UNIVERSITY OF CALIFORNIA

Los Angeles

Markets and the Internal Organization of Firms

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Economics

by

Jacob Matthew Kohlhepp

2023

ABSTRACT OF THE DISSERTATION

Markets and the Internal Organization of Firms

by

Jacob Matthew Kohlhepp
Doctor of Philosophy in Economics
University of California, Los Angeles, 2023
Professor Simon Adrian Board, Co-Chair
Professor Maurizio Mazzocco, Co-Chair

All three chapters of this dissertation study different aspects of internal organization, and two of the chapters go a step further to understand how decisions within the firm impact markets outside the firm.

The first chapter of this dissertation studies empirically how worker task assignments within firms interact with labor and product markets. Using data from a software company, I observe millions of task assignments within hundreds of hair salons, many of which are competitors. I develop a measure of organization complexity, which is the amount of information required to implement a given task assignment, to provide evidence of firm-specific organization costs, which grant complex salons a comparative advantage in producing high-quality products.

Based on these facts, I develop a model where oligopolistic firms with different organization costs choose their internal structure. Complexity is costly, but it allows firms to improve product quality by better matching workers with multidimensional skills to tasks. I characterize the profit-maximizing organization, and use results from the literature on rational inattention and information theory to identify and estimate the model for Manhattan hair salons. Counterfactuals reveal that allowing internal organization to be

heterogeneous and endogenous changes the equilibrium effects of policy. A sales tax cut increases specialization and therefore the productivity of all workers, while a minimum wage increase generates new types of wage spillovers.

The second chapter of this dissertation studies empirically how voluntary labor supply decisions within an organization impact workplace injury rates using novel data on the payroll and workers' compensation claims of Los Angeles traffic officers. I use the leave taken by coworkers as an instrument to estimate the causal effect of daily labor supply decisions on workplace injury. Self-selection via voluntary labor supply reduces injuries by 48 percent compared to the underlying injury rate. The majority of the effect is driven by private factors, implying decentralized overtime assignment mechanisms like shift auctions can be used to reduce injury rates.

The third chapter of this dissertation (joint with Stepan Aleksenko) studies theoretically how the use of recruiters impacts the types of workers hired. The chapter considers a model where a firm delegates search for a worker to a recruiter. Productivity is uncertain prior to hire with recruiter beliefs characterized by an expectation and variance. Delegation occurs using a refund contract which is common in the industry. We analyze how delegation in this setting shapes search behavior and the composition of hires. We demonstrate that delegation is theoretically equivalent to making the search technology less accurate. This generates inefficiency: search effort and social surplus are lower under delegation than in the first-best benchmark. We show this inefficiency is driven by moral hazard with a multitasking flavor. The recruiter wastes search effort finding low variance workers at the expense of high expectation workers. As a result, as workers become more homogeneous with respect to productivity variance, delegation becomes more efficient. Our model provides a microfoundation for variance-based statistical discrimination.

The dissertation of Jacob Matthew Kohlhepp is approved.

Till von Wachter

Moritz Meyer-ter-Vehn

Daniel Haanwinckel Junqueira

Maurizio Mazzocco, Committee Co-Chair

Simon Adrian Board, Committee Co-Chair

University of California, Los Angeles

2023

To my wife and daughter.

You are God's greatest gift to me.

You encourage me to new heights...

and keep me grounded when I reach them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I acknowledge the guidance and support of my advisors, in particular the co-chairs of my committee Maurizio Mazzocco and Simon Board who helped develop me into an economist over the last five years. I thank Stepan Aleksenko, my friend and main co-author for his brilliance, commitment to excellence on all of our projects and emergency assistance when I broke my leg. I am grateful for the support and advice of a community of economists both at UCLA and beyond, in particular Tara Sinclair and Tyler Ransom, who went out of their way to support a graduate student from another university.

I thank my parents, Robert and Laurie Kohlhepp, and my wife's parents, Francis and Sherry Wang, for their support and love as I pursued a P.h.D. I am specifically thankful for their help caring for my daughter Rosy. Without their help, this dissertation could not have been written.

I thank the UCLA Economics Department, for over a decade of support from undergraduate (2012) through graduate school (2023). I am grateful for financial support from the Graduate Research Mentorship Program and the Dissertation Year Fellowship. I thank the Institute for Humane Studies, for both financial support and guidance during my time as a graduate student.

VITA

2016	B.A. (Economics) and B.A. (Political Science), UCLA.
2016–2018	Associate Economist, Welch Consulting, Los Angeles, California.
2020	M.A. (Economics), UCLA, Los Angeles, California.
2019–2022	Teaching Assistant and Instructor, UCLA, Los Angeles, California.
2020–2021	Consultant, Boulevard Salon Management, Los Angeles, California.
2022–2023	Dissertation Year Fellowship, UCLA, Los Angeles, California.

CHAPTER 1

The Inner Beauty of Firms

> "Of all the things I've done, the most vital is coordinating those who work with me and aiming their efforts at a certain goal." - **Walt Disney**

1.1 Introduction

Greater specialization allows markets to better use the unique talents of individuals. As early as Adam Smith's pin factory, economists have recognized that much of this division of labor occurs within the firm, a process often referred to as internal organization. In practice, firms differ in their ability to organize people and use a wide variety of organization structures. How do firms choose their internal organization, and how does this choice interact with product markets, labor markets and government policy?

To answer this question, I propose a framework to study firms' equilibrium choice of internal organization. Using a set of stylized facts from management software data, I model firms as deciding which workers to hire and how to assign them to tasks. More complex assignments are costly, but they improve product quality through a better match of skills to tasks. Because firms differ in their organizational capabilities, they choose different internal structures. Additionally, because firms share a product and a labor market, the internal organization structures of competing firms are intertwined in equilibrium. I estimate the model for Manhattan hair salons, and show that allowing internal organization to be heterogeneous and endogenous qualitatively changes the effect of counterfactual policies. For example, a minimum wage raises equilibrium specialization for minimum wage workers, reduces specialization for non-minimum wage workers, and causes wage spillovers which are not monotone in initial wage.

In the first part of this paper, I use novel data to establish empirical patterns in firm internal organization. The data, which come from a management software company, allow me to observe the assignment of millions of tasks to individual workers across hundreds of hair salons. I view firms as choosing organization structures, which are matrices where rows represent workers, columns represent tasks, and each element is the fraction of total time assigned to each worker-task pair. I create a measure called organization complexity, which quantifies the amount of information that must be communicated within a firm in order to implement a given organization structure.

I document three facts about salon internal organization. First, complexity varies significantly across salons but very little across time, with few salons engaging in complete specialization. This is evidence of firm-specific and time-invariant organization costs which prevent full specialization. Second, complex firms have higher revenue and employment. This indicates firms with lower organizational costs have a competitive advantage in the product market. Third, complex firms have higher prices and more repeat customers. This is evidence the organizational competitive advantage operates through quality rather than quantity, meaning organizationally efficient salons have a comparative advantage at producing higher-quality products.

In the second part of the paper, I build a model consistent with these facts. In this model, firms with product market power choose product prices, the composition of their workforce, and worker task assignments. Workers differ in their skill at each task. Assigning tasks to the most skilled worker raises product quality but also increases organization complexity. Firms differ in the cost of complexity and their task-based production function, which causes them to choose different internal structures. Firms compete in a common product and labor market, so their choices of internal structures both shape and are shaped by wages, prices and qualities.

The main theoretical result in this paper is a characterization of the firm's optimal organization structure enabling analysis, identification and estimation. My model differs from past task-assignment models along three dimensions: firms face heterogeneous organization costs which prevent full specialization, firms have market power, and workers have horizontal skill heterogeneity.[1] Because of these differences, I cannot use existing approaches to make the firm's problem tractable. Instead, I show that the profit-maximizing organization also solves an equivalent rational inattention problem with mutual information attention costs. This equivalence allows me to weave together existing results to prove the other propositions in the paper.

Using this model, I analyze the theoretical forces that shape a firm's choice of internal structure. I show that firms navigate a trade-off between organization complexity, wage and quality, where they attempt to produce the highest-quality product using the simplest task assignment and the lowest wage bill. I prove that even though firms are choosing the task assignment of each individual worker, at a high level, the firm is choosing a point along a convex frontier that divides two dimensions: organization complexity and wages adjusted for product quality. The firm chooses the point along the frontier that is tangent to its isoprofit curves, which I show are straight lines with a slope determined by the firm's organization cost parameter.

In the third part of the paper, I identify and estimate a structural version of the model for hair salons in New York City. The distribution of organization costs is identified by the complexity of a firm's task assignments. Further, organization costs and structures are known functions of the data and the other parameters, and do not need to be estimated.[2] Variation in the interaction of task intensity and organization complexity across firms in the same market allows the identification of the other parameters. Intuitively, firms intense in task k and organizationally complex hire a large share of task k specialists and assign a large amount of task k to these specialists. The quality of these firms identifies the skill of task k specialists, while the cost of these firms identifies the wage of task k specialists. I provide a computationally light, nested fixed-point estimation procedure

1. In the beauty industry horizontal skill heterogeneity is important: one worker may be skilled at cutting and unskilled at coloring while another may be skilled at coloring and unskilled at cutting.

2. This is because complexity reveals organization costs, by the logic in the last paragraph.

which implements this identification strategy.

The estimated model reveals that even within a single industry (hair salons) and occupation (cosmetologists), variation in task specialization is large and depends on unobserved worker skills and unobserved firm organizational differences. Firms in the bottom quartile of organization costs (efficient salons) on average assign 90% of tasks to the associated specialist, while firms in the top quartile (inefficient salons), assign only 67%. Haircut specialists spend most of their time cutting, but blow-dry specialists spend less than half of their time blow-drying. I also show that internal organization is a large source of productivity differences across firms, accounting for 40% of variation in marginal costs.

In the fourth part of the paper, I study two counterfactual policy changes, one in the product market and one in the labor market. In both cases, the fact that internal organization is heterogeneous and endogenous introduces new economic forces and qualitatively changes the total economic impact of each policy. The structure of the model allows any policy to be cleanly decomposed into a reallocation effect, where labor shifts across firms but internal organization remains fixed, and a reorganization effect, where task assignments within the firm are allowed to adjust. The reallocation effect is driven by the heterogeneity of internal organization, while the reorganization effect is driven by the endogeneity of internal organization.

In the first counterfactual, I eliminate the 4.5% New York City sales tax on services. The reallocation effect improves the competitive position of complex salons who were initially providing high-quality services. The reorganization effect induces almost all salons to reorganize in order to increase quality. Both effects increase equilibrium task specialization across all workers and increase equilibrium labor productivity. Workers capture most of the productivity gains through higher wages.

In the second counterfactual, I increase the minimum wage from $15 to $20. The reallocation effect reduces the competitive position of firms with internal structures that rely on minimum wage workers. Thus, non-minimum wage workers initially employed alongside minimum wage workers see a reduction in labor demand. The reorganization effect causes firms to lay off more minimum wage workers and shift their tasks onto other workers. This increases task specialization for minimum wage workers but reduces it for other workers. Although the labor market is competitive, organizational heterogeneity and endogeneity allow the model to generate aggregate labor-labor substitution patterns that are not possible with standard models. For Manhattan hair salons, reallocation and reorganization together produce wage spillovers that are non-monotonic in initial wage, with high- and low-wage workers seeing wage increases and workers in the middle seeing wage decreases.

In this paper I draw insights from organizational economics and the task-based literature in labor economics in order to understand how internal organization decisions shape economic outcomes. The primary contribution of the paper is to build and estimate a model where organizationally unique firms make task assignment decisions which have labor and product market consequences.

The literature in organizational economics provides many ways in which firms can allocate talent better than markets do. These include monitoring (Alchian and Demsetz 1972, Baker and Hubbard 2003), relational contracts (Baker, Gibbons, and Murphy 2002), knowledge (Garicano and Wu 2012), coordination (Dessein and Santos 2006), trust (Meier, Stephenson, and Perkowski 2019) and culture (Martinez et al. 2015). Just as Holmstrom and Milgrom (1994) view the firm as an incentive system, in this paper I view the firm as a system of organizational practices. Once one adopts this view, firms should have heterogeneous organizational capabilities depending on their particular mix of practices (Argyres et al. 2012). Using firm-specific costs, I capture this heterogeneity in order to study its impact on market outcomes. I find that organizational heterogeneity is important both for determining the division of labor across the economy and for understanding the distributional impact of policy changes.

I model labor as being divisible into tasks which can be assigned to workers with different skills, a tradition that dates back to at least Sattinger (1975) but has seen growing use since Autor, Levy, and Murnane (2003). I incorporate features present in different parts of the literature, including multidimensional worker types (Lindenlaub 2017), firms with multiple worker types (Haanwinckel 2020), organization costs (Adenbaum 2021, Garicano 2000), and firm-specific task demands (Lazear 2009). I also incorporate product market power. This combination of features allows for flexible labor-labor substitution patterns that are determined by the distribution of skills, organization costs and task demands in the economy. This flexibility is why I find that a minimum wage generates non-monotonic wage spillovers even in a competitive labor market.[3] Additionally, my model generates jobs which are bundles of tasks and which vary from firm to firm even for the same type of worker. This makes my model more realistic than past models, which typically generate fully specialized jobs that are homogeneous within industry.

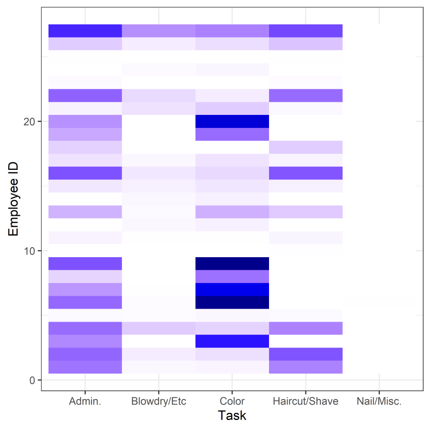
Finally, this paper makes a methodological contribution. In the majority of task-based models and hierarchy models (Garicano and Rossi-Hansberg (2006), Caliendo et al. (2012), Garicano and Hubbard 2016), workers are matched to tasks according to a single dimension, typically involving education. Since it is known prior to estimation that wages are increasing in this observed dimension, information on the wage distribution identifies features of the task-based production function. Direct information on tasks is then typically not used for estimation. I consider the opposite case, when information on wages is limited but information on tasks is rich. I show that the parameters of the task-based production function can be inferred based on differences in qualities and costs across firms intense in different tasks but operating in the same market. Further, task information allows the incorporation of workers that have unobserved horizontal differences in skill. This makes the model useful for incorporating skill differences which cannot be inferred from observed characteristics of workers.

I am able to maintain computational and analytical tractability by establishing that the profit-maximizing organization is the solution to a class of well-studied problems from the information theory and rational inattention literature. In a similar fashion, Ocampo (2022) and Adenbaum (2021) show the firm's task assignment problem is an optimal transport problem, while Freund (2022) uses the Fréchet distribution combined with re-

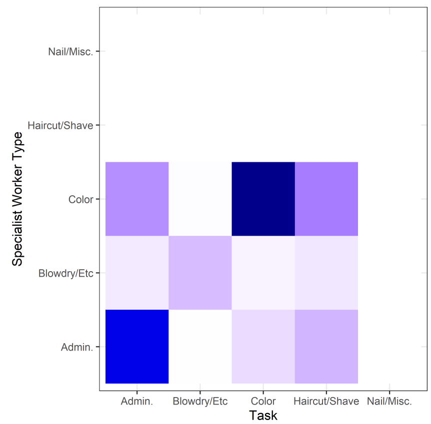
3. This is similar to how Teulings (2000) showed that imperfect substitution along a single dimension changes how we should analyze minimum wages.

sults from the trade literature. Despite this theoretical similarity, this paper is distinct in that it uses data on task assignments within firms directly for both the design and estimation of the model. An illustration of the methodological contribution is given in Figure 1.1.

Figure 1.1: Utilization of Task Assignment Data



(a) Observed



(b) Estimated

Note: Darker colors indicate a higher fraction of total work at the salon. The model in this paper takes in establishment-level data about the task assignments of employees with unknown skills (Panel A) and returns the task assignments of worker types with known skills (Panel B). Even though the displayed salon in New York City employs 26 people, the model infers these represent only three of the five worker types available in the market and that most specialization is within the color and administrative tasks.

The remainder of the paper is organized as follows. Section 2 describes the management software data. Section 3 describes three stylized facts about hair salons that are used to build the model. Section 4 specifies the theoretical model. Section 5 theoretically analyzes the model. Section 6 discusses the identification and estimation of a structural version of the model. Section 7 presents the parameter estimates and assesses the model fit. Section 8 performs two counterfactual policy experiments. Sections 9 discusses implications and Section 10 concludes.

1.2 Data

This section describes the salon management software data I use in this paper.

1.2.1 Context and Institutional Details

The data set was obtained from a data sharing agreement I negotiated with a salon management software company. The software facilitates running a beauty business, including scheduling, pricing, payments, inventory, staffing, business reporting, client profiling and marketing. As of July 19, 2022, a monthly subscription has a base price of $175. Although the company also markets its software to spas, tanning salons and massage parlors, hair salons and barbers make up the majority of its clients. For this reason, I analyze only hair salons and barbershops.

The software is sold to beauty businesses throughout the United States, but the data indicate uptake is largest in Los Angeles (where the company was founded) and New York City. An important aspect of the data set is that it allows me to observe the internal organization of salons that are geographically close and therefore likely to be direct competitors in labor and product markets. For example, I observe 10 salons in the lower Manhattan zip code 10013, which is a 0.55 square mile area.

The data document which stylist is assigned to each task and client, and record the duration of the appointment, the price paid, and a custom text description of each task. If more than one employee is assigned to a single client, this is recorded as multiple entries describing what each employee contributed. Although the data are de-identified, unique IDs allow a researcher to track employees and clients across time within a salon.[4]

A sample from the data is provided in Table 1.3, with IDs replaced with pseudonyms. This sample shows the different ways two salons coordinate employees to meet customer demand. Blake requested a cut, highlights and a treatment at salon 1A. The salon had a single employee, Rosy, perform all three services. Grace requested a cut and a single process (color) at salon 2A. Unlike salon 1A, salon 2A chose to assign each of these tasks to two separate employees, Tyler and Ben. Both of these salons are in the same zip code.

While the data are rich in terms of task content and worker assignments, informa-

4. IDs are salon specific, so I cannot track employees or clients if they move across salons.

tion about employee compensation is sparse. The software can track some compensation information (tips, commissions and employment relationship, etc.), but these additional functions are not used consistently by client salons, as discussions with the company and analysis of internal data revealed.

1.2.2 Mapping Descriptions to Tasks

The data contain 20,560 unique text descriptions of services. This section describes the process used to categorize these descriptions into five tasks.

A licensed cosmetologist was hired to group the tasks into a manageable number of categories. Appendix Section 1.11.14 describes the instructions sent to the cosmetologist and displays part of the final spreadsheet. I use the six-category grouping provided by the cosmetologist with one modification: I combine the extension task with the blow-dry task to create five final task categories, because the extension task is very sparse–for Manhattan in 2021 Q2, fewer than 10 hours were dedicated to this task. This sparsity leads to estimation problems, as parameters tied to this task have a negligible effect on observable outcomes.

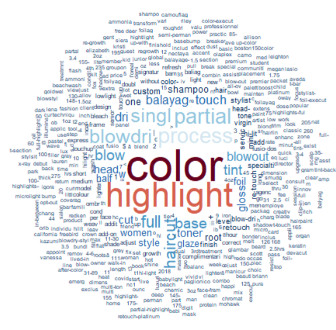
If a service is marked as multiple task categories, I divide the service into unique tasks in the following way. First, I compute the average amount of time spent on each task among services that are marked only as one task. Second, I compute the fraction of time to assign to each task as the corresponding task average divided by the sum of the averages of all other tasks marked for that service. Third, I distribute the total time spent on the service across the tasks using this imputed fraction. This process generates task categories that are mutually exclusive.

1.2.3 Descriptive Statistics

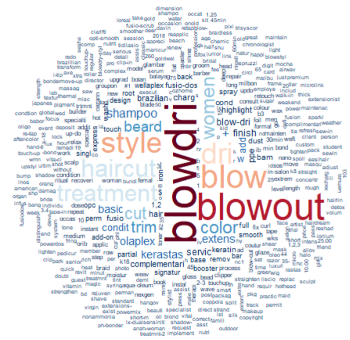
The data used in this section and the stylized facts include all observed firm-quarters where revenue per customer is positive. I exclude 2021 Q3, because I observe only part of the quarter. I also exclude an establishment in Kentucky with revenue that is implausibly

Figure 1.2: Task Categories

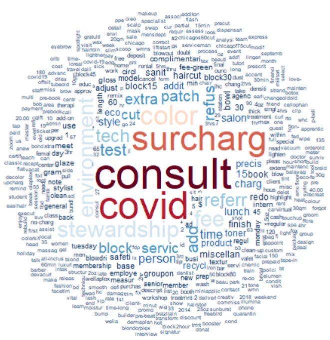
(a) "Haircut/Shave:" 38%



(b) "Color/Etc.:" 41%



(c) "Blowdry/Etc.:" 8.9%



(d) "Administrative:" 6.1%

(e) "Nail/Misc.:" 6.3%

Note: The 20,560 service descriptions grouped into task categories by a cosmetologist. Word size is proportional to the number of tasks which include the word.

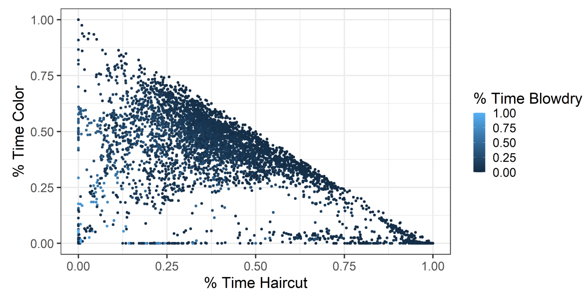
high. The data contain information on 445 hair salon establishments, which represent 316 unique businesses, 9,179 hair stylists, 1,654,233 customers and 10.8 million services performed. Establishments first appear in the data when they adopt the management software. The last complete month with available data is August 2021. Although the software is used by salons across the country, users are concentrated in New York and California.

I aggregate the data to the firm-quarter level, for analysis. Descriptive statistics at this level are provided in Table 1.4. Throughout the paper, I refer to the price as the average revenue per customer per quarter. The salons have an average price of $200. Even though there is significant variation in the relative intensity of tasks at different salons, most salons offer at least four of the five task categories in a given quarter. Throughout the paper, I refer to the task mix of a salon as the fraction of total time spent on each of the five tasks. Firm-quarter heterogeneity in the task mix is illustrated in Figure 1.3. Firms vary in their intensity in each task.[5]

Figure 1.3: Variation in Firm-Quarter Task Mix

Share of Labor	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Haircut/Shave	4,558	0.41	0.23	0.00	0.26	0.52	1.00
Color/Highlight/Wash	4,558	0.38	0.20	0.00	0.29	0.52	1.00
Blowdry/Etc	4,558	0.09	0.12	0.00	0.03	0.11	1.00
Administrative	4,558	0.05	0.11	0.00	0.002	0.04	1.00
Nail/Etc	4,558	0.06	0.16	0.00	0.00	0.05	1.00

(a) Summary Statistics



(b) Variation in 3 Main Tasks

Note: Panel A provides summary statistics about the share of time spent on each task across all firm-quarters. Panel B illustrates this variation for the three most common tasks. Each point is a firm-quarter.

