\ + \sum_{j=0}^{s-1} \frac{\partial F_j}{\partial \tilde{e}_j} \frac{\partial \tilde{e}_j}{\partial W_s} = 0. \end{aligned} \tag{OA.43}$$

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.

D.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.43) 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

$$MRPL_1 - W_1 + \sum_{k=2}^{N-1} \rho^{k-1} \prod_{m=1}^{k-1} \left[(1-\tau_m) G_{m+1}^{alt}(\bar{U}_{m+1})\right] (MRPL_k - W_k) = 0. \tag{OA.44}$$

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

Next, evaluate (OA.43) for $s = 2$, which yields

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

By (OA.44), 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 $s \geq 1$.

D.9.2 Productive Effort

Now assume that $\frac{\partial F_0}{\partial \tilde{e}_0} > 0$. Evaluating (OA.43) at $s = 1$ gives

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

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 level-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.29) yields

$$\tilde{e}_s = \frac{\rho\,\phi(\kappa^P(\tau_s))\left[\bar{\bar{U}}_{s+1} - \bar{U}_s^{alt}\right]}{2c}, \qquad s = 0, 1, \ldots, N-2. \tag{OA.45}$$

Thus, equilibrium effort at level-s is linear in continuation utility. As a consequence, using (OA.34) and (OA.36), we have that

$$\frac{\partial \tilde{e}_s}{\partial W_{s'+j}} = \rho^j \prod_{m=1}^{j} \left[(1-\tau_{s'+m-1}) G^{alt}_{s'+m}(\bar{U}_{s'+m})\right] \frac{\partial \tilde{e}_s}{\partial W_{s'}}, \tag{OA.46}$$

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

Meanwhile,

$$\frac{\partial \bar{U}_s}{\partial W_{s'+j}} = \rho^j \prod_{m=1}^{j} \left[(1-\tau_{s'+m-1}) G^{alt}_{s'+m}(\bar{U}_{s'+m})\right] \frac{\bar{U}_s}{\partial W_{s'}}, \tag{OA.47}$$

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

An implication of (OA.46) and (OA.47) is that effort and utility respond proportionally to changes in future compensation, reflecting the linear relationship between effort and continuation utility in (OA.45). 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.43) 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.43) evaluated at $s = 1$, together with (OA.46) and (OA.47), 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,

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

D.10 Entry Elasticity and the Exact Intertemporal Wedge

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

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

$$\psi_0 \equiv \frac{g^{alt}_0(\bar{U}_0)\, W_0}{G^{alt}_0(\bar{U}_0)} \cdot \frac{\partial \bar{U}_0}{\partial W_0}, \qquad \frac{\partial \bar{U}_0}{\partial W_0} = \alpha_0, \tag{OA.48}$$

where we normalize $\alpha_0 = 1$.

Also, define the steady-state *marginal revenue product of labor* of a level-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

$$\frac{\partial V(l,\epsilon)}{\partial l_{s,k}} = \rho^k MRPL_s, \quad k \in \{0, \ldots, N-2\}, \quad s \in \{k+1, \ldots, N-1\}. \tag{OA.49}$$

Substituting (OA.38) and (OA.48) on (OA.40), we obtain

$$MRPL_0 + \sum_{s=1}^{N-1} \rho^s \prod_{m=0}^{s-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\mathcal{L}\right] MRPL_s$$
$$= W_0 + \sum_{s=1}^{N-1} \rho^s W_s \prod_{m=0}^{s-1} \left[(1-\tau_m) G^{alt}_{m+1}(\bar{U}_{m+1})\mathcal{L}\right] + \frac{W_0}{\psi_0}. \tag{OA.50}$$

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

$$\text{NPV}_W \equiv W_0 + \sum_{s=1}^{N-1} \rho^s W_s \prod_{m=0}^{s-1} \left[(1-\tau_m) G_{m+1}^{alt}(\bar{U}_{m+1})\mathcal{L}\right], \tag{OA.51}$$

$$\text{NPV}_{\text{MRPL}} \equiv MRPL_0 + \sum_{s=1}^{N-1} \rho^s \prod_{m=0}^{s-1} \left[(1-\tau_m) G_{m+1}^{alt}(\bar{U}_{m+1})\mathcal{L}\right] MRPL_s. \tag{OA.52}$$

Then (OA.50) implies the exact wedge

$$\text{NPV}_{\text{MRPL}} \;=\; \text{NPV}_W \;+\; \frac{W_0}{\psi_0}. \tag{OA.53}$$

By defining the present-discounted-compensation elasticity of labor supply as

$$\psi_{PD} = \psi_0 \frac{NPV_W}{W_0}, \tag{OA.54}$$

and substituting on (OA.53), we obtain an equation equivalent to that in the statement of Proposition 2.

E Incorporating Random Productivity Shocks into the Model

We extend the theoretical model in Section 3 by introducing a random productivity shock into the firms' production functions. This aligns our theoretical analysis with the production function estimation procedure described in Section 4.2, which relies on such shocks.

Specifically, let the output attributable to level-s workers in firm f, period t be

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

where $\Omega_{s,f,t}$ denotes a level-firm-period-specific productivity state, and all other terms are as defined in (2). We assume that $\ln \Omega_{s,f,t}$ follows an AR(1) process.

The introduction of the productivity state leaves the workers' optimization problem and the resulting effort and retention conditions from Section 3 intact. On the firms' side, define the level-s revenue of firm f in period t as

$$R_s(L_{s,f,t}, L_{-s,f,t}, \bar{e}_{s,f,t}, \bar{e}_{-s,f,t}, \Omega_{s,f,t}) = d\left[Q_{s,f,t}\right]^{\frac{1+\eta}{\eta}}, \quad s \in \{1,2\}$$

where d and η are as in (4). Then the optimization problem defining the choice of the compensation–promotion policy $\{W_{1,f,t}, W_{2,f,t+1}, \tau_{f,t}\}$ for workers hired in period t (originally defined in (10) for the baseline model) becomes

$$\begin{aligned} V(L_{2,f,t}, \Omega_{s,f,t}) = \max_{W_{1,f,t}, W_{2,f,t+1}, \tau_{f,t}} \Big\{ & R_1(L_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t}, \Omega_{s,f,t}) \\ & + R_2(L_{2,f,t}, L_{1,f,t}, e^*_{2,f,t}, \tilde{e}_{f,t}, \Omega_{s,f,t}) - W_{1,f,t}L_{1,f,t} \\ & - \rho W_{2,f,t+1}(1-\tau_{f,t})G^{alt}_{2,f}(\bar{U}_{2,f,t+1})L_{1,f,t} + \rho \mathbb{E}\left[V\left(L_{2,f,t+1}, \Omega_{s,f,t+1}\right)|\Omega_{s,f,t}\right] \Big\}, \end{aligned} \tag{OA.55}$$

subject to (7); with $L_{1,f,t}$ and $L_{2,f,t+1}$ given by (8) and (9), respectively; and with the expectation in the last term defined over the future productivity state, $\Omega_{s,f,t+1}$. As made explicit in (OA.55), $\Omega_{s,f,t}$ enters as an additional state variable, so the optimal compensation–promotion policy offered by each firm depends on $\Omega_{s,f,t}$ in addition to $L_{2,f,t}$.

The first-order conditions associated with (OA.55) are:

$$\begin{aligned} W_{1,f,t}: \ & \frac{\partial \bar{U}_{1,f,t}}{\partial W_{1,f,t}} g^{alt}_{1,f}(\bar{U}_{1,f,t})\mathcal{L}\Big\{ MRPL_{1,f,t} - W_{1,f,t} - \rho W_{2,f,t+1}(1-\tau_{f,t})B_{2,f,t+1} \\ & + \rho(1-\tau_{f,t})B_{2,f,t+1}\mathbb{E}\left[V'\Big((1-\tau_{f,t})B_{2,f,t+1}B_{1,f,t}, \Omega_{s,f,t+1}\Big)|\Omega_{s,f,t}\right] \Big\} - B_{1,f,t} = 0, \end{aligned} \tag{OA.56}$$