The salons in the sample have an average quarterly revenue of \$213,201 and an average of 13 employees. Johnson and Lipsitz (2022) studies a sample of salon owners and reports an average annual (not quarterly) revenue of \$233,000 and an average of seven stylists. It is important to be cautious when comparing self-reported survey estimates from other sources with management data (like this source), but given the subscription fee of the software, it is reasonable to conclude that the salons in my sample skew toward larger and higher-end salons. This suggests the heterogeneity found in this paper underestimates the heterogeneity in the universe of U.S. salons.

1.3 Stylized Facts

The model I use to study the effect of internal organization on product and labor markets is inspired by three stylized facts. These facts require the definition of two concepts which will be used throughout the paper. To begin, denote workers by the index i, firms by the index j, and tasks by the index k.

5. To see variation across jobs, refer to Table 1.15 and Figure 1.32.

Definition 1 *A firm's organization structure, denoted by B_j, is a matrix where element $B_j(i,k)$ is the fraction of labor assigned to worker i and task k.*

An example of two different organizational structures is given in Figure 1.4. The left structure is staffed by specialists while the right structure is staffed by generalists.

The column totals represent the firm's task mix, that is, the amount of each task needed to produce one unit of output. This mix is assumed to be exogenous, determined either by technology constraints or demand. The row margins represent the composition of a firm's workforce: the fraction of total work that is assigned to each worker. The two panels provide two different organization structures which have the same task mix (column totals) but which assign work very differently. In such a structure, workers are exchangeable, and roles are not distinct.

Figure 1.4: Two Organization Structures

Specialist Salon

Employee	Tasks: 1	2	3	
A	1/2	0	0	1/2
B	0	1/4	0	1/4
C	0	0	1/4	1/4
Tot.	1/2	1/4	1/4	1

Generalist Salon

Employee	Tasks: 1	2	3	
A	1/6	1/12	1/12	1/3
B	1/6	1/12	1/12	1/3
C	1/6	1/12	1/12	1/3
Tot.	1/2	1/4	1/4	1

Note: Two organizational structures a firm with a task mix of $1/2, 1/4, 1/4$ could choose. Column sums represent the task mix, and row sums represent the fraction of work performed by each employee.

The second concept, complexity, measures the minimum amount of information that must flow through the firm in order for it to implement a given structure, and it is based on a literature in information theory starting with Shannon (1948).[6]

6. It is the mutual information of the joint distribution over workers and tasks. Mutual information is a well-understood way to measure information costs. It has many desirable properties and many microfoundations.

Definition 2 *The complexity of an organization structure B_j is*[7]

$$I(B_j) = \sum_{i,k} B_j(i,k) log\left(\frac{B_j(i,k)}{\underbrace{\sum_{k'} B_j(i,k')}_{\text{Worker } i \text{ Labor Share}} \underbrace{\sum_{i'} B_j(i',k)}_{\text{Task } k \text{ Labor Share}}} \right)$$

Consider the two structures in Figure 1.4. The firm can implement the chair-renter structure (right) by randomly assigning workers to tasks. This implementation does not require information about tasks or worker identities, so the complexity is 0 in this case. To implement the employee structure (left), the firm must tell each worker exactly which task to perform. The firm can write an employee manual, stating "assign the task to employee A if you observe '0', assign to B if you observe '01', and assign to C if you observe '10'." The expected number of bits (or amount of information) is $1 \times 1/2 + 2 \times 1/2 = 1.5$. This is the minimum information required to communicate this assignment, so the complexity in this case is 1.5.

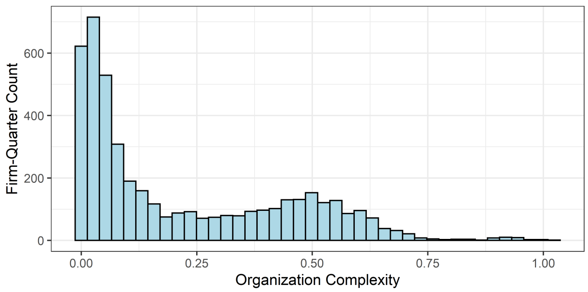
I now present three stylized facts about internal organization. Throughout the rest of the paper, complexity is assumed to be measured without error. Appendix Section 1.11.16 provides evidence that measurement error is small.

Fact 1 *Complexity varies significantly across firms and little across time.*

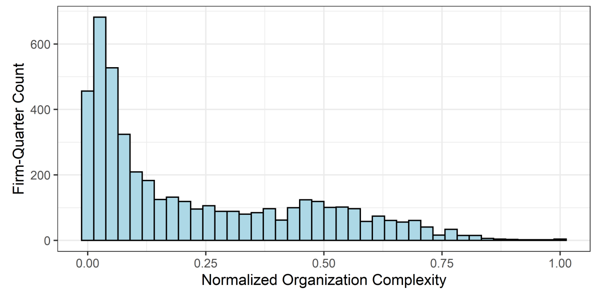
To establish this fact, I first compute I_j^{max}, which is the maximum value of complexity given a firm's task mix in a given quarter. I construct normalized complexity $\bar{I}_j$ as raw complexity divided by I_j^{max}. Normalized complexity $\bar{I}_j$ has a minimum of 0 (like raw complexity) and a maximum of 1 (unlike raw complexity). I plot a histogram of normalized complexity in Figure 1.5 and observe that complexity varies significantly across firm-quarters and has a long right tail. In particular, I observe that while some firms have very complex organizations (close to the upper bound), others have very simple organizations (complexity of 0). To understand whether the variation is across time or across salons, I

7. When computing this measure, I assume that $0log(0) = 0$.

Figure 1.5: Histogram of Normalized Complexity



(a) Raw



(b) Normalized

Note: Includes all firm-quarter observations. Both normalized and raw complexity vary significantly. Normalized complexity is observed achieving both its lower bound (0) and upper bound (1).

decompose complexity into a salon-specific component, a time-specific component and a residual component:

$$\bar{I}_{j,t} = \bar{I}_j + \bar{I}_t + e_{j,t}$$

I estimate the firm and year components by regressing normalized complexity on time and salon fixed effects. This allows me to decompose the total variance of complexity into the three components:

$$\underset{.0516}{Var(I_{j,t})} = \underset{.0464}{Var(\bar{I}_j)} + \underset{.0002}{Var(\bar{I}_t)} + \underset{-.0009}{2Cov(\bar{I}_j, \bar{I}_t)} + \underset{0.0059}{Var(e_{j,t})}$$

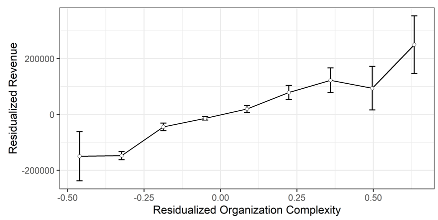
These results demonstrate that 90 percent of the variance in normalized complexity is attributable to the firm component and only 0.4 percent to the time component. Therefore, complexity varies significantly across firms but little across time. This is evidence the choice of complexity is driven by a time-invariant, firm-specific organization cost.

Fact 2 *Complex firms have higher revenue and employment.*

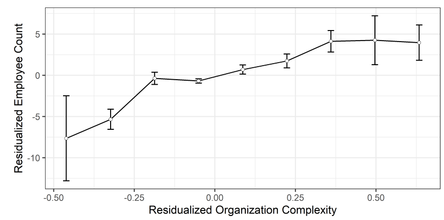
Complexity is positively correlated with revenue and employment, as well as several other measures of firm size. This correlation is depicted in Figure 1.6, which shows binned scatter plots of residualized complexity against residualized revenue employment, customers and visits. The plots control for the task mix, county, and quarter fixed effects.

Table 1.2 demonstrates via a series of regressions that the correlation is positive for all firm size variables and statistically significant at the 5 percent level for revenue and employment. The positive relationship between revenue and complexity is robust; it remains when only Manhattan hair salons are analyzed and when employee count is interacted with complexity.[8]

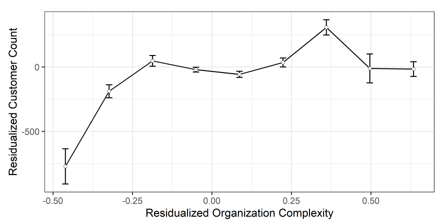
Figure 1.6: Organization Complexity and Firm Size



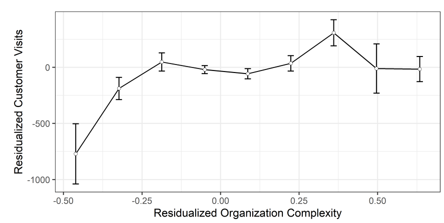
(a) Revenue ($)



(b) Employees



(c) Unique Customers



(d) Customer Visits

Note: Each panel illustrates the positive relationship between organization complexity and a different measure of firm size. All variables are residualized for quarter, county and task mix. Firm-quarters are grouped into equally spaced bins based on complexity.

The positive correlation between firm size and complexity suggests some salons have an organizational competitive advantage in that they find it easier than competitors to adopt productive organizational practices. This allows them to implement more complex task assignments at a lower cost.

Fact 3 *Complex firms have higher prices and more repeat customers.*

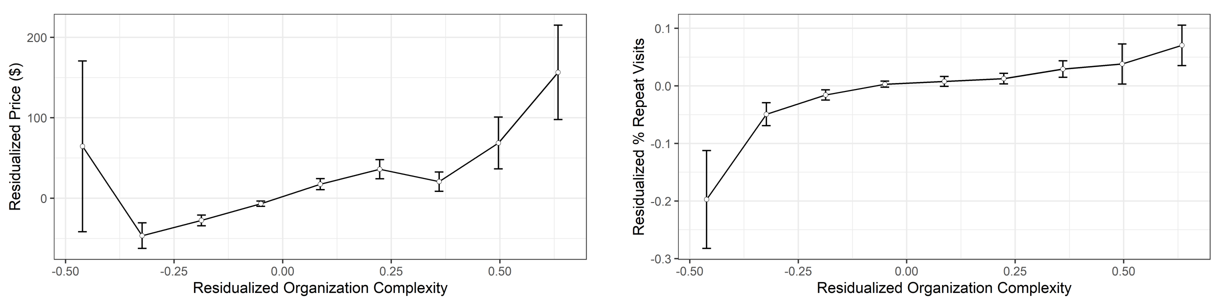
Complexity is positively correlated with price, as shown in Panel A of Figure 1.7. Appendix Section 1.11.12.2 proves that this pattern in the data is inconsistent with a model

8. See Appendix Section 1.11.17 for details on these additional results.

where organizational competitive advantages operate only through marginal cost reductions. In such a model, prices should be decreasing in complexity. The fact that the opposite is true suggests salons with higher internal complexity are producing services with higher unobserved quality and thus higher costs.[9]

To test this quality channel, I use the share of repeat visits as a proxy for quality in Panel B of Figure 1.7. It is reasonable to assume that a customer who returns was satisfied with the quality of the original service. The fraction of visits by return customers rises with complexity, evidence of a link between quality and organization. This suggests that the organizational advantage described in Fact 2 operates through unobserved quality rather than quantity. In the next section, I build a model inspired by this and the other two facts. Appendix Section 1.11.15 discusses the robustness of the stylized facts.

Figure 1.7: Organization Complexity, Prices and Repeat Customers



(a) Price

(b) Repeat Customers

Note: The positive relationship between organization complexity and price (panel A), and the relationship between organization complexity and the fraction of customers that return (panel B). All variables are residualized for quarter, county and task mix. Firm-quarters are grouped into equally spaced bins based on complexity.

1.4 Model

This section specifies a model where firms choose prices and organizational structures simultaneously in order to compete for consumers. Consistent with the stylized facts, firms choose their organization structure subject to heterogeneous organization costs. The

9. Kugler and Verhoogen (2012) use a similar argument to conclude that endogenous product quality is important.

main benefit of a complex organization is the ability to produce a higher-quality product. There are three important groups of objects in the economy: firms, indexed by $j = 1, ..., J$; worker types, indexed by $i = 1, ..., N$; and tasks, indexed by $k = 1, ..., K$.

Firms and Tasks. The J firms differ in their organization cost $\gamma_j \in \mathbb{R}_+$, discussed below. Each firm produces a single good[10] using a Leontief task-based production function described by $\alpha \in \mathbb{R}^K_+$, which I refer to throughout the paper as the task mix. The task mix is homogeneous in the theoretical section only for exposition: all results are obtained when it varies by firm. To produce one unit of the good, the firm must allocate α_k labor to task k, where I normalize $\sum_k \alpha_k = 1$. The firm can choose how these tasks are assigned to workers in a process that is described shortly.

Workers and Labor Markets. Each of the N worker types is characterized by inelastic total labor supply L_i and skill set vector θ_i. Element $\theta_i(k)$ is the quality with which worker i performs task k. The labor market is competitive with type-specific wages w_i, which I collect into a wage vector w.

Firm Strategies. Firms choose the price of their product $p_j \in \mathbb{R}_+$ and their organizational structure $B_j \in \Delta^{N\times K}$, where $\Delta^{N\times K}$ is a $N \times K$-dimensional unit simplex. Element $B_j(i, k)$ of an organization structure specifies the fraction of total labor allocated to worker type i and task k.[11] An organizational structure B_j is feasible if it is consistent with the task-mix vector: $\sum_i B_j(i, k) = \alpha_k \,\forall\, k$. The workforce composition, $E_j(i) = \{E_j(1), ..., E_j(N)\}$, is the

10. Considering only a single good allows me to focus on internal organization. Nocke and Schutz (2018) shows we can represent a pricing game with multi-product firms as one with single-product firms by adjusting qualities and costs when demand takes the multinomial logit form.

11. This definition of task assignment treats all workers with a given set of skills symmetrically. In the model brought to the data, workers with the same skills may have different labor supplies. I show in Appendix Section 1.11.6 that, due to an invariance property of the organization cost function, even if a firm could treat different workers with the same skills differently, it would not choose to do so in equilibrium. Thus this abstraction is without loss of generality under mutual information organization costs.

fraction of total labor demanded that is from each worker type. By definition,

$$E_j(i) = \sum_k B_j(i,k). \tag{1.1}$$

The cost of a firm's organization structure is the firm-specific parameter γ_j multiplied by the complexity of the organization structure $I(B_j)$. Recall complexity is defined as[12]

$$I(B_j) = \sum_{i,k} B_j(i,k) log\left(\frac{B_j(i,k)}{\underbrace{\sum_{k'} B_j(i,k')}_{\text{Type } i \text{ Labor Share, } E_i} \underbrace{\sum_{i'} B_j(i',k)}_{\text{Task-mix, } \alpha_k}}\right).$$

A firm's organizational structure determines the match between worker skills and tasks. As a result, it determines product quality ($\xi(B_j)$). I specify that product quality is a weighted average of task quality:

$$\xi(B_j) = \sum_{i,k} B_j(i,k)\theta_i(k).$$

Since quality is valued by consumers, increased quality is the main benefit of carefully assigning workers to tasks. A firm's organization structure also determines its per-unit wage bill:

$$W(B_j) = \sum_{i,k} w_i B_j(i,k)$$

Demand. Total market demand for good j is given by a function D_j which maps the prices and qualities of all firms into a quantity demanded for firm j. I assume that demand for good j depends on own-price and own-quality only through the quality-price index $\xi(B_j) - \rho p_j$, where ρ is a consumer price sensitivity parameter. I also assume demand for good j is strictly increasing in good j's quality-price index. This implies the

12. This mutual information-based functional form is used because it is the only cost function in a certain class where complexity over types will be equal to complexity over worker identities under a general matching process (Bloedel and Zhong 2021).

demand can be written as $D_j(\xi(B_j) - \rho p_j, p_{-j}, \xi_{-j})$.[13] I place further parametric assumptions on consumer utility only when the model is estimated.

The Firm's Problem. Per-unit organization costs and competitive labor markets imply marginal costs are constant. I denote the feasible set of organization structures $\mathbb{B} = \{B \in \Delta^{N\times K} | \sum_i B(i,k) = \alpha_k \,\forall\, k\}$. The firm's problem can now be defined:

$$\max_{p_j \in \mathbb{R}_+, B_j \in \mathbb{B}} D_j(\underbrace{\xi(B_j)}_{\text{quality}} - \rho p_j, p_{-j}, \xi_{-j}) \left[p_j - \underbrace{\left(\overbrace{\gamma_j I(B_j)}^{\text{org. cost}} + \overbrace{W(B_j)}^{\text{avg. wage}} \right)}_{\text{constant marginal cost, } MC_j} \right] \tag{1.2}$$

Equilibrium. An equilibrium consists of firm strategies $\{p_j, B_j\}_{j=1}^{J}$ and wages w such that:

1. Firms choose prices p_j and organizational structures B_j to maximize (1.2).

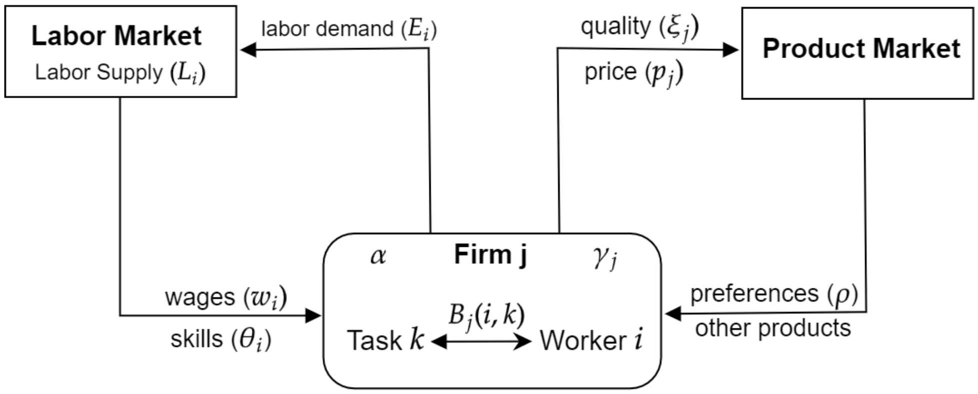
2. Labor markets for each worker type clear:

$$\sum_j D_j(\xi(B_j) - \rho p_j, p_{-j}, \xi_{-j}) \sum_k B_j(i,k) = L_i \,\forall\, i = 1, ..., N.$$

Model Summary. Figure 1.8 illustrates the model from the perspective of a single firm. The firm chooses B_j (i.e., determining who it hires and how hired workers are assigned to tasks) and prices taking into account internal factors (i.e., the task mix and organization costs), labor market factors (i.e., wages and skills), and product market factors (i.e., consumer price sensitivity, the prices and qualities of other products). The choice of B_j feeds back into the product market by determining product quality and prices, and into the labor market by determining labor demand across worker types.

13. There are several random utility models and representative consumer models consistent with this assumption, including multinomial logit, nested logit and mixed logit with a non-random price coefficient. A mixed logit model with consumer price sensitivity heterogeneity would violate the assumption.

Figure 1.8: Illustration of the Model



1.4.1 Discussion of Organization Costs

This section describes different ways of interpreting γ_j, the cost of increasing the complexity of firm j's organizational structure by 1 unit. The multifaceted nature of γ_j can account for several dimensions of organizational heterogeneity all at once. Unfortunately, this also means the mechanisms driving the value of γ_j at any particular firm can not be separated.

Coordination Costs. Under this interpretation, γ represents the fact that firms are "second-best solutions to transactions plagued by various forms of contractual incompleteness" (Gibbons 2020) and that "firms can never succeed in conquering the nonrational dimensions of organizational behavior" (Williamson 1984). As γ approaches 0, coordination costs disappear and a firm can design any organizational structure it chooses at 0 cost. When γ becomes sufficiently large, firms will resort to assigning every worker the same job. In the latter case, workers are essentially firms, since they perform all of the tasks the firm performs and do not rely on coworkers. In this way the distribution of γ traces out the value of firms. A firm with low γ is greater than the sum of its workers, producing a product superior to that which any of its workers could produce alone.[14]

14. Total organization cost increases linearly with the size of the firm for a fixed organization, and per-unit organization costs increase non-linearly with specialization. This is different than the coordination costs found in papers such as Becker and Murphy (1992).

Rational Inattention. The mutual information form of organization costs gives it a rational inattention micro-foundation. We can interpret γ as the level of "managerial talent" (Lucas 1978) which determines the attention cost needed to allocate tasks to workers. Similarly, organization costs also capture contractual inattention, such as those described by Tirole (2009). Different firms may find it more or less costly to write down complex contracts in order to support complex organizational structures.

Incentives for Teams. Under this interpretation, organization costs reflect losses due to free-riding. Dai and Toikka (2022) show that the profit of a firm managing a team of multiple workers increases with the productivity of the known technology. Thus, heterogeneity in γ_j reflects the fact that some firms know a larger number of technologies, or ways to combine tasks and workers.

Costly Specialization. Because complexity is a measure of distance from the random assignment of workers to tasks,[15] γ_j can be interpreted directly as a firm-specific specialization cost.

An example from the hair salon industry makes these ideas concrete. There are two main ways to organize a salon. In the chair-renter arrangement, stylists pay a fixed fee to the salon owner and keep all revenue. Chair renters set their own hours and develop their own client lists. In the employee arrangement, stylists are paid by the salon owner and do not run their own business. The chair renters are independent contractors and have little need for coordination; in the language of the model, there is little organizational complexity in this arrangement. In contrast, the employee arrangement exhibits complex contracts (including non-competes, commission-based compensation, etc.) and requires coordination. In the language of the model, there is significant organizational complexity in this arrangement.[16]

15. See Appendix Section 1.11.8 for a proof.

16. Johnson and Lipsitz (2022) survey salon owners and find that 48% are employee salons and 52% are independent contractor salons.

1.5 Theoretical Results

This section analyzes the theoretical model. I first show the profit-maximizing organization structure is also the solution to a simpler problem that is well studied in information theory and behavioral economics. I use this equivalence to understand the economic forces which determine each firm's internal structure.

1.5.1 Main Characterization

The firm's problem as written in Equation (1.2) appears complicated at first glance; there are $1 + N \times K$ choice variables and the objective is highly non-linear. The following theorem reveals the firm's problem can be greatly simplified.

Theorem 1 *An organizational structure (B_j^*) is profit-maximizing if and only if it solves*

$$\min_{B_j \in \mathbb{B}} \gamma_j I(B_j) + W(B_j) - \rho^{-1}\xi(B_j), \tag{1.3}$$

which is a rate-distortion problem and a rational inattention problem.

The proof of the result is provided in Appendix Section 1.11.2. The main idea of the proof is that if an organization structure does not solve Equation (1.3), the firm can switch to a structure that does and adjust its prices to strictly improve its profit. In this way, even though price and organization structure appear entangled in the firm's problem, they can be separated during analysis. The result relies on the fact that the quality-price index $\xi(B_j) - \rho p_j$ is sufficient for price and organization structure in demand, and demand is strictly increasing in the quality-price index. The result does not rely on the functional form of organization costs.

Theorem 1 is useful for three reasons. First, it allows the model to be taken to the data. Because (1.3) is a rate distortion and rational inattention problem, and these problems are well studied in information theory and behavioral economics, I can weave together results across the two strands of literature to identify firm-specific organization costs, prove a form of equilibrium existence and uniqueness, construct an estimation algorithm, and solve for counterfactual equilibria.

[17] **Proof.** In Appendix Section 1.11.1, I show that (1.3) is a rate-distortion problem and a rational inattention problem using a series of algebraic manipulations. Because (1.3) is a rational inattention problem with mutual information costs, I can leverage Matêjka and McKay (2015) and state that there always exists an organization structure which minimizes. Denote one such structure B_j^*. I now must show that B_j^* maximizes profit.