$$W_{2,f,t}: \frac{\partial \bar{U}_{1,f,t}}{\partial W_{2,f,t+1}} g_{1,f}^{alt}(\bar{U}_{1,f,t}) \mathcal{L} \Bigg\{ MRPL_{1,f,t} - W_{1,f,t} - \rho W_{2,f,t+1}(1-\tau_{f,t}) B_{2,f,t+1}$$
$$+ \rho(1-\tau_{f,t}) B_{2,f,t+1} \, \mathbb{E}\left[V'\big((1-\tau_{f,t}) B_{2,f,t+1} B_{1,f,t}, \Omega_{s,f,t+1}\big) | \Omega_{s,f,t} \right] \Bigg\}$$
$$- \rho(1-\tau_{f,t}) B_{1,f,t} B_{2,f,t+1} + \rho(1-\tau_{f,t}) B_{1,f,t} \frac{\partial \bar{U}_{2,f,t+1}}{\partial W_{2,f,t}} g_{2,f}^{alt}(\bar{U}_{2,f,t+1}) \Bigg\{ - W_{2,f,t+1}$$
$$+ \mathbb{E}\left[V'\big((1-\tau_{f,t}) B_{2,f,t+1} B_{1,f,t}, \Omega_{s,f,t+1}\big) | \Omega_{s,f,t} \right] \Bigg\}$$
$$+ \frac{1+\eta}{\eta} \frac{\partial \tilde{e}_{f,t}}{\partial W_{2,f,t+1}} \Bigg\{ d \left[F_1(B_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t}) \right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial \tilde{e}_{f,t}}$$
$$+ d \left[F_2(L_{2,f,t}, B_{1,f,t}, e^*_{2,f,t}, \tilde{e}_{f,t}) \right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial \tilde{e}_{f,t}} \Bigg\} = 0,$$

$$\tau_{f,t}: \frac{\partial \bar{U}_{1,f,t}}{\partial \tau_{f,t}} g_{1,f}^{alt}(\bar{U}_{1,f,t}) \mathcal{L} \Bigg\{ MRPL_{1,f,t} - W_{1,f,t} - \rho W_{2,f,t+1}(1-\tau_{f,t}) B_{2,f,t+1}$$
$$+ \rho(1-\tau_{f,t}) B_{2,f,t+1} \, \mathbb{E}\left[V'\big((1-\tau_{f,t}) B_{2,f,t+1} B_{1,f,t}, \Omega_{s,f,t+1}\big) | \Omega_{s,f,t} \right] \Bigg\}$$
$$+ \rho W_{2,f,t+1} B_{2,f,t+1} B_{1,f,t} - \rho B_{2,f,t+1} B_{1,f,t} \mathbb{E}\left[V'\big((1-\tau_{f,t}) B_{2,f,t+1} B_{1,f,t}, \Omega_{s,f,t+1}\big) | \Omega_{s,f,t} \right]$$
$$+ \frac{1+\eta}{\eta} \frac{\partial \tilde{e}_{f,t}}{\partial \tau_{f,t}} \Bigg\{ d \left[F_1(B_{1,f,t}, L_{2,f,t}, \tilde{e}_{f,t}, e^*_{2,f,t}) \right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial \tilde{e}_{f,t}}$$
$$+ d \left[F_2(L_{2,f,t}, B_{1,f,t}, e^*_{2_{f,t}}, \tilde{e}_{f,t}) \right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial \tilde{e}_{f,t}} \Bigg\} = 0.$$

These first-order conditions differ from those in Appendix A.1 only in that the terms associated with the firm's continuation value are taken in expectation over future productivity draws.

For completeness, we restate the relevant definitions adapted to the stochastic environment: define the marginal revenue product of level-s workers at firm f in period t as

$$\text{MRPL}_{s,f,t} = \frac{1+\eta}{\eta} d \left\{ Q_{1,f,t} \frac{\partial F_1}{\partial L_{s,f,t}} + Q_{2,f,t} \frac{\partial F_2}{\partial L_{s,f,t}} \right\},$$

and the expected net-present-discounted value of the marginal product of a new hire by firm f in period t as

$$\text{NPV}_{f,t}^{\text{MRPL}} = \text{MRPL}_{1,f,t} + \rho(1-\tau_{f,t}) G_{2,f}^{alt}(\bar{U}_{2,f,t+1}) \, \mathbb{E}\left[\text{MRPL}_{2,f,t+1} | \Omega_{s,f,t} \right],$$

where, as in (OA.55), the expectation in the last term is over $\Omega_{s,f,t+1}$. Then, one can follow the same steps used to prove Propositions 1 and 2 to verify that the results extend immediately to the case with random productivity shocks.