To do this, consider any structure and any price B_j', p_j' which is feasible but does not solve (1.3). Now examine the expression for profit:

$$D_j(\xi(B_j) - \rho p_j, p_{-j}, \xi_{-j})\Big[p_j - \Big(\gamma_j I(B_j) + W(B_j)\Big)\Big]$$

If the firm chooses B_j^* as its organization structure, it can construct the price:

$$p_j^* = p_j' + \Big(\gamma I(B_j') + W(B_j') - [\gamma I(B_j^*) + W(B_j^*)]\Big)$$

Notice that:

$$p^* - (\gamma I(B_j^*) + W(B_j^*)) = p' - (\gamma I(B_j') + W(B_j'))$$

Thus the second term of profit is the same under p_j', B_j' and p_j^*, B_j^*. Turning to the first term: I have that $\xi(B_j^*) - \rho p_j^* > \xi(B_j') - \rho p_j'$ because $\xi(B_j^*) - \rho p_j^* - (\xi(B_j') - \rho p_j')$ can be re-written as follows:

$$= \xi(B_j^*) - \rho\Big(\gamma I(B_j^*) + W(B_j^*) - [\gamma I(B_j') + W(B_j')]\Big) - \rho p' - (\xi(B_j') - \rho p_j') \tag{1.4}$$

$$= -\rho\Big[\Big(\gamma I(B_j^*) + W(B_j^*) - \rho^{-1}\xi(B_j^*)\Big) - \Big(\gamma I(B_j') + W(B_j') - \rho^{-1}\xi(B_j')\Big)\Big] \tag{1.5}$$

$$> 0 \tag{1.6}$$

17. Rate distortion theory and rational inattention are also related. See Denti, Marinacci, and Montrucchio (2020).

The final inequality follows from the fact that B_j^* was defined as a minimizer of (1.3). Since demand is strictly increasing in the price-quality index $(\xi(B_j') - \rho p_j')$, (1.6) implies demand is strictly higher under (p_j^*, B_j^*). Therefore firm profit from choosing (B_j^*, p_j^*) is strictly greater than choosing (B_j', p_j'). Since this argument holds for arbitrary (B_j', p_j'), B_j^*, is a profit-maximizing organization structure and the proof of the theorem is complete.

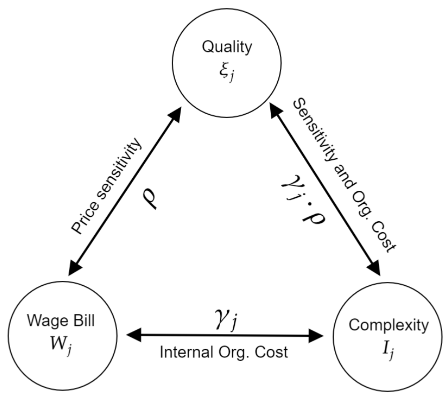
Second, the pricing and organization decisions can be separated when solving for an equilibrium for fixed wages. Specifically, a firm's internal organization *directly* affects own and competitor prices, but these do not directly affect internal organization. Additionally, one firm's internal organization does not *directly* impact a competitor's internal organization; however, each firm's internal organization is *indirectly* impacted by all prices and all competitor internal organizations via wages.

The separation implied by Theorem 1 has practical implications: it means equilibria are robust to timing. Although I assume firms choose organizations and prices simultaneously, if firms chose organizations first and then competed in prices, the outcomes would be the same. The separation implied by Theorem 1 also greatly improves tractability because for fixed wages, the organization problem can be solved first and then used to derive equilibrium prices. This simplifies counterfactual analyses and allows for policies to be decomposed in useful ways.

One question is whether the separation implied by the model is reasonable, or, equivalently, is it the case that wages are the main connection between different firms' internal organizations? The answer appears to come down to whether the labor market is well approximated by perfect competition and whether demand satisfies the index restriction. These assumptions seem reasonable in the case of hair salons, because they sell a horizontally differentiated product and are small in terms of employment, but they may not be in industries where differentiation is largely vertical (e.g., supermarkets) or where individual firms employ a large share of the labor market (e.g., manufacturing company towns). Appendix Section 1.11.12.2 discusses ways in which the model can be extended to accommodate these other contexts.

Third, Theorem 1 reveals the forces that shape a firm's internal organization. Examining Equation (1.3) shows firms face a triple trade-off, as depicted in Figure 1.9. Each firm wishes to achieve the lowest complexity and wages while achieving the highest quality. How it navigates this trade-off depends on its internal organization cost γ_j, consumer price sensitivity ρ and their interaction $\gamma_j \cdot \rho$.

Figure 1.9: The Complexity-Wage-Quality Trade-Off



If a firm wishes to increase quality, it has two options: (1) hire better workers and incur a wage cost or (2) rearrange its current workforce to better leverage existing worker skills and incur an organization cost. Intuitively, when consumers are price sensitive (ρ is high), the firm cannot pass on costs to consumers via prices. Thus, firms prioritize minimizing costs over maximizing quality by choosing less complex organizations.

To analyze how the firm navigates the complexity-wage-quality trade-off, I define the *organization frontier* as the set of all organization structures which minimize complexity for some quality-adjusted wages (Q). The frontier consists of the simplest organization that achieves some quality-adjusted wages. I wish to study the relationship between quality-adjusted wages and complexity along the frontier:

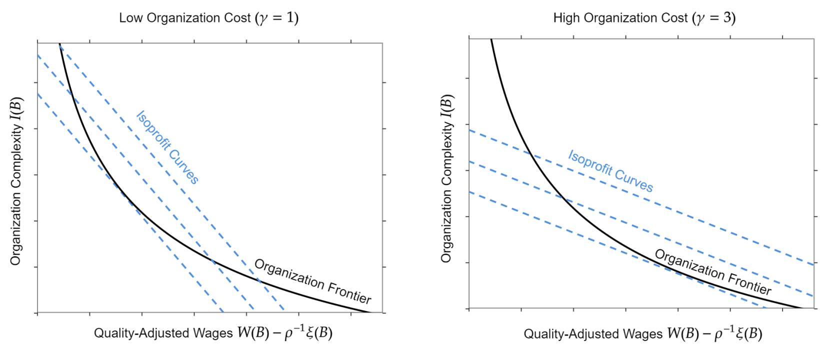
$$I^*(Q) = \min_{B_j \in \mathbb{B}} I(B_j) \text{ s.t. } W(B_j) - \rho^{-1}\xi(B_j) \leq Q.$$

The characterization provided in Theorem 1 allows me to apply existing results from information theory to understand the general shape of this relationship.

Proposition 1 *Organization complexity along the organization frontier ($I^*(Q)$) is continuous, convex and decreasing in quality-adjusted wages.*

The proof is provided in Appendix Section 1.11.3. The proposition implies the choice of a high-dimensional organization structure can be thought of as a two-dimensional choice, similar to a classic expenditure minimization problem from consumer theory, as illustrated in Figure 1.10. Although B_j (i.e., how a firm chooses its workers and how they are assigned to tasks) is a high-dimensional object, the firm essentially solves a two-dimensional trade-off between complexity and quality-adjusted wages. The firm's optimal structure will be the point of tangency between the organization frontier and the best possible (leftmost) isoprofit curve. The firm's isoprofit curves have a slope equal to $-\gamma_j^{-1}$. As the organization cost parameter (γ_j) rises, the curves become flatter, causing the tangent point to shift right and reducing organizational complexity while increasing quality-adjusted wages. A more complex organization allows a firm to produce a higher-quality good at a lower wage, but it requires a greater organization cost. An immediate consequence is that a lower organization cost parameter grants the firm an organizational competitive advantage in the product market.

Figure 1.10: Choosing an Organizational Structure



Note: Although B_j (i.e., how a firm chooses its workers and how they are assigned to tasks) is a high-dimensional object, the firm essentially solves a two-dimensional trade-off between complexity and quality-adjusted wages. The firm's optimal structure will be the point of tangency between the organization frontier and the best possible isoprofit curve.

Proposition 2 *In equilibrium, firms with a lower organization cost (γ_j) have higher organization complexity, market share and profits.*

The proof is provided in Appendix Section 1.11.3. Recall that γ_j represents the management technology, relationships, knowledge and practices specific to the firm which make it easier or harder for the firm to implement a given organizational structure. Proposition 2 implies more organizationally efficient firms are larger and more profitable, and can produce better-quality goods at a lower cost. Importantly, this proposition confirms that the model is consistent with Fact 2: complexity should be positively correlated with measures of firm size. This is in line with the findings of Kuhn et al. (2022), who use surveys and administrative data to show that more coordinated or specialized firms are more profitable.

1.5.2 Workforce Heterogeneity

The model assumes that workers are perfect substitutes in production, both in terms of quantity and quality. To see this, set $\gamma_j = 0$ and examine Equation (1.3). Without organization costs, the firm minimizes a constrained linear objective with weights determined by wages and skill sets. All complementarities between workers arise endogenously via organization costs. Because these costs are firm specific, this allows for rich heterogeneity within a product and labor market, both in terms of labor-labor substitution patterns and workforce composition.

I illustrate this with a simple version of the model with three worker types. Suppose wages are fixed at $w = (21, 20, 15)$, the task mix is $\alpha = (1/3, 1/3, 1/3)$, price sensitivity is $\rho = 1$, and worker skill sets are given by Θ (defined shortly). Under this worker-type space, there are two worker types that are specialists in task 1 and 3 relative to each other, but that have higher absolute skill in all tasks compared to a third type. When I "adjust" skills for wages, it can be seen that in relative terms, there are two workers who are optimal to hire for task 1 and task 3, and one jack-of-all-trades who is a safe option for all tasks:[18]

$$\Theta = \begin{bmatrix} 15 & 19 & 26 \\ 23 & 19 & 15 \\ 15 & 15 & 15 \end{bmatrix} \implies \theta - \rho w = \begin{bmatrix} -6 & -2 & 5 \\ 3 & -1 & -5 \\ 0 & 0 & 0 \end{bmatrix}.$$

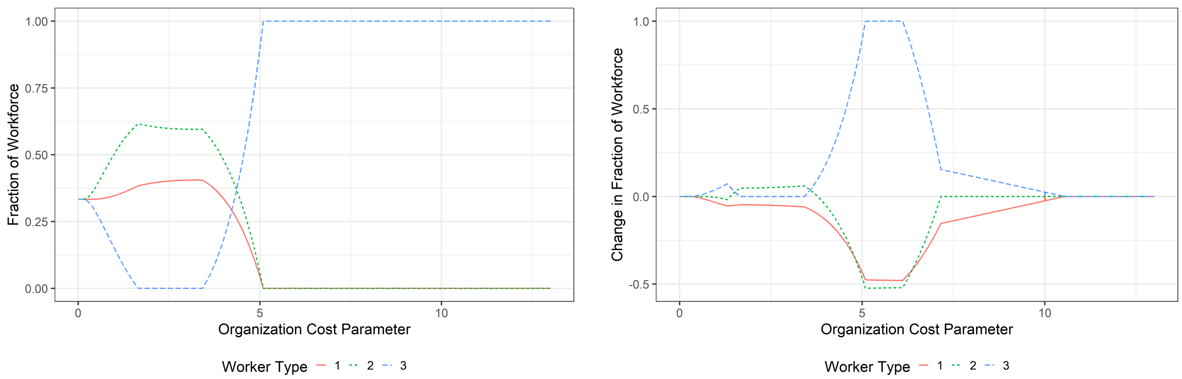
Firms facing the same market conditions and task mix can have heterogeneous workforce compositions, as illustrated in Figure 1.11 panel A. Organizationally efficient firms employ an equal share of each worker because they can fully utilize the specific skills of each worker type. Firms with intermediate organization costs hire only two worker types. Organizationally inefficient salons employ only type 3 workers (jacks-of-all-trades), because these firms cannot utilize the specific skills of the specialist types.

Additionally, firms facing the same market conditions and task mix can exhibit very different labor-labor substitution patterns, as demonstrated in Figure 1.11 panel B. When the wage of type 1 workers is increased by 1, firms with different organization costs react differently. Firms with very high or very low organization costs reduce the share of type 1 workers and increase the share of the other two types. Firms with intermediate costs reduce the share of both type 1 and type 2 workers. Thus, these two types are substitutes at extreme firms, but are complements at intermediate firms.

This section demonstrates that the model allows for heterogeneity both in labor-labor substitution patterns and workforce composition. The key driver of heterogeneity within a market is organization costs, which in the empirical application will be identified from the data. Thus the model allows for rich labor market responses to counterfactual policies, in contrast to many existing models, which generally restrict labor-labor substitution patterns prior to estimation.

18. These parameter values are based on an example in Csaba (2021).

Figure 1.11: Organizational Heterogeneity



(a) Workforce Composition

(b) Labor-Labor Substitution

Note: Panel A illustrates that as organization costs change, the composition of a firm's workforce changes in a non-monotone fashion. Panel B illustrates the change in the share of each worker type due to a 1-unit increase in the wage of type 1 workers. Type 1 and type 2 workers are substitutes at extreme firms (i.e., those with very high or very low organization costs) and complements at intermediate firms.

1.6 A Structural Model of Internal Organization

Understanding the quantitative relationship between internal organization and the labor and product market requires a structural model that can be taken to the data. This section describes such a model, which preserves the spirit of the theory developed in Section 1.5 while allowing for additional firm and worker heterogeneity. I prove the identification of the distribution of organization costs and provide a computationally light, nested-fixed point generalized method of moments estimation procedure.

1.6.1 Econometric Model

I define labor markets and product markets as counties, and time periods as quarters. I estimate the model for New York County (Manhattan) 2021 Quarter 2, the last full quarter with available data in my sample. I add several types of heterogeneity to the theoretical model introduced in Section 1.4 to better fit the data. The theoretical results in Section 1.5 continue to apply to the econometric model.

Consumers. I assume a parametric form for demand. There is a mass M of consumers interested in purchasing at most one of the J final products, where M is set to be the population of Manhattan. Consumer z's utility for good j is represented by the logit utility function

$$u_{z,j} = \xi(B_j) - \rho p_j + \epsilon_{z,j},$$

where $\epsilon_{z,j}$ is distributed i.i.d. Type 1 extreme value across consumers and products. The outside option for consumers is assigned index $j = 0$, and its utility is normalized to $u_{z,0} = \epsilon_{z,0}$. As in McFadden (1973), market demand for good j can be written as

$$D_j(\xi(B_j) - \rho p_j, p_{-j}, \xi_{-j}) = \frac{exp(\xi(B_j) - \rho p_j)}{\sum_{j'} exp(\xi(B_{j'}) - \rho p_{j'})}. \tag{1.7}$$

Expression (1.7) reveals the multinomial logit demand system satisfies the quality-price index restriction assumed during theoretical analysis.

Production Function Heterogeneity. The task mix is firm-specific and therefore indexed by j (α_j). This allows firms in the same product and labor market to have different organizational frontiers. Since tasks are observed, the distribution of time across task categories can be computed.[19]

Marginal Cost Heterogeneity. Marginal cost may depend on the firm-specific task mix (α_j) to capture the costs of materials relating to specific tasks (e.g., dyes) as well as an idiosyncratic marginal cost shifter ϕ_j, which has expectation 0 and is independent of γ_j and α_j. I measure $\bar{a}_j$ as the average number of hours salon j spends on a customer in a quarter. I specify that organization costs and wages are per hour of labor. This allows each firm to have a different required labor per unit ($\bar{a}_j$) so that I can capture traditional productivity differences across firms.[20] With these modifications, marginal cost can be

19. It is assumed that enough tasks are observed so that the computed shares are exactly equal to the underlying shares.

20. This is similar to specifying a production function of the form $\bar{a}_j \min\left\{\frac{a_1}{\alpha_1}, \ldots \frac{a_k}{\alpha_k}, \ldots, \frac{a_K}{\alpha_K}\right\}$.

expressed as

$$MC_j = \bar{a}_j\left[\gamma_j I(B_j) + W(B_j)\right] + \sum_k m_k \alpha_j(k) + \phi_j.$$

Quality Heterogeneity. In addition to endogenously chosen quality $\xi(B_j)$, each firm also has exogenous unobserved quality ν_j, which has expectation 0 and is independent of γ_j, α_j and $\bar{a}_j$. Exogenous unobserved quality ν_j represents reputation and other attributes that impact quality but are fixed in a given period and unrelated to labor. Inclusion of ν_j ensures only quality differences correlated with observed organization complexity (I_j) will be attributed to internal organization. Quality is now $\xi(B_j) + \nu_j$.

Worker Labor Supply Heterogeneity. Workers with the same skill set may differ in their labor supply. This clarifies the relationship between worker identities (observed in the data) and worker types in the model (unobserved). Specifically, in addition to being characterized by their skill set, workers are also characterized by an inelastic person-specific labor supply.[21]

Worker-Firm Matching. I augment the game by specifying that firms first demand an amount of labor of each skill set, and then an unspecified process matches workers to firms. The only assumption I place on this process is that the firm's labor demand from the first stage is exactly met. Thus if a firm demands 10 hours of a skill set, this amount may be met by any combination and number of workers, but no more or less than 10 hours is supplied in total. Following the matching process, firms then select an organization structure $\tilde{B}_j$, which is an assignment of worker identities to tasks.[22]

Worker Skill Sets. I assume there is one specialist worker type for each of the five tasks. Tasks are performed with a base skill level β_k when assigned to a non-specialist, and are

21. If the set of labor supplies is Λ, the worker type space is now $\Theta \times \Lambda$.

22. This means that in principle a firm may employ multiple workers of the same skill set and assign them different distributions of tasks. This allows me to bring the model to the data, where I observe worker identities but not worker types.

performed with an additive skill gap θ_k when assigned to a task k specialist. The matrix of skill sets, where each row denotes a worker type and each column a task, can be written as[23]

$$\Theta = \begin{bmatrix} \theta_1 + \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ \beta_1 & \theta_2 + \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ \beta_1 & \beta_2 & \theta_3 + \beta_3 & \beta_4 & \beta_5 \\ \beta_1 & \beta_2 & \beta_3 & \theta_4 + \beta_4 & \beta_5 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \theta_5 + \beta_5 \end{bmatrix}.$$

Sales Tax. The state of New York does not tax hair services. However, New York City levies a 4.5 percent tax on beauty services. Therefore, I denote the sales tax τ and assume it is 4.5% initially (New York City Department of Finance 2022).

Outside Option. I define consumers' outside option as not buying services from a salon. I use Consumer Expenditure Survey micro-data to compute the share of individuals from a county in a quarter who spend $0 at salons, and take this to be the share of people who choose the outside option. Based on this methodology, the share of New York County residents selecting the outside option in 2021 Q2 is 40%.

Profit. Under the econometric model, a firm's profit can be written as

$$\frac{exp(\xi(B_j) - \rho(1+\tau)p_j + \beta\alpha_j + \nu_j)}{\sum_{j'} exp(\xi(B_{j'}) + -\rho(1+\tau)p_{j'} + \beta\alpha_{j'} + \nu_{j'})}\left[p_j - \bar{a}_j\left(\gamma_j I(B_j) + W(B_j) + m\alpha\right) - \phi_j\right],$$

where the features added to the theoretical model are written in blue. Fixing an equilibrium, the parameters of the model can be divided into two groups. The first group is the firm-specific organization cost coefficients $\{\gamma_j\}_{j=1}^J$. The second group is the parameters that define worker skills (10 parameters), wages (5 parameters), material costs (5 parameters) and price sensitivity (1 parameter). I call these *market parameters* and denote them

23. In the absence of matched employer-employee data, which would allow the researcher to infer worker types, this restriction of the worker type space is likely necessary.

by Ω.

1.6.2 Equilibrium Existence and Uniqueness with Fixed Wages

In the empirical application of the model, I treat wages as fixed parameters to be estimated. Prior to identification and estimation, I establish that for fixed wages, there almost always exists a unique equilibrium.

Proposition 3 *Suppose wages are fixed parameters. A pure strategy equilibrium always exists, and it is unique except over a set of parameters with measure 0.*

The proof of this result is provided in Appendix Section 1.11.5, and it relies on the equivalence to a rational inattention problem established in Theorem 1. This result means that multiplicity arises only in knife-edge cases. Proposition 3 does not establish equilibrium uniqueness or existence in the full model with wages determined endogenously by labor market clearing. Nevertheless, Proposition 3 is crucial for proving Proposition 4, the main identification result in this paper.

Several aspects of the model make Proposition 3 surprising. First, each firm has 26 choice variables, and quality and marginal cost are endogenous. Many models where product positioning is endogenous (including the canonical two-stage Hotelling model) suffer from equilibrium existence and uniqueness problems.[24] This idea is captured well by Shaked and Sutton (1987): "It is notorious in models of product differentiation that equilibria may fail to exist for many reasonable specifications of the standard models."

The result is also useful for counterfactual analysis, because it means the model almost always delivers one and only one internal organization structure for each firm. The model will almost never suffer from the inverse identification problems, at least when wages are held fixed.

24. In a two-firm Hotelling model with product positioning, it is known that a pure strategy equilibrium does not exist for linear transportation costs. When transportation costs are quadratic, there are two equilibria (d'Aspremont, Gabszewicz, and Thisse 1979)

1.6.3 Identification of Firm-Specific Organization Costs

The organization costs $\{\gamma_j\}_{j=1}^J$ are important parameters, determining product quality, organization complexity and marginal costs for each firm. However, the fact that there is one parameter per firm and that I place no restrictions on the economy-wide distribution raises concerns for identification and estimation. I alleviate this concern with the following result.

Proposition 4 *Organization costs (γ_j) and organization structures (B_j) are a known function of firm task mixtures (α_j), complexities (I_j) and market parameters (Ω) for all firms with positive complexity, except for a set of market parameters with measure 0.*

The proof is fully described in Appendix Section 1.11.6, and it makes use of the essential equilibrium uniqueness result given in Proposition 3. A key hurdle is that I do not observe worker types, but worker identities within firms. Because I allow for a flexible matching process, a given firm may in principle employ multiple workers of the same skill set and assign these workers different tasks. However, a property of the mutual-information based organization cost ensures that if firms do employ multiple workers with the same skill set, they assign these workers the same tasks. This implies that the observed organization complexity based on worker identities is equal to the true organization complexity based on worker skill sets.

The intuition for the identification of γ_j is demonstrated in Figure 1.10. Suppose two firms with the same task mix (α_j) are observed in the same product and labor market. This means they have the same organization frontier. If firm A has a higher complexity, it must be that the slope of firm A's isoprofit curve is steeper than B's, which can only be because A has a lower organization cost. Therefore, I can order the firms by organization cost without knowing the market parameters. Once these parameters are known, I can find the cardinal values of each firm's organization cost.

The proposition implies that organization costs do not need to be estimated in the statistical sense. For any market parameters, there are unique organization cost parameters which rationalize the observed organization complexities and task mixtures. This is similar to the way unobserved product quality is a known function of market shares in Berry, Levinsohn, and Pakes (1995). Another similarity is the lack of a closed form for the known function. Instead, I provide a fixed-point algorithm as part of my suggested estimation routine.

Beyond estimation, Proposition 4 also implies that observing the task mix (a vector of length K) and organizational complexity (a scalar) is enough to estimate the model. It is not necessary to observe the individual assignments of workers to tasks; the researcher need only observe complexity. This means the model can be estimated in settings where rich assignment data are not available.[25] It also means that a researcher who has assignment data can estimate the model using only complexity and the task mix and use the rest of the data to conduct validation exercises. This is precisely what I do in Section 1.7.2.