F Extending the Model to Allow for 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.

F.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 L_t^{lat} denote the mass of senior workers hired laterally in period t. The total number of senior workers employed by the firm in period t, denoted by $L_{2,t}^{tot}$, is therefore

$$L_{2,t}^{tot} = L_{2,t} + L_t^{lat}.$$

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 $\{W_{1,t}, W_{2,t+1}, \tau_t\}$.

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 (4).

F.2 Worker and Firm Behavior

Junior workers' effort choices, given the compensation–promotion policy offered by the firm, continue to be characterized by (7). 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 utilities in period t, net of non-pecuniary components, are

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

for promoted and laterally hired seniors, respectively.

The number of junior workers hired by the firm remains given by (8). Substituting for promotions and lateral hiring, the total number of senior workers employed in period $t+1$ is

$$L_{2,t+1}^{tot} = (1-\tau_t) G_2^{alt}(\bar{U}_{2,t+1})\, L_{1,t} + G_2^{alt}(\bar{U}_{lat,t+1})\, \mathcal{H}.$$

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

$$\begin{aligned} V(L_{2,t}) = \max_{W_{1,t}, W_{2,t+1}, \tau_t, W_{lat,t}} \Big\{ & R_1\left(L_{1,t}, L_{2,t} + L_t^{lat}, \tilde{e}_t, e_2^*\right) + R_2\left(L_{2,t} + L_t^{lat}, L_{1,t}, e_2^*, \tilde{e}_t\right) \\ & - W_{1,t} L_{1,t} - \rho W_{2,t+1}(1-\tau_t) G_2^{alt}(\bar{U}_{2,t+1}) L_{1,t} - W_{lat,t} L_t^{lat} + \rho V\left[(1-\tau_t) G_2^{alt}(\bar{U}_{2,t+1}) L_{1,t}\right] \Big\} \end{aligned} \tag{OA.57}$$

with $\tilde{e}_t$ and $L_{1,t}$ given by (7) and (8), respectively, and

$$L_t^{lat} = G_2^{alt}(\bar{U}_{lat,t})\, \mathcal{H}.$$

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

$$V'(L_{2,t}) = \frac{1+\eta}{\eta} d \left\{ \left[F_1(L_{1,t}, L_{2,t} + L_t^{lat}, \tilde{e}_t, e_2^*)\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_{2,t}} + \left[F_2(L_{1,t}, L_{2,t} + L_t^{lat}, e_2^*, \tilde{e}_t)\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_{2,t}} \right\} \tag{OA.58}$$

analogous to the baseline model in Section 3.

F.3 Testable Implications

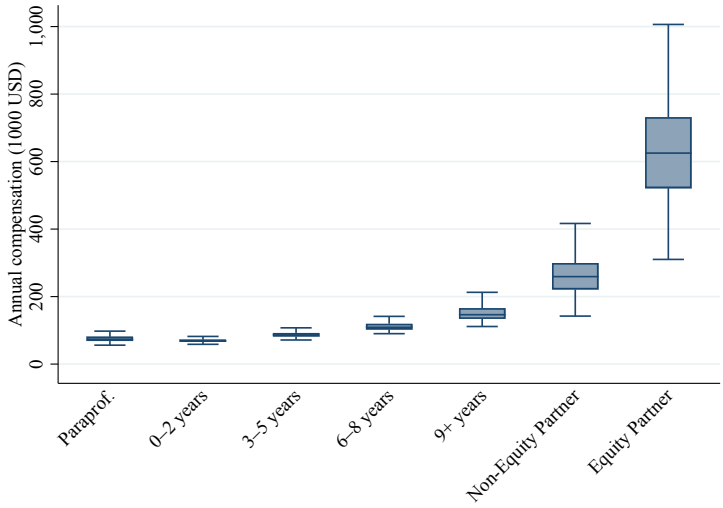
The equilibrium conditions mirror those in Section 3. Define