Economically, the result implies the task mix (a vector of length K) and complexity (a scalar) are sufficient statistics for a firm's internal organization structure (B_j, an $N \times K$ matrix). When all firms share the same task mix, a firm's internal structure can be fully inferred by competitors through complexity alone. On the one hand, this illustrates that multinomial logit demand and mutual information-based organization costs are imposing a large amount of structure. On the other hand, to the extent that multinomial logit demand and mutual information-based organization costs are reasonable, the heart of the model can remain tractable even if extended to settings with more complex strategy spaces.

1.6.4 Estimation of Market Parameters

I have established that organization costs are a known function of the data and market parameters. This section derives a set of moments and assumptions under which the market parameters can be estimated via the generalized method of moments. I note that material costs are not separately identified from $\mathbb{E}[\phi_j]$ and the skill base is not separately

25. For example, privacy concerns may often prevent the disclosure of employee-client assignments.

identified from $\mathbb{E}[\nu_j]$. I therefore estimate $\mathbb{E}[\nu_j] + \beta_{cut}$ and $\mathbb{E}[\phi_j] + m_{cut}$ and call them the demand and cost intercepts, respectively. This means all skills parameters and material costs parameters are relative to the haircut/shave task.

To construct moment conditions, I follow a common approach in the literature and use one demand-side and one supply-side equation. Starting with the supply side, the equilibrium pricing equation can be written as

$$p_j = \frac{1}{\rho(1+\tau)(1-s_j)} + \bar{a}_j\left[\gamma(\Omega, I_j, \alpha_j)I_j + W(\Omega, I_j, \alpha_j)\right] + m\alpha_j + \phi_j. \tag{1.8}$$

Because the demand system takes a multinomial logit form, market shares can be expressed as

$$log(s_j) - log(s_0) = \xi(\Omega, I_j, \alpha_j) - \rho(1+\tau)p_j + \beta\alpha_j + \nu_j. \tag{1.9}$$

I interact firm-level covariates with Equations 1.8 and 1.9. I use covariates that are relevant to the determination of prices and market shares but also independent of ν_j, ϕ_j. The firm organizational complexity (I_j) and task-mix vector (α_j) fit these requirements, because they change organization costs but are not impacted by ν_j, ϕ_j. Additionally, I include their interaction, $\alpha_j \cdot I_j$. A discussion of how this variation identifies specific parameters is provided in Section 1.6.6.

I add one additional wage moment. For each county and quarter, I compute the average wage of hair stylists using the Quarterly Census of Employment and Wages. I take the total quarterly wages of establishments with NAICS code 812112 and divide them by the number of establishments. This generates the average total wage bill, which corresponds to $W(\Omega, I_j, \alpha_j)$ multiplied by the number of customers. I allow for classical measurement error (e_j) which yields

$$W_j = Ms_ja_jW(\Omega, I_j, \alpha_j) + e_j.$$

Taken together, the moment conditions used for estimation are

$$\mathbb{E}\left[\begin{pmatrix}\phi_j(\Omega, I_j, \alpha_j, p_j, s_j)\\ \nu_j(\Omega, I_j, \alpha_j, p_j, s_j)\end{pmatrix}\begin{pmatrix}\alpha_j & \alpha_j I_j\end{pmatrix}\right] = 0 \qquad \mathbb{E}[e_j(\Omega, I_j, \alpha_j)] = 0.$$

For a single market and quarter, I obtain a total of 21 moments to estimate 21 market parameters. The model is globally identified if I assume that Ω is the unique vector of parameter values which satisfies the moment conditions. With this assumption, I estimate the model using the generalized method of moments (GMM). Denote the sample analogue of the moments as $G(\Omega)$. Then, to estimate Ω, I solve

$$\arg\min_{\hat{\Omega}} G(\hat{\Omega})' W G(\hat{\Omega}), \tag{1.10}$$

where W is a weighting matrix. Note that to evaluate this GMM objective, I must recover the vector of organization costs implied by the data and each guess of the market parameters. This requires solving each firm's internal organization problem many times per evaluation of the GMM objective. The next section provides a result which greatly reduces the computational resources needed to do this.

I take the weighting matrix W to be a diagonal matrix, where each diagonal element is the sample variance of the independent variable involved in the moment. This standardizes the moments in the objective function, preventing variables with large nominal values (i.e., average hours per unit) from dominating during estimation. I constrain wages to be between \$15 (the minimum wage) and \$200 per hour. I require that the algorithm search only over parameters values yielding a positive demand share for each type of labor.

1.6.5 A Computationally Light Estimation Procedure

Although organization costs are a known function of the data, there does not exist a closed-form expression for this function. This is a problem for estimation, because it is necessary to numerically solve each firm's internal organization problem many times until the model-produced complexities match the complexities in the data. This is computationally intensive when there are many tasks and many firms, because each firm's problem is a high-dimensional non-linear minimization problem. This section solves this problem by providing a nested-fixed point algorithm which efficiently solves the firm's problem and is proven to globally converge.

To derive the algorithm, I use the equivalence to a rate-distortion problem proved in Theorem 1.

Lemma 1 *Given market parameter values, the Blahut–Arimoto algorithm with Lagrangian multiplier* $(\bar{a}_j\gamma_j)^{-1}$ *delivers an organizational structure* B_j *which maximizes firm profit.*

The lemma follows directly from Theorem 1 and well-known results in information theory.[26] The Lagrangian multiplier involves $\bar{a}_j$ because marginal costs are $\bar{a}_j(\sum_i w_i E_i + \gamma_j I(B_j)) - \rho^{-1}\xi(B_j)$. The Blahut–Arimoto algorithm (Blahut 1972) is a fixed-point algorithm which iterates on two optimality conditions and can be described as follows (suppressing firm subscripts):

0. Guess some labor demand E^0. Create matrix V:

$$V_{i,k} = exp[(\bar{a}\gamma)^{-1}(\rho^{-1}\theta_{i,k} - w_i)].$$

1. Compute B^t as

$$B(i,k)^t = \alpha_k \frac{V_{i,k}E^t(k)}{\sum_i E^t(i)V_{i,k}}.$$

2. Compute E^{t+1} as

$$E^{t+1}(i) = \sum_k B(i,k)^t.$$

3. If converged, exit; else return to Step 1 and advance t.

26. See Tishby, Pereira, and Bialek 2000 or Blahut 1972.

The Blahut–Arimoto algorithm is proven to converge to a global optimum from any feasible starting point (Tishby, Pereira, and Bialek 2000). It avoids the need to repeatedly use nonlinear optimization routines, and because it is a fixed-point method, acceleration routines such as Du and Varadhan (2020) can be used to increase the speed of convergence. The algorithm also delivers the entire internal organization of the firm, B_j. Practically, a researcher can use the algorithm to search for the γ_j which makes the model-generated complexities match the complexities observed in the data. Because complexity is monotone in γ_j, a researcher can use the bisection method for this task. Thus, the full estimation procedure is as follows:

1. Given a guess of the market parameters $\hat{\Omega}$, use the Blahut–Arimoto algorithm to find the organization costs $\gamma_j(\hat{\Omega})$ which rationalize each firm's observed organizational complexity I_j.

2. Using $B_j(\hat{\Omega}), \gamma_j(\hat{\Omega})$, compute firm-specific wage bills and endogenous quality.

3. Evaluate the GMM objective given by Equation 1.10.[27] If the objective is minimized, stop; otherwise, return to step 1 with a new market parameter guess, $\hat{\Omega}$.

This estimation algorithm is similar in spirit to the demand estimation procedures that have become popular in industrial organization since Berry, Levinsohn, and Pakes (1995). Just as those procedures invert market shares using a contraction mapping to derive unobserved product qualities, my procedure inverts organization complexities using a contraction mapping to obtain unobserved organization costs. Implicit in this inversion procedure is that complexity is measured without error. Appendix Section 1.11.16 provides evidence that measurement error is small.

1.6.6 Identifying Variation

Proposition 4 establishes that given fixed values for the market parameters, organization costs (γ_j) are identified by variation in complexity and the task mix across firms. The

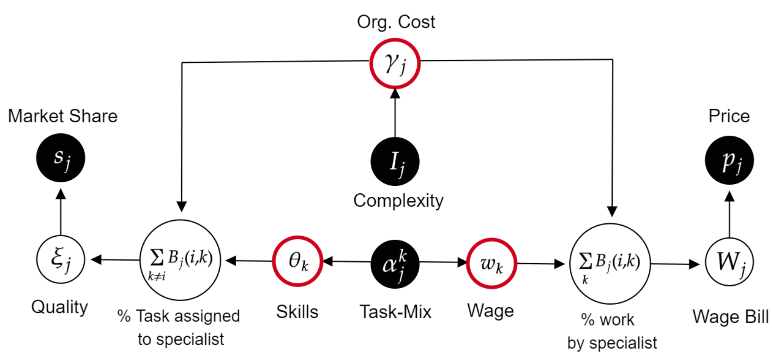
27. I use the R package "gmm" described in Chausse (2021) to perform estimation.

purpose of this section is to discuss the sources of identifying variation for the market parameters.

Consumer price sensitivity is identified by the pass-through of average wages to consumers. If wage costs are passed through to consumers via higher prices, consumers are not price sensitive and ρ is low. Once price sensitivity (ρ) is known, the marginal cost of each firm can be obtained by subtracting the markup from prices. Similarly, service quality can be obtained from observed prices and market shares. For this reason, I discuss identification of the other parameters as if quality and cost were observed.

The other market parameters are identified using variation across firms in the task mix and complexity. The relationship between market parameters and the observed data is demonstrated in Figure 1.12.

Figure 1.12: Identifying Variation for Skills and Wages



Note: The diagram shows the way market parameters (skill sets and wages) are identified from observable data. Black objects are observed data, while white objects are parameters to be estimated or unobserved variables, such as organization costs. Organization costs mediate the relationship between particular skill and wage parameters and costs and qualities. Since organization costs are known functions of complexity, it is possible to use variation in the interaction of complexity and the task mix to identify the different parameters.

The base skill parameters (β) and the material costs (m) are identified by variation in the task mixtures (α_j) across firms. When I observe a firm that is intense in task k, I its cost is informative about m_k and its quality is informative about β_k. This is why α is interacted with the demand supply side residuals to obtain a first set of moments.

Recall that a complex firm generally has a specialized workforce. When I observe a firm that is complex and is intense in a task k, this implies two things. First, the firm likely uses a large share of task k specialists. Therefore, the observed price (and thus cost) of that firm largely reflects the wage of task k specialists (w_k). Second, the firm assigns a large amount of task k to those specialists. Therefore, the observed market share (and thus quality) largely reflects the skill gap of task k specialists (θ_k). This is why $\alpha_j \cdot I_j$ is interacted with the demand supply side residuals to obtain a second set of moments.

1.7 Empirical Results

This section summarizes parameter estimates and uses the model to analyze the sources of variation in task content for hair stylists in Manhattan.

1.7.1 Parameter Estimates

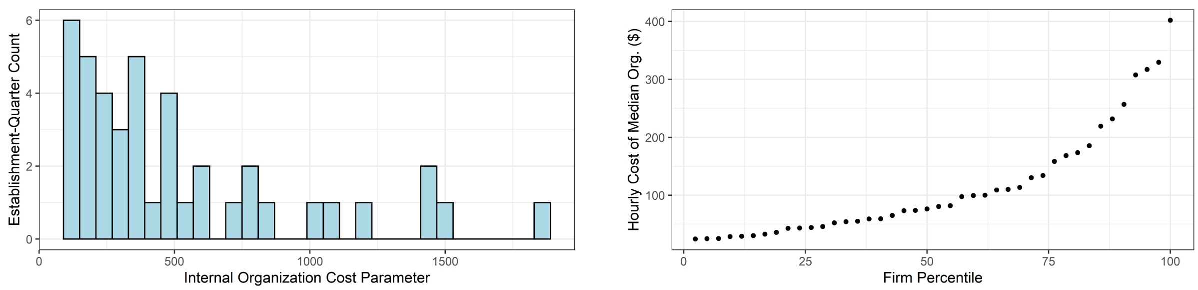
Estimated wages, skill parameters and material costs are organized by task in Table 1.5. Price sensitivity and the intercepts are provided in Table 1.6. Standard errors are computed as the sample standard deviation of the parameter estimates from 500 bootstrap replications, with the procedure described in Appendix Section 1.11.22.

The coefficients associated with the color and haircut tasks are the most precisely estimated. This is not surprising, as these tasks are the most common and, as a consequence, their associated parameters will have the most statistical power. Across all tasks, the skill-gap parameters are positive, indicating that assigning the task to the associated specialist increases quality. The skill-gap parameters can be interpreted as the dollar value to a consumer of increasing task specialization in that task by 4 percentage points. Wages are in 2021 dollars per hour. Material costs are in terms of 2021 dollars per service.

Wages for color specialists are more than double the wages for haircut specialists, and the skill gap for color specialists is nearly double the skill gap for haircut specialists. This is in line with folk wisdom in the industry that it is hard to master coloring. The material costs are largest for the color task, in line with the fact that coloring is intense in expensive non-labor supplies, such as hair dye.

The organization costs (γ_j) for each firm are shown in Figure 1.13a. To provide a magnitude for these estimates, I plot the cost of implementing the median-complexity organization structure across all firms in Figure 1.13b. There are large differences in organization costs. Firms in the bottom quartile of organization costs can implement the structure for less than \$50 an hour. It would cost firms in the top quartile over \$150 an hour. The estimates imply that variation in organization costs explains 40% of total variation in prices across firms.[28]

Figure 1.13: Estimated Organization Costs



(a) Organization Costs (γ_j) **(b)** Cost of Median Org. Structure

Note: Panel A displays the estimated organization cost (γ_j) parameters for Manhattan. These can be interpreted as a measure of organization frictions at each firm, with lower values indicating less friction. Panel B displays the magnitude of these differences, by plotting the cost (in dollars) to each firm of implementing the median-complexity organization structure.

For each firm, I recover the unobserved, equilibrium organization structure B_j. Four examples are visualized in Figure 1.14. These matrices represent the amount of time allocated to each task and each worker type.[29]

1.7.2 Model Fit and Validation

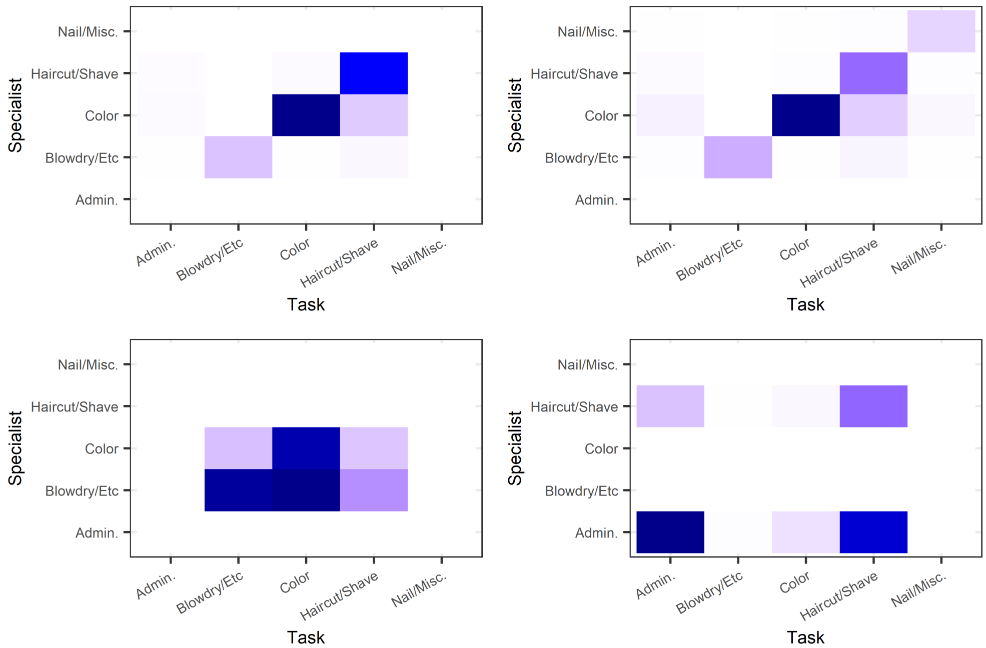
I assess model fit by comparing the predicted and actual relationship between prices and various organizational variables in Figure 1.15. The model captures the shape of the relationships.

Although the model delivers an entire predicted organization structure (B_j) for each

28. This is obtained as the R-squared of regressing price on $\bar{a}_j \gamma_j I(B_j)$.

29. I present two estimated structures in tabular form in Appendix Table 1.17.

Figure 1.14: Estimated Organization Structures

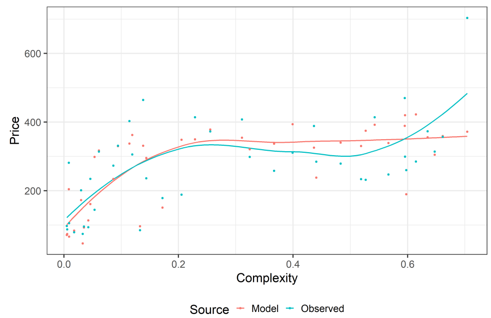


Note: Darker colors indicate a higher fraction of total labor was allocated to that worker-task pair. The first two structures are for organizationally efficient salons, while the last two are for organizationally inefficient salons.

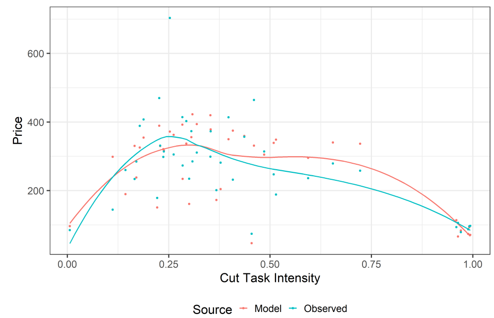
firm, the estimation procedure uses only some of this information. I use the additional predicted information to validate the model. In particular, I compare the model-generated distribution of task content to the observed distribution of task content. Recall that the jobs within firm j are denoted by b_j, which is a matrix where element i, k denotes the time worker type i spends on task k. Using the model, I can compute b_j for each firm among worker types that it hires. In the data, I can compute $\tilde{b}_j$, which are jobs within firm j, where element i, k denotes the time worker i spends on task k. The main difference between $\tilde{b}_j$ and b_j is that the first is with respect to worker identity, and the second is with respect to worker types. To make them comparable, I can weight each job by the total amount of labor it represents. Combining all J firms yields an unobserved and model-based distribution of job task content for each of the five tasks, where jobs are weighted by their effective labor.

Tables 1.7a, 1.7b and 1.7c compare the model and observed mean, median, variance, 25th percentile and 75th percentile of job task content. The estimated results exactly match

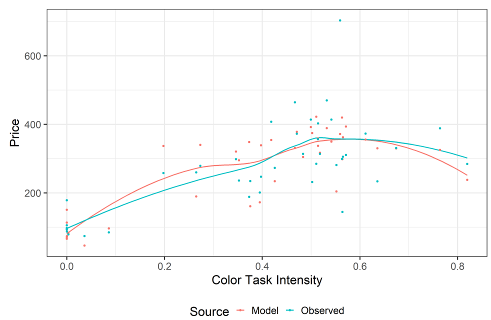
Figure 1.15: Model Fit



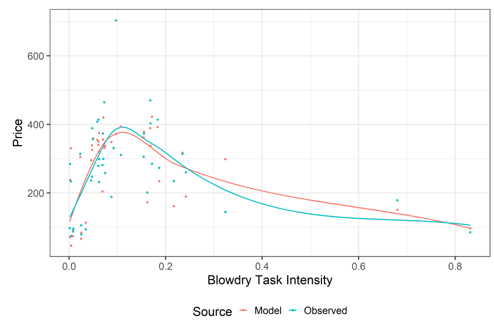
(a) Complexity



(b) Task Mix 1 (Haircut/Shave Task)



(c) Task Mix 2 (Haircut/Shave Task)



(d) Task Mix 3 (Haircut/Shave Task)

Note: Each panel plots the model and observed relationship between price and different firm variables. Dots represent individual firms, while lines are Loess smoothed fitted curves.

the mean and between-firm variance of job task content because the model imposes that organization structures must be consistent with the task mix α_j, which is exactly the average amount of time spent on each task at each firm. The estimates are also reasonable approximations of the total variances of task content and the 75th percentile of the job task-content distribution. The model is not able to match the median and the 25th percentile.

The statistics related to the color/highlight/wash task are the hardest for the model to replicate, because the empirical distribution for this task is triple peaked. The empirical and model-generated task-content distributions are presented in full in Appendix Figure 1.31.[30]

30. Comparing the entire distribution of actual and model-predicted task content is a demanding test of the model.

1.7.3 The Determinants of Task Specialization

The estimated model allows the researcher to understand how worker skills and firm internal organization determine the task specialization of jobs. Measuring task specialization as the amount of time a worker spends on their specialty task, I find that 45% of the variation in task specialization is attributable to firms, while 55% is attributable to worker skills. [31]

I calculate the variation in task specialization due to the firm component by computing the fraction of total firm labor spent on any worker's specialty. Firms with higher organization costs exhibit less task specialization. The magnitude of this effect is large: firms in the bottom quartile of organization costs (efficient firms) assign on average 90% of tasks to the associated specialist, while firms in the top quartile (inefficient firms), only 67%.

There is also significant variation in specialization across worker types. Haircut/shave specialists work the most specialized jobs, spending 95% of their time on their specialty task. Blow dry/extension/style specialists work the most generalized jobs, spending only 48% of their time on their specialty task.

1.8 Counterfactuals

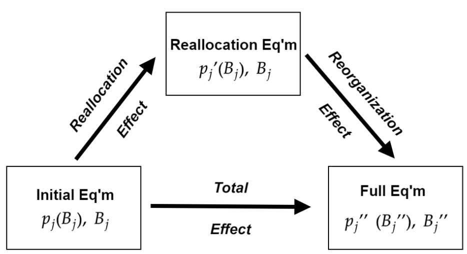
This section uses the estimated model to study two counterfactual policy changes, one impacting the product market and one impacting the labor market. Internal organization qualitatively alters the responses to these well-studied policies. The procedure used to solve for equilibria and conduct the analyses is described in Appendix Section 1.11.24. The model allows me to distinguish between two effects of any policy: a reallocation effect and a reorganization effect.

31. In the decomposition, I separate task-specialization variance into a within-worker type and across-worker type component. Since the only difference across worker types is firms, I can call the across-worker type component the firm component of variance.

To do this, I first define the reallocation equilibrium. It is the outcome when firms are allowed to adjust prices (p_j) but organization structures (B_j) are fixed at the initial equilibrium choices. Because prices control quantities, this equilibrium allows firms to adjust the total labor they hire, but not the division of labor within the firm. The *reallocation effect* of any policy change is the change in outcomes between the reallocation equilibrium and the initial equilibrium. It captures changes due to the reallocation of labor across firms. Because firms differ in their organization costs and task mixtures, reallocation will change the task content of jobs, relative wages, and other outcomes.

The *reorganization effect* of any policy is the change in outcomes between the full equilibrium and the reallocation equilibrium. It captures changes due to reorganization of labor within firms. I define the *total effect* of any policy change as the change in outcomes from the initial to the full equilibrium. These relationships are summarized in Figure 1.16.

Figure 1.16: Reallocation, Reorganization and Total Effect



Note: The diagram shows how the total effect of a counterfactual policy can be decomposed into two parts. The first part fixes the organization structure but allows firms to update prices. This is the reallocation effect, because labor flows across firms but does not change within firms. The second part allows organization structures to adjust. This is the reorganization effect, because firms are now responding to wages and prices by shifting their internal structure.

In the reallocation equilibrium, firms are acting as if they employ a composite worker. The worker's skills and wage are determined exogenously by the initial internal organization of the firm, B_j. The firm has the option of adjusting the total amount of labor it demands from this composite worker, but cannot adjust the worker's skills and wages. In the full equilibrium, the firm is free to fully adjust its internal structure.

1.8.1 Minimum Wage Increase

I study a counterfactual increase in the minimum wage in Manhattan from \$15 (the minimum in 2021) to \$20. An increase to \$20 is similar in magnitude to the increases that would occur if the minimum wage were pegged to inflation, as proposed in several pending pieces of legislation.[32]

To implement the counterfactual, I require that all equilibrium wages be at least \$20, and that markets clear for all worker types for which the wage is not binding. I allow there to be excess labor supply (unemployment) for those worker types facing a binding minimum wage. The model is well suited for studying large increases in the minimum wage because it allows salons to reorganize as well as raise prices. There are technical details that must be addressed when implementing the minimum wage counterfactuals, including the possibility of multiple equilibria and numeraire goods. I address these in Appendix Section 1.11.24.2.

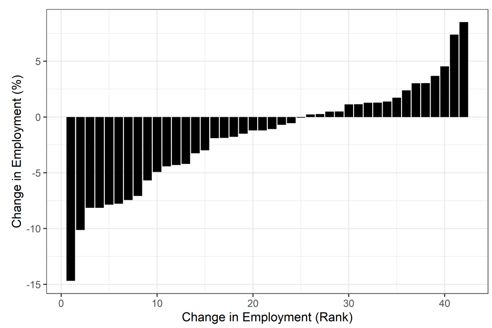
I find that the minimum wage binds for the haircut/shave specialist only. The new wages and employment levels across worker types are given in Table 1.19 (including values for the reallocation equilibrium). I first discuss the reallocation and reorganization effects of this policy change. I then analyze the overall impact of the new policy, using the reallocation and reorganization effects to understand the underlying forces.

1.8.1.1 The Reallocation Effect

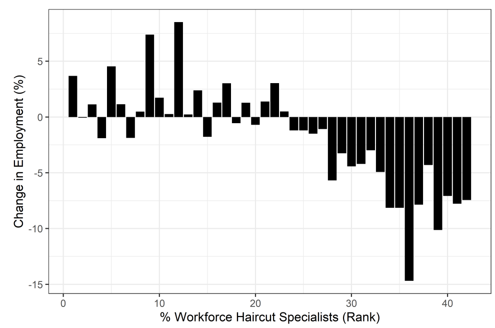
The impact of the minimum wage on individual salons depends partly on their initial internal structure. As a result, the minimum wage changes the competitive positions of salons and reallocates labor. By comparing the initial and reallocation equilibrium, I can hold each firm's internal structure fixed but allow firms to adjust prices. This captures the extensive margin adjustment of salons but prevents internal reorganization. Figure 1.17 presents the reallocation effect of the minimum wage in a series of three panels.

32. Senate Bill S3062C and Assembly Bill A7503B.

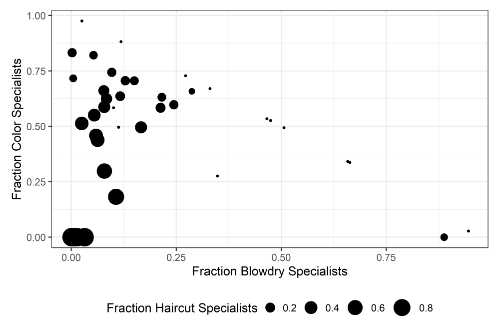
Figure 1.17: The Minimum Wage Reallocation Effect



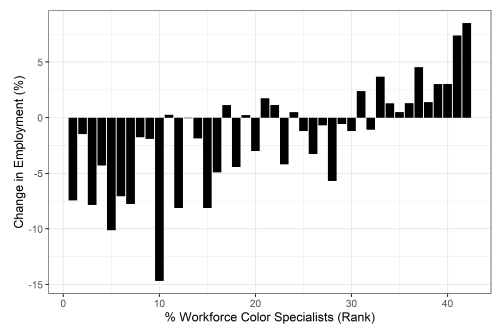
(a) Ordered by Employment Loss



(b) Ordered by Share Binding Workers



(c) Initial Coworkers



(d) Ordered by Share Color Specialists

Note: In panels A, B and D each bar is a firm, and employment changes are comparing the reallocation equilibrium to the initial equilibrium, holding fixed internal organization. Panel A orders firms by employment losses. Panel B reorders firms by the fraction of the workers that are haircut specialists in the initial equilibrium. Panel C plots firms by their initial workforce composition. Panel D orders firms by their share of color specialists.

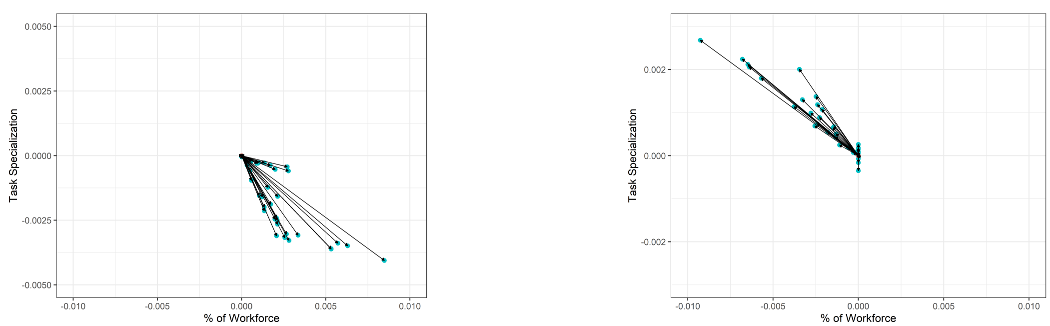
The minimum wage has a disproportionately negative impact on salons whose internal organization relies heavily on minimum wage workers. These salons see the largest increases in marginal costs and thus the largest decreases in output and employment. Because the minimum wage increases some salons' costs more than others', it changes the competitive position of firms in the product market. As can be seen in the figure, this effect is heterogeneous enough that some salons see employment increases.

Workers that are often employed alongside minimum wage workers initially see negative wage spillovers because the minimum wage erodes the competitive position of these firms. In the opposite way, workers employed at salons with few minimum wage workers initially see positive wage spillovers, because the minimum wage improves the competitive position of these firms. The effects of the minimum wage is contagious, and are spread across workers based on firm internal organization. In equilibrium, the minimum wage reallocates labor towards high-complexity, task-specialized salons and away from low-complexity, task-generalized salons, raising industry task specialization and average worker productivity.

1.8.1.2 The Reorganization Effect

By comparing the full equilibrium and the equilibrium where firms can adjust only prices, I can study the effect of internal reorganization. In Figure 1.18, I plot the vectors representing firms, where the length and direction of the vector represents the change in the firm's relative labor demand for that worker type and the change in task specialization of that worker type at the firm.

Figure 1.18: Reorganization Effect Under a Minimum Wage Increase



(a) Non-Binding (Color/Highlight Specialists) **(b)** Binding (Haircut/Shave Specialists)

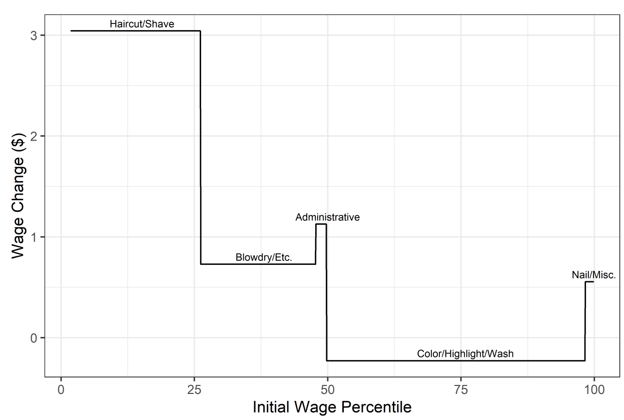
Note: Each arrow in both panels is a firm, with the blue dot at the end of the arrow representing the firm after the reorganization effect (the final position). Panel A displays the change in task specialization and relative employment for color/highlight specialists, a type for which the minimum wage is not binding. Relative employment increases, while task specialization decreases. Panel B displays the change in task specialization and relative employment for haircut/shave specialists, a type for which the minimum wage is binding. Relative employment decreases, while task specialization increases. This illustrates how firms are asking surviving workers to "pick up the slack."

The figure illustrates a general pattern. Salons reduce relative employment and increase task specialization of minimum wage workers. Salons reduce task specialization and increase relative employment for workers above the minimum wage. I call this a "pick-up-the-slack" effect. Intuitively, the minimum wage reduces the comparative advantage of workers for which the minimum wage binds in all tasks relative to other (non-binding) workers. Firms respond by laying off minimum wage workers and shifting tasks performed by them onto the relatively less expensive non-binding workers. Only minimum wage workers that are sufficiently productive survive, which are those who are task specialized. This implies that the minimum wage increases the absolute productivity of binding workers, but decreases the absolute productivity of non-binding workers.

1.8.1.3 Total Impact

Although the minimum wage is binding for only one worker type, all workers see wage changes. Table 1.8a shows that there are both positive and negative wage spillovers. The largest positive spillover is for administrative specialists, who see a wage increase of 4.2% (+$1.13). Color/highlight/wash specialists see a small wage decrease, of 0.7% (-$0.23). What is notable about these spillovers is that they are non-monotonic in initial wage. To see this, I plot the wage change experienced by different workers ordered by initial wage in Figure 1.19. Non-monotone spillovers occur because substitution patterns in the model are determined endogenously based on the distribution of firm organization costs and task mixtures.

Figure 1.19: Minimum Wage Spillovers Across the Initial Wage Distribution



Note: This figure plots the wage change experienced by different workers ordered by the initial wage of the worker. Haircut/shave specialists are the only binding type, so their wage increase is due to the direct effect of the minimum wage. All other wage changes are spillovers. While the majority of workers see wage increases, some see decreases. Spillovers are not monotone in initial wage.

These non-monotonic wage spillovers illustrate that internal organization can link workers that are very far apart in the initial wage distribution. Workers that differ horizontally in their specialty may be quite likely to work alongside each other, and they may have quite different wages depending on other factors. In this way the reallocation effect can cause large wage increases or decreases even for high-wage workers: indeed, this is exactly what I observe with haircut and nail specialists. Similarly, because the initial wage distribution is not determined by vertical skill differences, the reorganization effect will induce firms to shift tasks from minimum wage workers to workers across the wage distribution.

In some empirical work it is common to use the top of the wage distribution as a placebo test for studying minimum wage increases (Cengiz et al. 2019). The idea, which holds in a wide class of task-based models, is that workers at the top of the distribution should be unaffected by minimum wage increases because they are not substitutes for workers at the bottom of the distribution. This idea completely breaks down under my model.

In Table 1.9 I decompose wage spillovers into those arising from the reallocation and the reorganization effect. As discussed in the prior two subsections, spillovers for each worker type are a combination of forces, with the reorganization and reallocation effects sometimes moving in opposite directions. For example, color specialists see negative wage spillovers because they are employed alongside minimum wage workers and the minimum wage increase disadvantages the salon where they work. But they also see positive wage spillovers because firms shift tasks from minimum workers to them during reorganization. The total wage spillover for color specialists is negative, as the reallocation effect is about double the reorganization effect. For binding workers (haircut/shave specialists) the two effects work in the same direction, increasing unemployment. In this sense, internal reorganization amplifies unemployment losses.

Table 1.9b shows that reorganization wage spillovers follow a pattern. Workers that see an increase in task specialization see a wage decrease (or unemployment increase), while workers that see a decrease in task specialization see a wage increase. Because task specialization determines worker productivity, this implies that internal reorganization causes absolute productivity and wages to move in opposite directions.

Table 1.10 summarizes the welfare impacts of the policy. Wage gains for employed workers amount to \$1.6 million, and they are greater than salon losses (-\$714,413) and unemployed wage losses (\$600,240) combined. A little less than a quarter of the wage gains (\$389,847) come from positive spillovers. Welfare losses are concentrated among consumers, who see welfare reductions due to an average 1.7% price increase and an average 0.5% quality decrease. The minimum wage has heterogeneous impacts across firms, with most firms seeing employment losses but some firms seeing employment gains.

1.8.1.4 Productivity and the Minimum Wage

The minimum wage literature has put forward the idea that minimum wage increases may increase worker productivity. A rationale for this is efficiency wage theory, which holds that in response to better pay, workers may exert more effort. My model provides an alternative, firm-driven reason for the same phenomena.

The model in this paper does not feature efficiency wages, and labor markets are competitive. Nonetheless, increasing the minimum wage increases the productivity of low-wage workers. This improvement in productivity comes from the reorganization effect. Increasing the minimum wage induces firms to reduce the labor they demand from minimum wage workers. They do this by first shifting the tasks minimum wage workers are least good at to other workers. This increases specialization among minimum wage workers that are employed, and thus raises their productivity. Notably, even though minimum wage workers become more productive in an absolute sense, they become less productive relative to other workers given the prevailing wages.

1.8.2 Sales Taxes

New York City is unique in that it levies a 4.5% sales tax on certain services, including those performed at hair salons. This section studies the effect of eliminating this sales tax. Formally, I estimate a new equilibrium with $\tau^{NEW} = 0$. The wages in this new equilibrium are provided in Table 1.20.[33] I first discuss the reallocation and reorganization effects of the policy. I then analyze the overall impact of the policy, using the reallocation and reorganization effects to understand their driving mechanisms.

1.8.2.1 Reallocation Effect

Eliminating the sales tax confers a competitive advantage on firms producing high-quality services in the initial equilibrium. Since salons with low organization costs tend to produce high-quality services, eliminating the sales tax reallocates labor towards organizationally efficient firms, as seen in Figure 1.20. These firms produce high-quality services using a more task-specialized internal structure. Thus the reallocation effect increases market-wide task specialization because more workers are working at task-specialized firms.

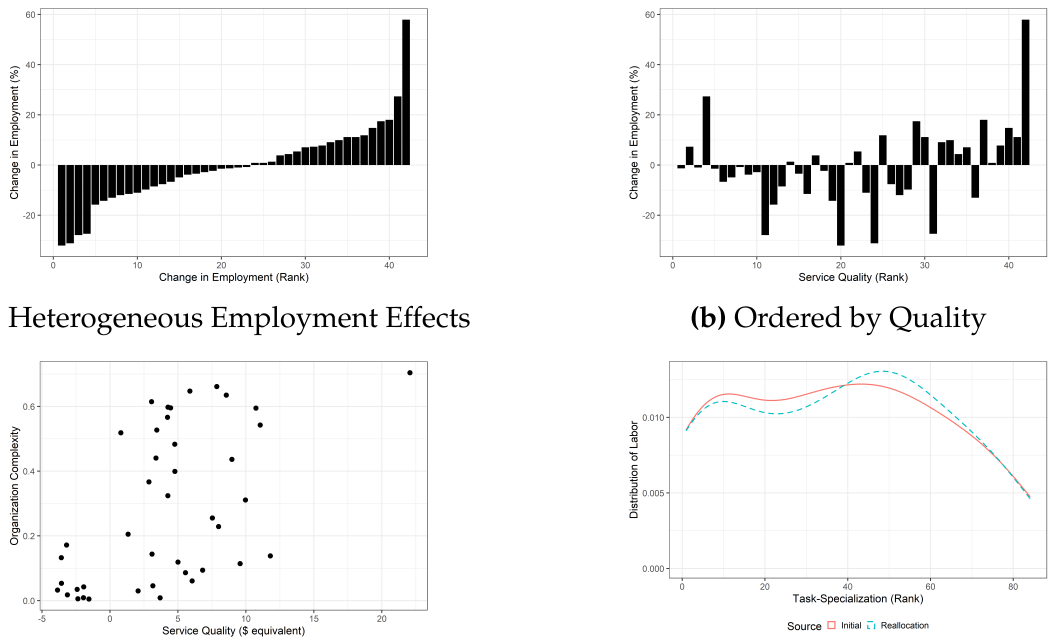
1.8.2.2 Reorganization Effect

Eliminating the sales tax makes producing higher-quality products more attractive. In order to produce higher-quality products, firms choose internal organizations which are on average 5.5% more complex, increasing average labor market task specialization by 0.9%. In terms of the three-way trade-off introduced in Figure 1.9, eliminating the sales tax has the same effect as reducing consumer price sensitivity (ρ). Average firm service quality rises by 10%. This is consistent with the quality-complexity-wage three-way trade-off discussion in the theoretical section.

Figure 1.21 illustrates that these market-wide patterns also happen at the firm level. However, the extent to which salons increase quality and increase task specialization depends on the firm's internal organization costs and its particular task mix. Thus, the slopes and lengths of the arrows in Figure 1.21 differ. Changing sales tax, a product market policy, influences what workers do and what workers are paid in the labor market.

33. Employment remains the same before and after the policy because labor supply is inelastic and the minimum wage is assumed not to be binding.

Figure 1.20: Sales Tax Reallocation Effect



(a) Heterogeneous Employment Effects

(b) Ordered by Quality

(c) High-Quality Salons Are Complex

(d) Shift in Distribution of Labor

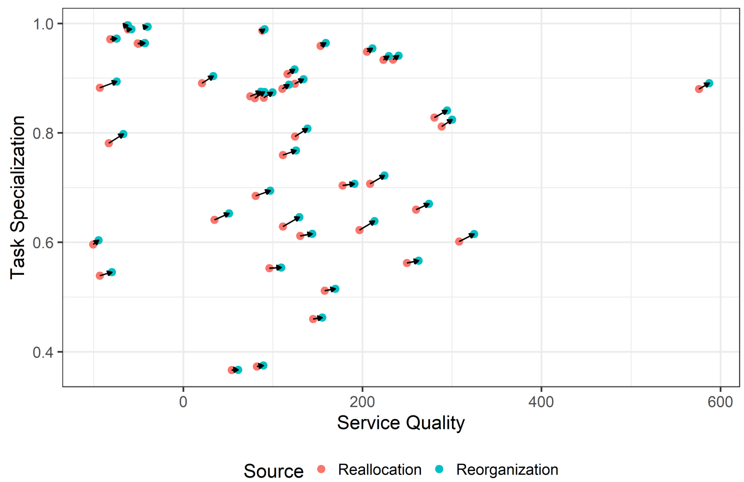
Note: Each bar is a salon. The sales tax elimination decreases employment at some salons and increases it at others (Panel A). Salons with high-quality service see an improvement in their competitive position (Panel B). These salons have a complex, task-specialized internal structure (Panel C). As a result, labor is reallocated to task-specialized firms, and workers become more specialized in equilibrium (Panel D).

1.8.2.3 Total Impact

Table 1.12a summarizes the effect of the policy on wages and task specialization. All worker types see wage increases and task-specialization increases. Wage increases are not proportional to task-specialization increases: even though blow-dry specialists see the largest increase in specialization, they see the lowest increase in wages. This is because the size of wage increases is partly driven by how the policy impacts the competitive position of firms.

The welfare effects of the policy are summarized in Table 1.12b. Overall, eliminating the sales tax leads to a small welfare increase, of 0.19%. However, the effects are quite different for different actors in the model. Firms respond to the sales tax elimination by increasing quality by 10%. Firms capture the surplus from improved quality and reduced taxes from consumers by raising prices by 8.7%. Firm profit increases by a modest 0.58%

Figure 1.21: Reorganization Effect Under a Sales Tax



Note: Each pair of dots connected by an arrow represents a firm, with red representing the firm before the sales tax and blue representing the firm after the sales tax. The direction of the arrows indicates that most salons increase quality by raising task specialization internally. The magnitude of this change (given by the length and angle of the arrow) depends on the firm's particular organization costs and task mixture.

because workers capture most of the surplus from firms through higher wages, which rise by a dollar amount that is comparable to the total lost tax revenue. This is consistent with workers capturing almost all of the productivity improvements from increased task specialization.

Eliminating the sales tax reduces consumer welfare. Why does this occur? In the model, salons can control only two aspects of their products: prices and vertical quality. In the reallocation equilibrium I hold fixed quality and allow only price adjustment. I see that prices rise by 4.7%, as salons both pass on higher wage costs to consumers and increase markups. This reduces consumer welfare by 0.18%. The remaining 0.57% of lost consumer welfare is due to quality over-provision. When the sales tax is eliminated, salons reorganize to increase quality. Reorganization increases organization costs and increases wages, which salons pass on to consumers via higher prices. Consumers would prefer a cheaper, lower-quality product. Quality over-provision in imperfectly competitive markets is not common but can occur.[34]

34. Crawford, Shcherbakov, and Shum (1997) present a case where it occurs in cable television markets.

1.9 Discussion

This paper provides a model which, when estimated, allows the researcher to study firms' internal organization in equilibrium. The theoretical section highlights that common forces govern internal organization across firms. It also provides tractable ways to think about the complex choice of organization. The counterfactual exercises emphasize that while common forces govern internal organization, and the researcher can think about these forces tractably within the model, the equilibrium effects of policies are quite rich and depend on market structure. Internal organization provides new mechanisms for policies to change market outcomes, such as wages, prices and task specialization. In the following subsections I discuss the implications of my results for workers and elaborate on ways in which the model can be applied to other contexts.

1.9.1 Implications for Workers

An area for future work is the welfare impact of task specialization on workers. This paper shows that product and labor market policies change the task composition of jobs in an industry. Minimum wage increases raise specialization for some workers and lower it for others, while cutting the sales tax increases specialization for all workers. What are the welfare implications of these changes for workers?

This is ultimately an empirical question. On the one hand, greater task specialization could deepen worker experience in certain tasks, improving the production possibilities when workers bring their skills together. Greater task specialization also means workers are less exchangeable with other workers at the same firm. On the other hand, greater task specialization may limit task exploration. It may also make jobs more skill-specific, making it more difficult for a worker to find a new job and limiting a worker's outside options. Workers may also intrinsically value generalized jobs more than specialized jobs.

The relative importance of these forces depends crucially on the size of job search frictions, the nature of task-based human capital accumulation, and the amenity value of specialization. Thus, understanding the effect of minimum wage increases and sales tax decreases on worker welfare requires extending the model and better data on wages and employee outcomes.

1.9.2 Model Generality

Although the main empirical application in this paper is hair salons, the model can be used to study internal organization in other settings. In many cases, such as restaurants and hotels, the model can be applied as is. In other settings, such as manufacturing, the model can be adjusted while preserving the core features. The flexibility of the model makes it ideal for studying questions like the adoption of information and communication technology, immigration and robotics.

Appendix Section 1.11.12.2 shows how to adjust the model to accommodate quantity-based (rather than quality-based) productivity, continuous task spaces, labor market power and more sophisticated demand systems. I discuss the complications with some of these extensions and some potential ways to get around these complications. In addition to proving the usefulness of the model, the extensions highlight that the central insights of the model are robust.