$$\begin{aligned} \text{MRPL}_{1,f,t} \equiv \frac{1+\eta}{\eta} d \Bigg\{ & \left[F_1(L_{1,f,t}, L_{2,f,t}^{tot}, \tilde{e}_{f,t}, e_{2,f,t}^*)\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_{1,f,t}} \\ & + \left[F_2(L_{2,f,t}^{tot}, L_{1,f,t}, e_{2,f,t}^*, \tilde{e}_{f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_{1,f,t}} \Bigg\}, \\ \text{MRPL}_{2,f,t} \equiv \frac{1+\eta}{\eta} d \Bigg\{ & \left[F_1(L_{1,f,t}, L_{2,f,t}^{tot}, \tilde{e}_{f,t}, e_{2,f,t}^*)\right]^{\frac{1}{\eta}} \frac{\partial F_1}{\partial L_{2,f,t}^{tot}} \\ & + \left[F_2(L_{2,f,t}^{tot}, L_{1,f,t}, e_{2,f,t}^*, \tilde{e}_{f,t})\right]^{\frac{1}{\eta}} \frac{\partial F_2}{\partial L_{2,f,t}^{tot}} \Bigg\}. \end{aligned}$$

One can verify that the first-order conditions associated with (OA.57) coincide with those in Section A.1. Propositions 1 and 2 therefore continue to hold in the presence of lateral hiring.

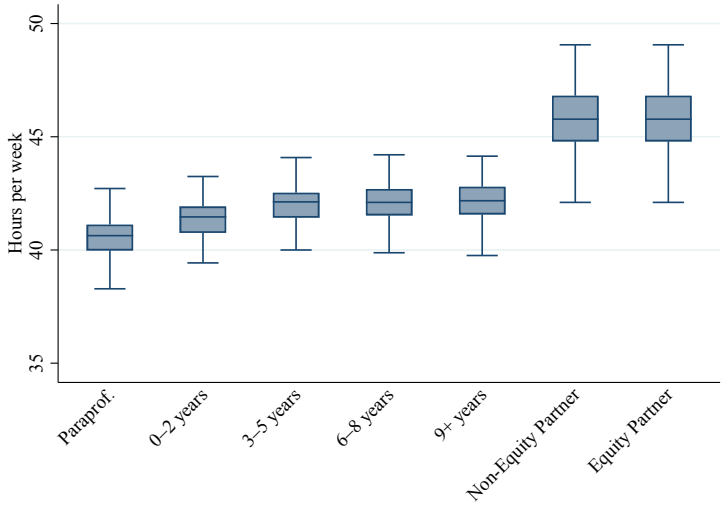
G Additional Figures

Figure OA-9: Annual Compensation



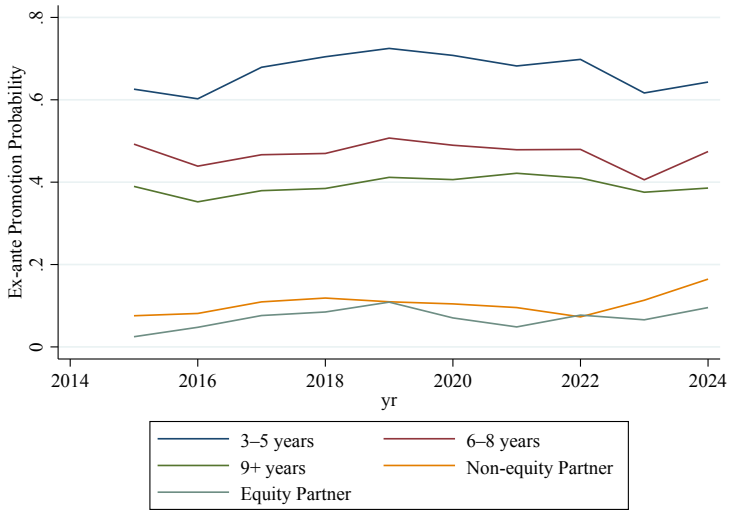
Notes: Boxplots of annual compensation by seniority level, in 1000 USD.

Figure OA-10: Recorded Hours per Week



Notes: Boxplots of hours per week by seniority level. Assumes 49 weeks worked per year, based on an average of 3 weeks of PTO per year (Inside Public Accounting, 2023).

Figure OA-11: Average Employment Shares, by Year



Notes: Average number of employees per seniority level divided by total partnership-track employees at the firm.

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.