A key assumption throughout the paper is the quality-price index restriction on demand. This assumption is most clearly violated by pure vertical demand systems, where consumers agree perfectly about the ranking of products and differ only in their willingness to pay for quality. While hair salons can be described well without pure vertical models of demand, other industries, such as grocery stores, can not be. Indeed, Ellickson (2007) uses hair salons as an example of an industry where pure vertical differentiation is largely absent, in order to show why this differentiation is so important for understanding the grocery store industry. Because the internal organization component of the model could be very helpful in understanding grocery stores and similar industries, in Appendix Section 1.11.12.2 I provide preliminary steps towards relaxing the index restriction.

When estimating the model for Manhattan hair salons, I restrict the worker type space to five specialists (horizontal types). This restriction is not theoretical: the results in the first part of the paper hold under any worker-type space, including a space where one worker has higher skills at all tasks than another (vertical types). Rather it is empirical: I restrict the type space because I do not have data on wages or worker demographics. If I had such data, the model could be estimated with a richer type space.

1.10 Conclusion

This paper studies how internal organization decisions within firms interact with markets outside firms. I develop a structural model, grounded in a set of stylized facts, which allows firms to differ in their internal organization and to change it in response to market conditions. Workers have multidimensional skills and different wages. Internal organization matters because the match between workers and tasks determines wage costs and product qualities. Firms in the same market choose different internal organizations because firms vary both in their ability to internally organize and in their task-based production functions.

The model allows me to look inside the black box of the firm and understand how organizational decisions are made in equilibrium. Firms face a trade-off wherein they want to design the simplest organization that achieves the lowest wage cost and the highest product quality. In equilibrium, the aggregate assignment of workers to tasks is determined both by worker skills and the internal organization decisions of many competing firms. Despite the richness of the model, I am able to identify and estimate it using data on hair-salon task assignments. This allows me to analyze policy counterfactuals.

The counterfactual exercises illustrate that allowing internal organization to be endogenous and heterogeneous qualitatively changes the impact of policy. Minimum wage increases generate new types of wage spillovers that cannot occur in many other models of the labor market. Sales tax cuts induce firms to reorganize their workforce, changing the task composition of jobs. Although these effects are specific to the salon industry, they indicate that internal organization is an important force that deserves careful study in a variety of contexts.

The framework provided in this paper provides a starting point for researchers to do exactly this. The approach in this paper can be extended to accommodate quantity-based (rather than quality-based) productivity, continuous task spaces, labor market power and more sophisticated demand systems. These extensions, combined with traditional employer-employee matched data will be important to answer future questions, including the effect of internal organization on human capital accumulation and the welfare implications of endogenous task assignment for workers.

Tables

Table 1.1: Regressions of Worker Specialization on Organization Complexity

Dependent Variable:	Worker Task Specialization		
Model:	(1)	(2)	(3)
Variables			
Organization Complexity	0.2853***	0.2862***	0.2922***
	(0.0313)	(0.0310)	(0.0392)
Fixed-effects			
Quarter-Year		Yes	Yes
County			Yes
Observations	62,452	62,452	62,452
R^2	0.10184	0.10901	0.21483

Standard errors clustered at the salon level.
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05*

Note: Task specialization is measured as the maximum fraction of time spent on a single task by a worker. Complexity is measured at the salon level. Across all specifications, complexity (a salon-level measure) can account for 10% of the variation in worker specialization.

Table 1.2: Regressions of Salon Size on Organization Complexity

Dependent Variables:	Revenue	Employees	Utilized Labor	Customers	Visits
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Org. Complexity	347549.2***	9.75**	26481	334.6	731.7
	(79546.2)	(3.016)	(35653.2)	(259.6)	(450.1)
Fixed-effects					
Quarter-Year	Yes	Yes	Yes	Yes	Yes
County	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	4,558	4,558	4,558	4,558	4,558
R^2	0.32465	0.34319	0.28918	0.34901	0.35004

Standard-errors clustered at the salon level.
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05*

Note: Observations are salon-quarters. Regressions illustrate a positive correlation between complexity and several measures of salon size after controlling for county and quarter fixed effects and the composition of tasks performed at the salon in the quarter.

Table 1.3: Salon Activity Data Sample

Firm	Salon	App.	Cust.	Service	Staff	Time Stamp	Price	Duration
1	1A	123	Blake	Advanced Cut	Rosy	3/26/2021 16:15	100	72
1	1A	123	Blake	Full Head - Highlights	Rosy	3/26/2021 16:15	243	127
1	1A	123	Blake	Treatment Add On (Olaplex)	Rosy	3/26/2021 16:15	39	72
2	2A	9982	Grace	Women's Cut	Tyler	3/17/2021 11:00	225	43
2	2A	9982	Grace	Single Process	Ben	3/17/2021 11:00	200	77

Note: This table is a snapshot displaying two actual appointments at salons in the same zip code from the data used for the estimation. Customer IDs are replaced by pseudonyms.

Table 1.4: Summary Statistics for All Salon-Quarters

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Revenue	4,558	213,201.30	248,359.90	5	58,912.5	271,236.5	2,559,703
Price	4,558	199.73	135.16	0.20	111.71	261.88	3,180.44
Employees	4,558	13.38	10.79	1	6	17	92
Customers	4,558	1,159.23	1,098.45	1	397	1,619	16,768
Task Categories	4,558	4.45	0.86	1	4	5	5
Labor per. Customer	4,558	2.15	1.63	0.10	1.52	2.57	61.33

Note: The table displays summary statistics for the main variables of interest with data aggregated at the salon-quarter level. There is significant variation across salons in complexity, number of employees, revenue and many other dimensions.

1.11 Appendix

1.11.1 Rate Distortion and Rational Inattention Equivalence

Equation (1.3) from Theorem 1 can be rewritten as

$$\gamma_j \min_{B_j \in \mathbb{B}} \left\{ I(B_j) + \gamma_j^{-1} \Big[W(B_j) - \rho^{-1}\xi(B_j) \Big] \right\}. \tag{1.11}$$

I can rewrite (1.11) as a maximization problem:

$$\max_{B_j \in \mathbb{B}} \left\{ \sum_{i,k} B_j(i,k)(\rho^{-1}\theta_{i,k} - W_i) \Big] - \gamma_j I(B_j) \right\}. \tag{1.12}$$

Comparing (1.12) to formulations in papers such as Jung et al. (2019) illustrates that this is a rational inattention problem with mutual information attention costs. I rewrite Equation

Table 1.5: Parameter Estimates, Tasks

Task	Associated Specialist Skill Gap	Associated Specialist Wage	Skill Base	Material Cost
Administrative	43.29*	26.99	-16.16	-147.60*
	(21.66)	(63.75)	(14.58)	(13.47)
Blowdry/Etc.	141.69*	20.91	-70.56*	12.39
	(36.67)	(40.22)	(13.57)	(16.65)
Color/Highlight/Wash	60.03*	37.75*	-9.69	56.49*
	(21.24)	(7.00)	(11.97)	(15.79)
Haircut/Shave	32.45*	16.96*	.	.
	(13.07)	(8.32)	.	.
Nail/Spa/Eye/Misc.	66.48	81.16	-252.58*	-1061.12*
	(37.72)	(53.52)	(11.47)	(10.73)

Note: Standard errors from 500 bootstrap replications in parentheses; * indicates significance at the 0.05 level. For each task, the table lists the skill gap and wage of the associated specialist in 2021 dollars. The skill gap is the change in quality when a task is assigned to a specialist. Also listed are the skill base, the quality when the task is performed by a non-specialist, the material cost, and the non-wage costs associated with the task (e.g., dye for coloring). Material costs and skill base are relative to the haircut task. Wages are per hour, while material costs and skills are per unit.

Table 1.6: Parameter Estimates, Other

Parameter	Estimate
Price Sensitivity	0.04*
	(0.01)
Cost Intercept	27.95
	(15.21)
Utility Intercept	-24.77*
	(8.36)

Note: Standard errors from 500 bootstrap replications in parentheses; * indicates significance at the 0.05 level. Consumer price sensitivity (ρ) is the main determinant of demand elasticities.

1.11 one last time:

$$\gamma_j \min_{B_j \in \mathbb{B}} \left\{ I(B_j) + \gamma_j^{-1} \sum_{i,k} B_j(i,k)(W_i - \rho^{-1}\theta_{i,k}) \Big] \right\}. \tag{1.13}$$

Comparing Equation (1.13) to formulations such as Equation 6 in Tishby, Pereira, and Bialek (2000) demonstrates this is a well-understood minimization problem from information theory called a rate-distortion problem.

Table 1.7: Model Validation: Estimated vs. Observed Job Task Content

(a) Mean and Median

Task	Mean		Median	
	Model	Observed	Model	Observed
Haircut/Shave	0.4094	0.4094	0.2816	0.3357
Color/Highlight/Wash	0.4058	0.4058	0.3067	0.4042
Blowdry/Style/Treatment/Extension	0.1179	0.1179	0.0162	0.0704
Administrative	0.0278	0.0278	0.0050	0.0040
Nail/Spa/Eye/Misc.	0.0391	0.0391	0.0049	0.0000

(b) Variance

Task	Total Variance		Between Firm Variance	
	Model	Observed	Model	Observed
Haircut/Shave	0.1110	0.1268	0.0597	0.0597
Color/Highlight/Wash	0.1127	0.1105	0.0365	0.0365
Blowdry/Style/Treatment/Extension	0.0472	0.0194	0.0111	0.0111
Administrative	0.0098	0.0080	0.0063	0.0063
Nail/Spa/Eye/Misc.	0.0120	0.0171	0.0050	0.0050

(c) Interquartile Range

Task	p25		p75	
	Model	Observed	Model	Observed
Haircut/Shave	0.1583	0.0469	0.8013	0.7577
Color/Highlight/Wash	0.0417	0.0388	0.7020	0.6383
Blowdry/Style/Treatment/Extension	0.0004	0.0110	0.0726	0.1892
Administrative	0.0027	0.0000	0.0166	0.0108
Nail/Spa/Eye/Misc.	0.0000	0.0000	0.0329	0.0106

Note: The table compares model-generated and observed job task content along several dimensions. The model is designed to exactly match the average market-wide amount of time spent on each task and the between-firm variance. The other moments were not targeted, and assessing their match serves as a validation exercise.

Table 1.8: Total Effects of Increasing the Minimum Wage

(a) Wage Changes by Worker Type

Type	Wage Change	Total Wages Gained/Lost
Haircut/Shave - UNEMPLOYED	-100.00%	-$600,240
Haircut/Shave - EMPLOYED	17.95%	$1,528,205
Color/Highlight/Wash	-0.61%	-$228,453
Blowdry/Style/Treatment/Extension	3.48%	$323,374
Administrative	4.17%	$47,154
Nail/Spa/Eye/Misc.	0.68%	$19,319

(b) Welfare Breakdown

Source	Change	Percent Change
Salon Profit	-$714,413	-0.472%
Consumer Welfare	-$2,528,784	-1.671%
Employed Wages	$1,689,600	1.116%
Unemployed Wages	-$600,240	-0.397%
Total Welfare	-$2,153,838	-1.423%

Note: Increasing the minimum wage generates both positive and negative wage spillovers for workers on whom it is not binding. Positive spillovers are larger and occur for most worker types. Overall, wage increases for employed workers are more than salon profit losses and wage losses of unemployed workers combined. Total welfare declines, as consumers see higher prices and slightly lower quality.

1.11.2 Proof of Theorem 1

For any given organization structure, the firm will choose prices only weakly above marginal cost; otherwise, it receives negative profit. Without loss, I therefore restrict the set of price-structure pairs considered to be those where price exceeds marginal cost.

First, I prove that if an organization structure B_j^* solves the simpler problem (Equation 1.3), then it is profit-maximizing ("only if" direction). I need to show that for any price-organization structure pair (p_j', B_j') there exists p such that profit under (p_j, B_j^*) is weakly higher than profit under (p_j', B_j'). I do this by construction. Denote B_j^* as a structure which solves Equation (1.3). Such a structure always exists because Equation (1.3) is a

Table 1.9: Spillovers from an Increase in the Minimum Wage

(a) Reallocation Effect

Type	Reallocation Change		
	Employment	Task-Spec.	Wage
Haircut/Shave	-5.85%	-0.04%	17.95%
Color/Highlight/Wash	0%	-0.17%	-1.13%
Blowdry/Style/Treatment/Extension	0%	-0.40%	4.63%
Administrative	0%	0.09%	5.22%
Nail/Spa/Eye/Misc.	0%	-0.03%	0.58%

(b) Reorganization Effect

Type	Reorganization Change		
	Employment	Task-Spec.	Wage
Haircut/Shave	-0.73%	0.12%	0%
Color/Highlight/Wash	0%	-0.33%	0.52%
Blowdry/Style/Treatment/Extension	0%	0.03%	-1.15%
Administrative	0%	0.03%	-1.05%
Nail/Spa/Eye/Misc.	0%	-0.00%	0.10%

Note: The minimum wage increase has positive spillovers for some workers and negative spillovers for others. These spillovers can be further decomposed to those resulting from reorganization and those resulting from reallocation. Most spillovers come from the fact that the policy favors salons that have internal organizations intense in binding workers initially (reallocation). Some spillovers occur because the policy induces firms to shift tasks from binding to non-binding workers (reorganization).

Table 1.10: Summary of All Minimum Wage Increase Effects

Statistic	Reallocation	Reorganization	Total
Avg. Price	1.96%	-0.29%	1.67%
Avg. Complexity	0.00%	-0.46%	-0.46%
Avg. Quality	0.00%	-0.54%	-0.54%
Avg. Hourly Wage	3.40%	0.20%	3.60%
Std. Dev. Wage	-8.91%	1.03%	-7.88%
Task Specialization	-0.61%	-0.18%	-0.79%
Employment	-1.53%	-0.19%	-1.72%
Market Served	-2.69%	-0.12%	-2.81%
Total Profit	-2.69%	-0.12%	-2.81%
Consumer Welfare	-2.64%	-1.19%	-3.83%
Total Wages	1.81%	0.00%	1.82%
Total Welfare	-0.88%	-0.54%	-1.42%

Note: This table summarizes the impact of increasing the minimum wage from \$15 to \$20 on different actions and market outcomes in the Manhattan hair-salon market.

rate-distortion/rational inattention problem, as shown in Appendix Section 1.11.1.

For any price p'_j and any structure B'_j, I can construct $p_j = p'_j + \gamma_j I(B^*_j) + W(B^*_j) - \gamma_j I(B'_j) - W(B'_j)$. The price p_j is positive and therefore feasible. Recall that profit evaluated at (p_j, B^*_j) is

$$D_j(\xi(B^*_j) - \rho p_j, p_{-j}, \xi_{-j})\Big[p_j - \gamma_j I(B^*_j) - W(B^*_j)\Big].$$

The second multiplicative term of profit is equal under (p_j, B^*_j) and (p'_j, B'_j). The first term (demand) is strictly increasing in the quality-price index $\xi(B_j) - \rho p_j$; therefore, it is sufficient to show that this index is weakly higher for (p_j, B^*_j). I show this by rewriting

Table 1.11: Summary of All Sales-Tax-Elimination Effects

Statistic	Reallocation	Reorganization	Total
Avg. Price	4.70%	3.99%	8.68%
Avg. Complexity	0.00%	5.53%	5.53%
Avg. Quality	0.00%	10.03%	10.03%
Avg. Hourly Wage	18.32%	1.02%	19.34%
Std. Dev. Wage	22.67%	-6.32%	16.35%
Task Specialization	0.90%	0.93%	1.83%
Total Profit	4.32%	-0.60%	3.71%
Consumer Welfare	-0.18%	-0.57%	-0.75%
Total Wages	18.32%	1.02%	19.34%
Total Welfare	0.14%	0.05%	0.19%

Note: This table summarizes the impact of eliminating the service sales tax on different actions and market outcomes in the Manhattan hair-salon market.

Table 1.12: Total Effects of a Sales-Tax Elimination

(a) Wage Changes by Worker Type

Type	Wage Change	Task-Spec. Change
Haircut/Shave	31.99%	0.29%
Color/Highlight/Wash	20.09%	2.57%
Blowdry/Style/Treatment/Extension	6.06%	3.01%
Administrative	17.99%	1.03%
Nail/Spa/Eye/Misc.	12.74%	2.39%

(b) Welfare Breakdown

Source	Change	Percent Change
Salon Profit	\$942,740	0.58%
Consumer Welfare	-\$494,199	-0.30%
Wages	\$11,603,777	7.12%
Tax Revenue	-\$11,739,300	-7.20%
Total Welfare	\$313,017	0.19%

Note: Eliminating the sales tax raises wages most in percentage terms for haircut specialists. Workers gain the most from the elimination of the sales tax: wage increases are almost equal to the lost revenue to the government.

$\xi(B_j^*) - \rho p_j$:

$$= \xi(B_j^*) - \rho[p_j' + \gamma_j I(B_j^*) + W(B_j^*) - \gamma_j I(B_j') - W(B_j')] \tag{1.14}$$

$$= \xi(B_j^*) - \rho[p_j' + \gamma_j I(B_j^*) + W(B_j^*) - \gamma_j I(B_j') - W(B_j')] + \xi(B_j') - \xi(B_j') \tag{1.15}$$

$$= \xi(B_j') - \rho[p_j' + \gamma_j I(B_j^*) + W(B_j^*) - \gamma_j I(B_j') - W(B_j') - \rho^{-1}\xi(B_j^*) + \rho^{-1}\xi(B_j')] \tag{1.16}$$

$$= \xi(B_j') - \rho p_j' - \rho\underbrace{[\gamma_j I(B_j^*) + W(B_j^*) - \rho^{-1}\xi(B_j^*) - \{\gamma_j I(B_j') + W(B_j') - \rho^{-1}\xi(B_j')\}]}_{\leq 0 \text{ because } B_j^* \text{ minimizes}} \tag{1.17}$$

$$\geq \xi(B_j') - \rho p_j'. \tag{1.18}$$

This proves the "only if" direction. I now prove that if a structure B_j^* is profit maximizing, it solves Equation (1.3) (the "if" direction). Suppose for sake of contradiction there exists B'_j which is profit maximizing but does not solve Equation (1.3). Then, as in the first part of the proof, there exists B^*_j which does solve Equation (1.3). Then I can construct p_j as before for any p'_j that is weakly higher than marginal cost under B'_j. However, because B'_j does not minimize Equation (1.3), $\xi(B^*_j) - \rho p_j > \xi(B'_j) - \rho p'_j$, and thus profit is strictly higher under B^*_j, p_j. This contradicts optimality of B'_j and concludes the proof.

1.11.3 Proof of Proposition 1 and 2

I have already shown in Theorem 1 that optimal B solves a rate-distortion problem.

- Denote by Q the quality-adjusted wages. Denote by $I^*(Q)$ the optimal complexity as a function of quality-adjusted wages.
- RD equivalence $\implies$ $I^*(Q)$ is continuous, convex and decreasing. It is also strictly decreasing above some threshold $\bar{Q}$ (Chen, n.d.).
- The firm's choice of quality-adjusted wages solves

$$V := \min_{Q} \gamma I^*(Q) + Q.$$

- The envelope theorem implies the index and thus profit are increasing in γ:

$$\frac{\partial V}{\partial \gamma} = I^*(Q) \geq 0.$$

- Examine the FOC:

$$\frac{dI^*(Q) + \gamma^{-1} Q}{dQ} = \frac{dI^*(Q)}{dQ} + \gamma^{-1} = 0 \implies \frac{dI^*(Q)}{dQ} = -\gamma^{-1}.$$

- Because I^* is decreasing and convex, its derivative is negative and increasing.
- Therefore, Q which solves is increasing in γ.

- Thus profit and complexity will be positively correlated via γ.

1.11.4 Optimal Jobs Within the Firm

The last result shows the originally high-dimensional problem of the firm can be reduced to a tractable two-dimensional trade-off. However, one of the goals of the model is to understand how firms assign workers to tasks. This section describes the properties of task assignments within the firm and shows that the firm customizes the bundles of tasks it assigns individual worker types. For this, I define the job of worker type i at firm j as a vector $(b_j(i,\cdot))$, where element k denotes the amount of i's time spent on task k. The jobs at a firm are the rows of the organization structure divided by the total labor of worker type i:

$$b_j(i,k) = \frac{B_j(i,k)}{\sum_{k'} B_j(i,k')}.$$

Proposition 5 *The profit-maximizing organizational structure satisfies the following properties.*

1. ***Law of Demand:*** *The share of workers of type* i *(*$E_j(i)$*) decreases as their wage increases.*
2. ***Incomplete Specialization:*** *All hired worker types spend a positive amount of time on each task whenever* $\gamma_j > 0$*.*
3. ***Optimal Jobs:*** *Jobs take the following logit-like form:*

$$b_j(i,k) = \alpha_k \frac{exp(-\gamma^{-1}w_i + (\rho\gamma)^{-1}\theta_{i,k}))}{\sum_{i'} E_j(i')exp(-\gamma^{-1}w_{i'} + (\rho\gamma)^{-1}\theta_{i',k}))}.$$

I prove this result by appealing to the rational inattention literature. I derive the expression for optimal jobs by manipulating the first-order conditions and the constraints. The proof is provided in Appendix Section 1.11.4.1. Even though at a high level the firms are trading off complexity and quality-adjusted wages, under the surface, they customize jobs for individual workers and tasks. The proposition illustrates that task assignments depend on skills through $\theta_{i,k}$, wages through w_i, consumer price sensitivity through ρ, the task mix through α_k, and organization costs through γ_j. This proposition highlights two important features of the model. First, whenever there are some organizational frictions within a firm, complete specialization will not occur. Every "job" will be a bundle of multiple tasks. Second, because jobs depend on organization costs, where someone works matters for what they do. That is, two identical workers will not perform the same tasks even in the same product and labor market. The tasks included in any job will depend on the firm where a worker is employed.

1.11.4.1 Proof of Proposition 5

For the purposes of this proof only, I define $h_{i,k}$ as the fraction of task k performed by worker i. Then the optimal job of worker i is given by

$$h_{i,k} = \frac{E_i}{Z(k,\lambda)} exp\Big(- \lambda(\rho w_i - \theta_{i,k})\Big).$$

Summing over i yields

$$\sum_i h_{i,k} = \frac{1}{Z(k,\lambda)} \sum_i E_i exp\Big(- \lambda(\rho w_i - \theta_{i,k})\Big) = 1.$$

Therefore,

$$Z(k,\lambda) = \sum_i E_i exp(-\lambda(\rho w_i - \theta_i \delta^{\mathbb{I}\{\kappa_i \neq k\}}))$$

and

$$h_{i,k} = \frac{e_i exp(\lambda(-\rho w_i + \theta_{i,k}))}{\sum_{i'} e_{i'} exp(\lambda(-\rho w_{i'} + \theta_{i,k}))}.$$

Substituting for λ yields

$$h_{i,k} = \frac{E_i exp(-\gamma^{-1} w_i + (\rho\gamma)^{-1}\theta_{i,k}))}{\sum_{i'} E_{i'} exp(-\gamma^{-1} w_{i'} + (\rho\gamma)^{-1}\theta_{i',k}))}.$$

By the definition of $h_{i,k}$,

$$B_{i,k} = \alpha_k h_{i,k}.$$

To get to jobs, I divide by E_i:

$$b_{i,k} = B_{i,k}/E_i = \alpha_k/E_i h_{i,k} = \frac{\alpha_k exp(-\gamma^{-1}w_i + (\rho\gamma)^{-1}\theta_{i,k}))}{\sum_{i'} E_{i'} exp(-\gamma^{-1}w_{i'} + (\rho\gamma)^{-1}\theta_{i',k}))}.$$

This illustrates that optimal jobs take a multinomial logit form. I can also derive this result by applying Theorem 1 from Matêjka and McKay (2015).

The fact that all hired worker types spend a positive amount of time on each task is a direct application of Lemma 1 from Jung et al. (2019). An increase in wage corresponds to a decrease in the "payoff" to the firm of using workers of type i in all tasks (i.e., states of the world in the rational inattention literature). This means I can apply Proposition 3 from Matêjka and McKay (2015) to say that an increase in w_i leads to a decrease in E_i all else constant. I can even say that E_i is strictly decreasing in w_i whenever the initial share of worker i is strictly interior, i.e., $0 < E_i < 1$.

1.11.5 Proof of Proposition 3

To recover the best responses of the firm's problem, I use the fact that the joint maximization of any function is equivalent to the sequential maximization. Thus I can write the firm's problem as

$$\max_{B_j \in \mathbb{B}} \max_{p_j \in \mathbb{R}_+} \underbrace{\frac{exp(\overbrace{\xi(B_j)}^{\text{quality}} - \rho p_j)}{\sum_{j'} exp(\xi(B_{j'}) - \rho p_{j'})}}_{\text{market share, } s_j} \left[p_j - \underbrace{\left(\overbrace{\gamma_j I(B_j)}^{\text{org.}} + \overbrace{\sum_{i,k} w_i B_j(i,k)}^{\text{avg. wage}} \right)}_{\text{constant marginal cost, } MC_j} \right].$$

I first study the inner pricing problem. Fixing an organization structure, the model reduces to a logit Bertrand game with heterogeneous costs and qualities. Proposition 7 of Caplin and Nalebuff (1991) proves that such a game has a unique pure-strategy Nash equilibrium in prices. Therefore, for any chosen organizational structure, there is a single best-response price. In the proof of Theorem 1, I substituted the equation characterizing the optimal price into profit, and showed that the best response B_j also solves

$$\min_{B_j \in \mathbb{B}} I(B_j) + \gamma_j^{-1} \sum_{i,k} B_j(i,k)(W_i - \rho^{-1}\theta_{i,k}).$$

The best-response structure will therefore depend on other actions of the firm only through wages. The theorem also establishes that this is equivalent to a rational inattention problem with a mutual information cost function. With the equivalence to a rational inattention problem, I can establish existence. I can then appeal to Matêjka and McKay (2015) to say that there exists an organization structure which maximizes profit for each firm. This establishes equilibrium existence.

To obtain uniqueness, note that a rational inattention problem with mutual information costs is a special case of the problems considered by Lipnowski and Ravid (2022). A stochastic choice rule in their language is an organization structure in mine. Proposition 1 of their paper implies that if γ_j is known, the set of quality-adjusted wages which generate multiple organization structures is "meager and shy." Since I consider the case of finite tasks (finite Ω in their language), "meager and shy" implies a null set. This is only for one firm with a specific γ_j. The set of quality-adjusted wages which generate multiple profit-maximizing organization structures for at least one firm will be the union of all sets which generate multiplicity for each individual firm. The union of countable null sets is also null; therefore, the set of quality-adjusted wages that generate multiplicity is null.

Denote the set of quality-adjusted wages which generate multiplicity as $\mathbb{M}$. The mapping from market parameters Ω to quality-adjusted wages is defined by a multivariate, vector-valued function $F : \mathbb{R}_+^{N\times K+N+1} \to \mathbb{R}_+^{N\times K}$. It can be shown that if F is smooth and the rank of the Jacobian of F is at least $N \times K$, then the measure of the pre-image of any measure 0 set is 0.

I now prove that F satisfies the rank condition. Recall that the quality-adjusted wage of worker i and task k has the form $w_i - \rho^{-1}\theta_{i,k}$. Collapse i, k into a single index, $y = 1, ..., N \times K$, where $\mathcal{I}(\cdot)$ and $\mathcal{K}(\cdot)$ return the task and worker type associated with the index y. Then that element y of F is $F(\Omega) = w_{\mathcal{I}(y)} - \rho^{-1}\theta_{\mathcal{I}(y),\mathcal{K}(y)}$. The Jacobian of this function has a rank of at least $N \times K$ because each skill parameter $\theta_{i,k}$ impacts only one quality-adjusted wage. Formally, there exist at least $N \times K$ columns of the Jacobian which are linearly independent of each other. Thus the pre-image of the null set $\mathbb{M}$ on F will be measure 0. Since the pre-image is the set of parameters which generate multiplicity, the set of parameters which cause at lease one firm to have multiple profit-maximizing organization structures is measure 0.

Whenever all firms have a unique organization structure, they also have a unique cost and quality. It remains to be shown that equilibrium prices are also unique. To do this, I appeal to Caplin and Nalebuff (1991) and note that demand is multinomial logit, so whenever organization structures are unique, so are Nash equilibrium prices.

An implication of this result is that a pure-strategy Nash equilibrium exists for any fixed wages. I conjecture that this proof could be extended to show equilibrium existence in the full model, that is, when wages are determined by market clearing. One approach would be to prove that excess labor demands satisfy Kakutani's fixed point theorem. Extending the uniqueness result to say that the equilibrium is unique for almost any total labor supplies may not be possible. This is because, in general, worker types may be complements or substitutes depending on their skill sets. If firms are homogeneous with respect to task mixtures and organization costs, the wages that clear the market may very well be the wages which induce indifference across multiple organization structures and multiple equilibria.

1.11.6 Proof of Proposition 4

For simplicity, firm index j is suppressed throughout this section. I denote by $I(\tilde{B})$ the organization complexity based on worker identities. This is observed in the task assignment data. I denote by $I(B)$ the organization complexity based on worker skill sets. This is unobserved. I denote by $I^*(\gamma)$ the firm's complexity predicted by the model, where market parameters Ω and the task mix α are assumed to be known and thus are incorporated into the function and not left as arguments.

First, I prove that observed organization complexity based on worker identities ($I(B)$)

is equal to unobserved true complexity based on worker skill sets ($I(\tilde{B})$). Consider the augmented model proposed in Section 1.6.1. In particular, recall that workers with different labor supplies match to firms by some unspecified matching process. I then prove the following:

Lemma 2 *All workers with the same skill set are assigned the same distribution of tasks regardless of their labor supply.*

Proof of Lemma. A well-known property of mutual information attention costs is that they satisfy compression monotonicity or are "distraction-free " (Tian 2019). I will use this in the proof.

Suppose for the sake of contradiction the firm assigned two workers of the same skill set different distributions of tasks. Consider a different assignment of work such that the same amount of each task is accomplished, and both workers still are assigned the same total amount of work. Such an assignment always exists: I can just take the total time spent on each task by both workers and split it based on effective units of labor. By the strict distraction-free property of mutual information, this new assignment reduces organization costs. This does not impact the wage bill, since both workers have the same wage. Also, it does not impact quality, because the total amount of each task accomplished remains the same, and both workers have the same skill set. Thus quality-adjusted cost strictly decreases, so profit strictly decreases, contradicting the optimality of the original assignment. Therefore, all workers with the same skill set are assigned the same distribution of tasks regardless of their effective units of labor.■

This lemma means that the firm treats workers with different labor supplies but the same skill sets as if they were a single, aggregate worker. Denote worker identities as indexed by n, and worker skill sets by i. Denote the organizational structure over worker identities as $\tilde{B}$. Then

$$\frac{\tilde{B}_{n,k}}{\sum_{k'} \tilde{B}_{n,k'}} = \frac{B_{i,k}}{\sum_{k'} B_{i,k'}} \forall i, k \text{ s.t. } \theta_n = \theta_i.$$

Because the total amount of each task is fixed at α_k,

$$\sum_{n'} \tilde{B}_{n',k} = \alpha_k = \sum_{i'} B_{i',k}.$$

Plugging these results into organization complexity yields

$$I(\tilde{B}) = \sum_{n,k} \tilde{B}_{n,k} log\left(\frac{\tilde{B}_{n,k}}{\sum_{k'} \tilde{B}_{n,k'} \sum_{n'} \tilde{B}_{n',k}}\right) = \sum_{n,k}\sum_{i} \frac{B_{i,k} \sum_{k'} \tilde{B}_{n,k'}}{\sum_{k'} B_{i,k'}} \mathbb{I}\{\theta_n = \theta_i\} log\left(\frac{B_{i,k}}{\sum_{k'} B_{i,k'} \sum_{i'} B_{i',k}}\right).$$

And rearranging terms yields

$$= \sum_{i,k} \frac{B_{i,k}}{\sum_{k'} B_{i,k'}} log\left(\frac{B_{i,k}}{\sum_{k'} B_{i,k'} \sum_{i'} B_{i',k}}\right) \sum_{n,k'} \tilde{B}_{n,k'} \mathbb{I}\{\theta_n = \theta_i\}.$$

The sum of all $\tilde{B}_{n,k}$ of workers with the same skill set but different labor supply is E_i, which is exactly equal to $\sum_{k'} B_{i,k'}$. Therefore, I can write

$$I(\tilde{B}) = \sum_{i,k} \frac{B_{i,k}}{\sum_{k'} B_{i,k'}} log\left(\frac{B_{i,k}}{\sum_{k'} B_{i,k'} \sum_{i'} B_{i',k}}\right) \sum_{k'} B_{i,k'} = \sum_{i,k} B_{i,k} log\left(\frac{B_{i,k}}{\sum_{k'} B_{i,k'} \sum_{i'} B_{i',k}}\right) = I(B).$$

Therefore, organization complexity based on worker identities is equal to organization complexity based on worker skill sets. Since I observe identities, this implies that I can compute organization complexity as the mutual information between worker identities and tasks.[35]

I next show that γ is identified. This requires that there be a unique γ such that $I^*(\gamma) = I(\tilde{B})$. Define $Q_j := W(B_j) - \rho^{-1}\xi(B_j)$. Applying Theorem 1, I can write the firm's problem in the following way:

$$V := \min_{B \in \mathbb{B}} \gamma I(B) + W(B) - \rho^{-1}\xi(B) = \min_{Q \in \mathbb{Q}} \gamma \tilde{I}(Q) + Q,$$

35. One can also appeal to the data-processing inequality (which holds with equality) to avoid much of this algebra.

where $\tilde{I}$ is a continuous, decreasing and convex function. Further, it is strictly decreasing whenever $\tilde{I}(Q) > 0$ (Chen, n.d.). Consider only the case when $\tilde{I}(Q) > 0$. Then the FOC $\frac{dV}{dQ} = \gamma\frac{d\tilde{I}(Q)}{dQ} + 1 = 0$ and convexity imply the optimally chosen Q is strictly increasing in γ. This implies $\tilde{I}(B)$ is strictly decreasing in γ. Since $\tilde{I}(B) = I^*(\gamma)$ for optimal B, $I^*(\gamma)$ is strictly decreasing and identification is achieved whenever $I(\tilde{B}) > 0$.[36]

Theorem 1 established that the firm's problem is a rate-distortion problem. As a result, Blahut (1972) provides an algorithm that can be used to arbitrarily approximate $I^*(\gamma)$. Thus, because I^* is strictly decreasing, I can use this algorithm to invert complexity to retrieve γ as a known function of complexity and all other parameters.

To identify organization structures (B_j), I appeal to Proposition 3. Since wages are parameters during estimation, the proposition can be applied exactly, and I have that all organization structures are identified except over a set of market parameters with measure 0. Further, the algorithm given in Blahut (1972) constructs optimal B_j for each firm in the process of computing I^*. In the knife-edge cases where more than one structure is optimal for a firm, the algorithm will return one of them. Thus, organization structures are also a known function of the data and market parameters, except for a set of market parameters with measure 0.

1.11.7 Welfare

Preferences take a random utility form with Type 1 extreme value distribution for the horizontal taste heterogeneity $\epsilon_{i,j}$ in the population. I assumed throughout that this heterogeneity is distributed i.i.d. across consumers and alternatives. Therefore, expected utility of consumer i has the well-known closed form

$$V_i = \mathbb{E}[\max_j\{\xi_j - \rho p_j + \epsilon_{i,j}\}] = ln\left[\sum_{j=1}^{J} exp\Big(\xi_j - \rho p_j\Big)\right] + C,$$

36. Whenever complexity is 0 (it cannot be negative), any sufficiently large γ is consistent with the data.

where C is Euler's constant. There is a mass M of consumers; therefore, total consumer expected utility is $M \cdot V_i$.[37] I can then denominate this in dollar terms by dividing by the coefficient on price, ρ. My measure of total consumer welfare in dollar terms is

$$CS = \frac{M}{\rho}\left\{ln\left[\sum_{j=1}^{J} exp\left(\xi_j - \rho p_j\right)\right] + C\right\}.$$

With a sales tax τ, it is

$$CS = \frac{M}{\rho}\left\{ln\left[\sum_{j=1}^{J} exp\left(\xi_j - \rho(1+\tau)p_j\right)\right] + C\right\}.$$

Total welfare is measured as the sum of consumer surplus, firm profits and worker wages. This assumes an additive welfare function which weights all consumers, firms and workers equally.

1.11.8 Organization Complexity as Task Specialization

This section illustrates that complexity is a measure of average task specialization. To see this, first define a job as a vector, where component k is the fraction of a worker's total labor spent performing task k:

$$b_i(k) = \frac{B(i,k)}{E_i}.$$

I can measure the specialization of any job by comparing it to a benchmark "generalist job." I define the generalist job as the job where all workers are assigned exactly the task mix:

$$b_j^G(k) = \alpha_k.$$

Notice that when the firm gives all workers the generalist job, each worker is working as a miniature version of the firm itself. There is no sense in which a worker needs a coworker in order to produce output. With these two concepts in hand, I obtain the following result.

37. This assumes an additive welfare function which gives equal weight to all consumers.

Proposition 6 *Complexity ($I(B_j)$) is the weighted-average Kullback-Leibler divergence between the jobs at a firm and the firm's generalist job $b_j^G(k)$, where the weights are the share of each worker type.*

Proof. Using the definition of mutual information, I can write complexity as

$$
\begin{aligned}
I(B_j) &= \sum_{i,k} B(i,k) log\left(\frac{B(i,k)}{\sum_{k'} B(i,k') \sum_{i'} B(i',k)}\right) \\
&= \sum_{i,k} E_i \frac{B(i,k)}{E_i} log\left(\frac{B(i,k)}{E_i \alpha_k}\right) \\
&= \sum_{i} E_i \sum_{k} b_i(k) log\left(\frac{b_i(k)}{\alpha_k}\right) \\
&= \sum_{i} E_i \sum_{k} b_i(k) log\left(\frac{b_i(k)}{b_j^G(k)}\right) \\
&= \sum_{i} E_i D_{KL}(b_i || b_j^G).
\end{aligned}
$$

1.11.9 Closed-Form Logit Price Expression

Demand for a product j is given by

$$s_j(p_j) = \frac{exp(-\rho p_j + \xi_j)}{\sum_{j'=0}^{J} exp(-\rho p_{j'} + \xi_{j'})}.$$

Optimal pricing in a Bertrand Nash equilibrium with single-product firms is then given by

$$p_j = MC_j + \frac{1}{\rho(1 - s_j(p_j))}.$$

I now follow the arguments laid out in Aravindakshan and Ratchford (2011). I rewrite this expression as

$$p_j = c_j + \frac{1}{\rho(1 - \frac{exp(-\rho p_j + \xi_j)}{exp(-\rho p_j + \xi_j) + \sum_{j' \neq j} exp(-\rho p_{j'} + \xi_{j'})})}.$$

I rewrite it again as

$$p_j = c_j + \frac{1}{\rho} + \frac{exp(-\rho p_j + \xi_j)}{\rho \sum_{j' \neq j} exp(-\rho p_{j'} + \xi_{j'})}.$$

Multiplying by ρ and subtracting ξ_j yields

$$\rho p_j - \xi_j = \rho c_j + 1 + \frac{exp(-\rho p_j + \xi_j)}{\sum_{j' \neq j} exp(-\rho p_{j'} + \xi_{j'})} - \xi_j.$$

Now denote

$$E_j = \sum_{j' \neq j} exp(-\rho p_{j'} + \xi_{j'})$$

$$\frac{exp(-\rho p_j + \xi_j)}{E_j} + \xi_j - \rho p_j = -1 - \rho c_j + \xi_j$$

$$exp\Bigg(\frac{exp(\xi_j - \rho p_j)}{E_j}\Bigg) exp\Bigg(\xi_j - \rho p_j\Bigg) E_j^{-1} = exp\Bigg(-1 + \xi_j - \rho c_j\Bigg) E_j^{-1}$$

and

$$\tilde{W} = exp\Bigg(\xi_j - \rho p_j\Bigg) E_j^{-1}.$$

Then the expression becomes

$$\tilde{W} e^{\tilde{W}} = exp\Bigg(-1 + \xi_j - \rho c_j\Bigg) E_j^{-1}.$$

The left-hand side expression is the form required by Lambert's W, so the $\tilde{W}$ which solves is given by Lambert's W function of the right-hand side by definition. Thus optimal price solves

$$W\Bigg(exp\Bigg(-1 + \xi_j - \rho c_j\Bigg) E_j^{-1}\Bigg) = exp\Bigg(\xi_j - \rho p_j\Bigg) E_j^{-1}.$$

A property of this function is that $log(W(x)) = log(x) - W(x)$. Using this fact yields

$$-1 + \xi_j - \rho c_j - log(E_j) - W\Bigg(exp\Bigg(-1 + \xi_j - \rho c_j\Bigg) E_j^{-1}\Bigg) = \xi_j - \rho p_j - log(E_j),$$

which can be solved for the optimal price:

$$\frac{1}{\rho} + c_j + \rho^{-1} W\Big(exp\Big(-1 + \xi_j - \rho c_j\Big) E_j^{-1}\Big) = p_j^*. \tag{1.19}$$

1.11.10 Other Organization Costs

Theorem 1 does not rely on organization costs taking the mutual information form. However, identification of the structural model relies heavily on organization costs taking this form: it allows me to equate the observed complexity over worker identities to the true complexity over types. This is one of the main reasons why the mutual information functional form is used for organization costs in this paper.

However, imposing mutual information costs ex ante imposes behavioral assumptions on firms. In particular, it makes assumptions about how firms trade off complexity with other concerns, and it imposes symmetry conditions on worker types. Working out the implications of these assumptions for substitution patterns is a matter of consulting the rapidly growing literature on information costs and mapping these results to the labor context. This is a non-trivial task that is beyond the scope of this paper but an area for future work.

Empirically, it would be interesting to identify the correct organization cost function. As Pomatto, Strack, and Tamuz (2022) note, the mutual information cost and the log-likelihood ratio cost imply quite different behavior. However, as Lipnowski and Ravid (2022) suggest, it will be necessary to observe more information about firm choice probabilities to distinguish between different costs. In my setting, this amounts to better information about worker skills (education, demographics, prior experience, etc.). Ideally, such a project would use matched employer-employee data with detailed demographic information alongside task information.

1.11.11 Extensions

Many of the modeling assumptions are made solely to achieve tractability or to match the hair-salon application. The core idea behind the model is general, and this section outlines several extensions which accommodate other contexts and additional economic forces.

1.11.11.1 Labor Market Power

The model presented in this paper focuses on situations where firms have product market power but not labor market power. These assumptions are realistic when the product market is small relative to the labor market, either geographically or because of workers' ability to work in multiple industries. In many situations, such assumptions are not realistic, and one might expect firms to hold labor market power as well.

Introducing labor market power raises an interesting theoretical question which could make it the most important area for future work. Firms with labor market power have an incentive to reduce the number of workers they hire in order to mark down wages. How does this incentive interact with internal organization, and how does it change competition? Unlike firms in a competitive labor market, firms with labor market power will realize that demanding more of a certain type of worker increases wages.

Such an extension has the potential to help us understand two features of modern labor markets. First, we can measure the amenity value of task specialization to workers. In some industries, workers may find a specialized job unfulfilling or limiting, restricting their long-term career goals by pigeonholing them. However, in others, workers may find specialized jobs valuable because they deepen expertise. Second, in highly concentrated industries, we can study how internal organization choices are driven by a desire to make workers scarce for competitors. Anecdotal accounts in the technology sector suggest such talent wars occur. My model provides a way to study the trade-off between over-hiring a certain type of worker and trying to operate the firm.

A model with labor market power could be made tractable by assuming monopolistic competition in the product market and monopsonistic competition in the labor market. The labor market could be modeled using the framework introduced in Card et al. (2018). The novel internal organization cost introduced in this paper extends to such a model. However, because output would impact marginal costs (through wage markdowns), the characterization in Theorem 1 will no longer hold. New tools to solve and estimate the model would be needed. Because the new problem will be a non-linear rational inattention problem, results from Jung et al. (2019) may be helpful in this regard.

1.11.11.2 Large Firms

The model can be extended to the case where firms are "large," with a continuum of tasks and worker types.

Consider a firm which must complete a continuum of tasks to produce the final good. The task mix is now a distribution, which I assume to be normal: $k \sim N(0, \sigma^2)$. Suppose workers have a single specialty task, and that they are indexed by i in the order of their specialty task. An organization structure B is now a continuous bivariate joint distribution.

Suppose the quality of a performed task is given by the squared distance between the specialty of the worker and the task assigned, that is, $\xi = -\int (i-k)^2 dB(i,k)$, and denote $D = -\xi$. For simplicity, assume all workers have the same wages (skills are not priced by the market) and $\rho = 1$.

It can be shown that the organizational frontier in this special case has a closed form, and an organization structure B which maximizes profit is

$$\begin{pmatrix} i \\ k \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 - ln(4)\gamma & \sigma^2 \\ \sigma^2 & \sigma^2 \end{bmatrix} \right).$$

To interpret this result, note that as γ approaches 0, the correlation between tasks and workers approaches 1 and the marginal distribution of hired worker types widens and approaches the distribution of tasks. In other words, the firm assigns each task entirely to the appropriate specialist. Whenever there are positive organization costs, the task content of a worker of type i is a normal distribution centered on that worker's specialty with variance $\sigma^2 - ln(4)\gamma$. Greater organization costs reduce task specialization. This illustrates two things. First, it is easy to extend the model to accommodate the large-firm case, where the task space is uncountable (as is the worker-type space). Second, the key role of organizational frictions is a deep property of the model that does not go away when the researcher makes organizations large and less "lumpy."

1.11.12 A Quantity-Based Model

In some contexts, such as manufacturing, one may wish to model organizational efficiency as allowing firms to produce greater quantity rather than greater quality. Indeed, this is the default definition of productivity in economics. The model can also be extended to accommodate this: one can simply interpret the skill sets as denoting the amount of time required by the worker to complete task k (therefore smaller $\theta_{i,k}$ are better). Then the production function becomes a function of organization structure:

$$F_{\alpha,B}(a_j) = \min\left\{\frac{a_1}{\alpha_1 \sum_i \theta_{i,1} B_j(i,1)}, ...\frac{a_k}{\alpha_k \sum_i \theta_{i,k} B_j(i,k)}, ..., \frac{a_K}{\alpha_K \sum_i \theta_{i,K} B_j(i,K)}\right\}.$$

Given any fixed organizational structure, the efficient way to produce a single unit of output is to set $a_k = \alpha_k \sum_i \theta_{i,k} B_j(i,k)$. Thus the per-unit wage bill is given by

$$\sum_i W_i \sum_k \alpha_k \sum_i \theta_{i,k} B_j(i,k).$$

Marginal costs are constant and consist of the per-unit wage bill and organization costs:

$$MC_j = \sum_i w_i \sum_k \alpha_k \sum_i \theta_{i,k} B_j(i,k) + \gamma_j I(B_j).$$

All of the benefits of a more complex organization come through a reduction in the per-unit wage bill. In this way, the intuition from the original model extends directly to the quantity case: firms with greater organizational efficiency (lower γ_j) can produce more of the good with the same workforce. I did not use this as the main model because the following property is not compatible with the empirical application to hair salons:

Proposition 7 *Under a quantity model with multinomial logit demand, prices are decreasing with organizational complexity.*

The proof of this proposition is given in the next paragraph. Intuitively, under the quantity model with logit demand, all the benefits of a complex organization come from greater output rather than from greater revenue per unit. The reduction in marginal cost outpaces the increase in the markup, resulting in lower prices. This implies a negative correlation between prices and complexity, which is shown not to be true for hair salons. However, for manufacturing firms, it appears to be true. Caliendo et al. (2020) finds that prices (revenue-based productivity) decline when manufacturing firms reorganize.

Proof. Equation 1.19 from Appendix Section 1.11.9 provides a closed-form expression for price in any Nash Equilibrium under logit demand:

$$\frac{1}{\rho} + c_j + \rho^{-1} W\left(exp\left(-1 + \xi_j - \rho c_j\right) E_j^{-1}\right) = p_j^*.$$

Taking the derivative w.r.t. c_j yields

$$\frac{\partial p_j^*}{\partial c_j} = 1 - exp\left(-1 + \xi_j - \rho c_j\right) E_j^{-1} W'\left(exp\left(-1 + \xi_j - \rho c_j\right) E_j^{-1}\right).$$

A property of the Lambert W function is that

$$W'(x) = \frac{W(x)}{(1 + W(x))x}.$$

Thus, I can simplify the expression to

$$\frac{\partial p_j^*}{\partial c_j} = 1 - \frac{W(exp[-1+\xi_j - \rho c_j]E_j^{-1})}{1+W(exp[-1+\xi_j - \rho c_j]E_j^{-1})}.$$

The Lambert W function is weakly positive for values which are weakly positive; therefore, the derivative is positive, and price is decreasing in cost. The firm minimizes cost:

$$\min_{B \in \mathbb{B}} \gamma I(B_j) + W(B_j).$$

This is again a rate-distortion problem. Denoting the optimal wage-bill as $D = W(B_j^*)$, I can reformulate the problem as before, with the firm choosing D given some optimal organization cost and wage bill:

$$\min_{D} \gamma I(D) + W(D),$$

where I and W are expressed as functions of D instead of B_j. Then, as before, there is a negative cross-partial derivative:

$$\frac{\partial \gamma I(D) + W(D)}{\partial D \partial \gamma} = I'(D) < 0$$

with strict inequality whenever $I(D)$ is strictly positive. This establishes strict decreasing differences of D in γ; thus D is strictly decreasing in γ, and since $I(D)$ is a strictly decreasing function, it is also strictly decreasing in γ. Therefore, prices should be decreasing as γ decreases, while complexity should be increasing.

1.11.12.1 Non-Additive Quality

The model developed in this paper required the effect of the quality of each individual task to have an additive impact on overall quality. This assumption is natural in some settings, but unnatural in others. An excellent example is the launching of space shuttles. A single task performed poorly can be catastrophic, as illustrated by the Challenger explosion. In these contexts, nonlinear quality aggregation is necessary to model the production process. I can accommodate this within the model using multiplicative quality, similar in spirit to Kremer (1993):

$$\xi_j = \prod_{i,k} \theta_{i,k}^{B_{j,i,k}}.$$

Rewriting it using logarithms yields

$$\xi_j = exp\bigg(\sum_{i,k} B_{j,i,k} log(\theta_{i,k}) \bigg).$$

This is now an f-separable distortion measure, meaning I can apply recent work in information theory (Shkel and Verdú 2018) to adapt the Blahut–Arimoto algorithm and other tools to work with this extended model.

1.11.12.2 Quality Positioning and Richer Demand Systems

One surprising result from the theoretical section is that the choice of organization structure depends only on other firms' choices via wages. This derives from the quality-price index assumption placed on demand systems in the model. In some contexts this may be unrealistic, and one may believe that there is a "positioning effect," where the return to higher quality depends in part on how many other firms are also producing high quality. This section illustrates that this effect can be incorporated using mixed logit demand systems if a researcher is willing to sacrifice analytical and computational tractability.

Suppose consumers differ in their taste for quality. The utility of consumer z for product j is now given by

$$u_{z,j} = q_z \xi_j - \rho p_j + \epsilon_{z,j},$$

where q_z is distributed i.i.d. across consumers according to some distribution G. This utility specification now nests both pure vertical and pure horizontal differentiation models. When I specify that $\epsilon_{z,j}$ is a Type 1 extreme value and q_z is normally distributed, the result is a random coefficients logit model. Market share for product i among consumer segment z is given by

$$s_{j,z} = \frac{exp(q_z \xi_j - \rho p_j)}{\sum_{j'} exp(q_z \xi_{j'} - p_{j'})}.$$

To understand the effect of quality on market share, I can compute the derivative:

$$\frac{\partial s_j}{\partial \xi_j} = \int q_z s_{j,z}(1 - s_{j,z}) dG(z).$$

Two facts are apparent from this expression. First, the marginal revenue from increasing quality now depends on the quality position of other firms. Firms will find it more beneficial to raise quality when high-quality segments are relatively untapped by other firms. Second, the optimal organizational structure B_j will depend on the equilibrium quality choices of other firms.

The cost of this more-flexible demand system is tractability. Because of the dependence on the quality choices of other firms, the characterization in Theorem 1 no longer holds. In particular, the firm's problem is not a rate-distortion problem. Estimation requires solving the model for each firm using nonlinear convex optimization. Additionally, demand no longer takes a multinomial logit form, so there does not exist a closed-form solution relating market shares, prices and unobserved qualities. Estimation now also requires numerical integration and a BLP-style contraction mapping to invert market shares.

1.11.12.3 Richer Attention Elasticities

As noted by Csaba (2021), the mutual information cost restriction "attention elasticities" should be constant. To translate this to the context in this paper, consider a firm which is deciding how to split some amount of task k between two workers. Constant attention elasticities means that regardless of the task, and regardless of the initial skills and wages of the two workers, a 1-percent increase in the relative quality-adjusted cost of one of the workers relative to the other increases the probability task k is performed by that worker relative to the other by a constant percentage. Thus the mutual information cost function is building in symmetry and constant percentage changes when counterfactuals change the quality-adjusted costs of different worker types in equilibrium.

Csaba (2021) show that we can use α-mutual information cost functions to allow for more varied elasticities. Exploring such cost functions may be interesting, because they allow for a form of organizational inertia, where how elastic an organization structure is depends on the initial structure.

1.11.13 Knowledge Hierarchies

Rosen (1982) and Garicano (2000) envision firms as characterized by knowledge hierarchies. This idea is incorporated into a quantitative equilibrium model in Caliendo et al. (2012). In this conception, workers differ in their knowledge, tasks differ in their complexity or frequency, demand to each firm is exogenous, and firms choose the number of levels of their organization. There are also communication costs so that sending a problem up to a higher level of the organization has a cost.

As noted by Haanwinckel (2020), there are similarities between task-based models and knowledge hierarchy models. In particular, both models generally result in full specialization, where workers perform non-overlapping sets of tasks. Additionally, tractability in both settings is often maintained by ordering tasks and workers along a common dimension. Finally, estimation and worker types are often inferred from some combination of demographic information and wages. For example, following Caliendo, Monte, and Rossi-Hansberg (2015) it is common practice in knowledge hierarchy papers to group workers into management layers based on occupation and wages, where occupations with similar wages form a layer. In this sense, the difference between my model and knowledge hierarchy models is similar to the differences between my model and task-based models.

In another sense, knowledge hierarchy models seek to explain the hierarchical structure of firms, while my model tries to explain the task assignments within a firm. Because the models have different goals, they are designed differently. One way of bridging the gap is by introducing management worker types and introducing a management task into my framework. A key ingredient is that the management task generally must impact the other tasks multiplicatively. This can be accommodated, but requires new derivations.

1.11.14 Task Classification Process: Further Details

A licensed cosmetologist was paid to categorize 20,560 salon services performed according to their descriptions. As part of the agreement, the person provided a picture of their cosmetology license. The cosmetologist was provided with a blank spreadsheet with predefined subcategories and was instructed to mark all subcategories where the description matched with a 1. They were instructed that some subcategories may not be mutually exclusive, so they should mark all that applied. The initial job description was as follows:

> I have a list of approx. 20,560 short descriptions of salon services (mainly hair salons, but also some nail/spas). I would like someone with knowledge of the industry to mark whether each descriptions fits into one of several categories (male/female service, coloring, cutting, highlighting, washing, etc). This amounts to putting a 1 in each column that fits the description.

In a follow-up message I further clarified the instructions:

> Here are the descriptions. I did the first few to give you a sense of the task. Basically read the description and then put a 1 in all categories that fit. Sometimes a description may match many, sometimes 1, rarely none. If you start reading them and see that it may be worth adding a separate category let me know. The idea though is to capture the core "tasks" or services performed at hair salons, like cut, color, highlight, style, etc and also to get some info on gender and typos.

After the first draft was submitted, I checked the coding, looking for any mistakes or missed descriptions, and sent the document back to the cosmetologist several times for revision. A sample from the final spreadsheet is displayed in Figure 1.22.

Since the subcategories were very detailed, I hired the same cosmetologist, at a rate of $100, to classify the subcategories into six task categories. The specific instructions given to the cosmetologist were as follows:

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Service Description	Typo	Cut	Shave	Style	Extensions	Color/Bleach	Highlight/Balayage	Other Treatment	Blowdry	Wash/Shampoo	Eyebrow/Eye Lash Service	Admin/Consult	Spa Service	Nail service	Male	Female	Child	No Info
1319	'Add-on Camo								1				1						
1320	'Add-on Clipper cut (for booking with other Fa		1										1			1			
1321	'Add-on Color (1OZ)						1						1						
1322	'Add-on Color Melt						1						1						
1323	'Add-on Curling Iron				1								1						
1324	'Add-on Extra 20oz Lightener						1						1						
1325	'Add-on Haircut		1										1				1		
1326	'Add-on Haircut - (Thick hair)		1										1				1		
1327	'Add-on Haircut - with Color		1										1				1		
1328	'Add-on Haircut -(Fine Hair)		1										1				1		
1329	'Add-on Iron Service				1								1				1		
1330	'Add-on Kids Cut (when added to service of ar		1										1					1	
1331	'Add-on Lightener						1						1						
1332	'Add-on Olaplex						1						1						
1333	'Add-on Pixie cut		1										1				1		
1334	'Add-on Salon Metal Detox												1						
1335	'Add-on Scalp Treatment								1				1						
1336	'Add-on Service Only												1						
1337	'Add-on Toner						1						1						
1338	'Add-on Treatment								1				1						
1339	'Added Base						1						1						
1340	'Added Biolage Color						1						1						
1341	'Added Color For Eyebrows											1	1						
1342	'Added Deep Conditioner								1				1						
1343	'Added Elumen Glaze								1				1						
1344	'Added Glaze								1				1						
1345	'Added Highlight							1					1						
1346	'Added K18								1				1						
1347	'Added Nectaya Single												1						

Figure 1.22: Final Task Subcategorization Spreadsheet from Cosmetologist

> Please categorize the 13 tasks from before into "groups." For the 6 group column, put the 13 tasks into 6 groups that are most similar in terms of who would do them/tasks they would require. for example, if color and highlight are similar, mark both as number 1. Number the groups 1 through 6. For the four group column, make 4 groups, etc. Underneath, please write a small note describing why you put the tasks together the way you did.

I use the six-category grouping provided by the cosmetologist with one modification: I combine the extension task with the blow-dry task to create five final task categories, because the extension task is very sparse–for Manhattan in 2021 Q2, fewer than 10 hours were dedicated to this task. This sparsity leads to estimation problems, as parameters tied to this task have a negligible effect on observable outcomes.

If a service is marked as multiple task categories, I divide the service into unique tasks in the following way. First, I compute the average amount of time spent on each task among services that are marked only as one task. Second, I compute the fraction of time to assign to each task as the corresponding task average divided by the sum of the averages of all other tasks marked for that service. Third, I distribute the total time spent on the service across the tasks using this imputed fraction. This process generates task categories that are mutually exclusive.

1.11.15 Robustness of Stylized Facts

The first concern is one of reverse causality. Perhaps firm size allows firms to be organizationally complex and thus have a product market advantage.[38] Appendix Section 1.11.18 shows that while this cannot be ruled out, it is not generating all of the observed relationships. Even among firm-quarters with the same number of employees, there is significant variation in complexity, and there is a positive association between complexity and the main market outcomes (i.e., revenue, prices and repeat customers).

A second concern is that the correlations are driven by demand-side factors, such as consumer preferences for particular stylists rather than firm choices. The software records when customers request a particular staff member. It would be concerning if there was a strong positive correlation between the request rate at a salon and complexity. Appendix Section 1.11.19 shows that while many customers request specific staff, the rate of requests across salons is not correlated with organization complexity. Further, the correlation between the request rate and firm size is either zero or negative.

A third concern is that the correlations are driven by the specific functional form chosen for complexity. Appendix Section 1.11.20 shows that the main patterns persist when complexity is replaced by within-visit specialization. Within-visit specialization is measured as the fraction of multi-service visits which are performed by a team (i.e., more than one employee).

1.11.16 Measurement Error in Organization Complexity

Complexity is estimated based on the observed task assignments within firm, yet the empirical part of this paper treats complexity as if it were observed or measured without error. One justification is that many assignments are observed per firm per quarter, so estimation error should be small. If estimation error at the quarter level is small, the correlation between complexity measures at the month level within quarter should be

38. Li and Tian (2013) provide a theoretical mechanism for such an effect.

large. This section illustrates that this is indeed the case.

To do this, I recompute complexity for each month within a quarter so that I have three measurements of complexity per firm-quarter observation. In the full sample, the pairwise correlation between the first and second month is 0.945, the first and third is 0.98, and the second and third is 0.939. When 2020 (the onset of the coronavirus pandemic) is excluded, the pairwise correlations are 0.978, 0.962 and 0.976, respectively. The high correlation between complexity measurements within quarters suggests that complexity at the quarter level is measured precisely.

The assumption that complexity is measured without error allows me to "invert" complexity to obtain the underlying organization cost for each firm in a market. In a similar way, researchers in industrial organization often assume market shares are measured without error in order to invert them to obtain mean utilities for each firm. It is possible to relax this assumption and use the panel nature of the data to estimate each firm's organization cost parameter.

1.11.17 Firm Size and Complexity Associations

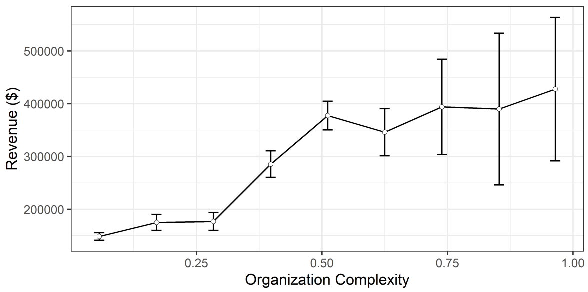
Table 1.13: Regressions of Firm Size on Complexity, Manhattan Only

Dependent Variables:	Revenue	Employees	Utilized Labor	Customers	Visits
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Org. Complexity	430406.6*	12.55	-17733.9	277.2	876.9
	(179977.4)	(6.531)	(70765.2)	(600)	(907.1)
Fixed-effects					
Quarter-Year	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	595	595	595	595	595
R^2	0.33485	0.21039	0.20359	0.44164	0.48831

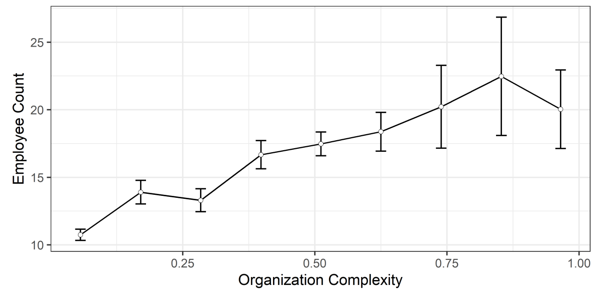
Clustered standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05*

Note: This table repeats the regressions of revenue and other measures of firm size on complexity, but only for New York County (Manhattan). The positive relationship between revenue and complexity remains statistically significant.

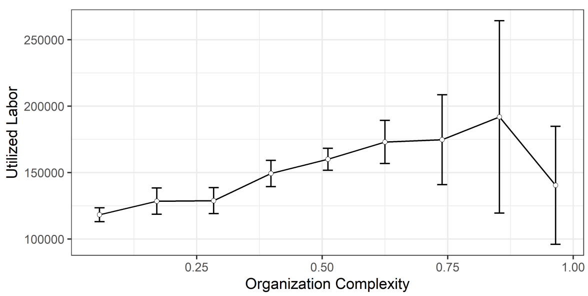
Figure 1.23: Organization Complexity and Firm Size



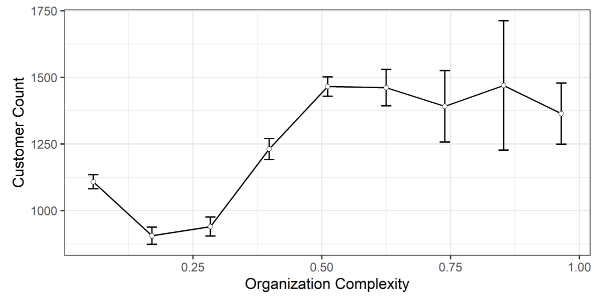
(a) Revenue ($)



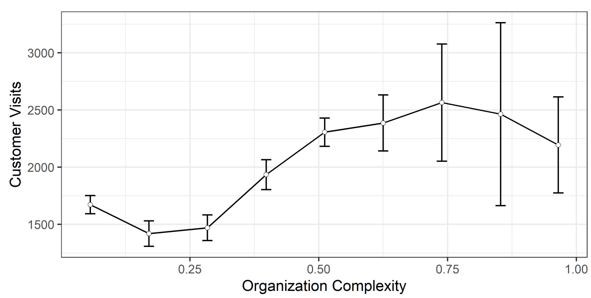
(b) Employees



(c) Utilized Labor (minutes)



(d) Unique Customers



(e) Customer Visits

1.11.18 Complexity Relationships Among Similar-Size Firms

The main text of the paper established that complexity is correlated with the number of employees as well as other outcomes. This raises concerns about the direction of causality: are firms larger because they are more internally complex, or are larger firms naturally able to design more internally complex structures? The model in this paper specifies a common organization cost, which generates jointly both larger and more complex firms. In this sense, complexity does not cause a firm to be larger; rather a common, unobserved productivity heterogeneity generates both.

Table 1.14: Regressions of Revenue on Complexity and Employee Count Interacted

Dependent Variable:	Revenue		
Model:	(1)	(2)	(3)
Variables			
(Intercept)	79487.9***		
	(19103.3)		
Complexity	-226181.8*	-242961.5*	-320973.6**
	(111684)	(110939.4)	(117545.2)
Employee Count	5652.8*	4871.6*	3878.9
	(2315.3)	(2257)	(2192.2)
Complexity × Employee Count	29487.9***	30187.8***	35052.8***
	(8587.8)	(8507.4)	(8528.5)
Fixed-effects			
Quarter-Year		Yes	Yes
County			Yes
Fit statistics			
Observations	4,558	4,558	4,558
R^2	0.4913	0.52042	0.61654

Clustered standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05*

Note: This table presents regressions of revenue on complexity interacted with employee count. The mean number of employees is 13.38, so the marginal effects in all specifications evaluated at the mean are positive.

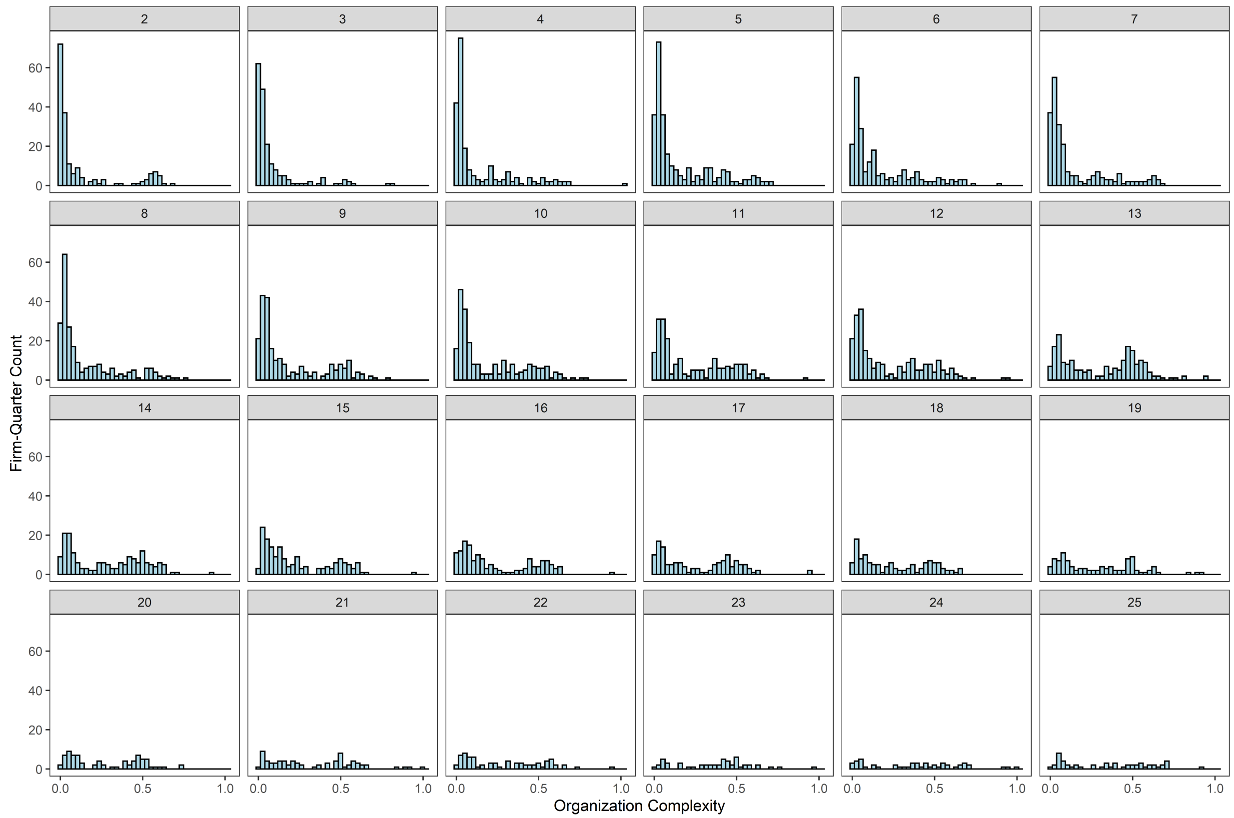
This answer still leaves several questions open. In particular, perhaps organization costs are more like fixed costs, so larger firms are better able to afford more complex organizations. Additionally, maybe larger firms have more organizational possibilities, and thus the relationships discussed are mechanical. I alleviate this concern by analyzing many of the outcomes among firms with the same number of employees.

The positive correlation between complexity and revenue, prices and repeat customers persists among firm-quarters with the same number of employees. There is a positive correlation within almost all firm sizes and for almost all variables. The exception is repeat customers among firms with 2–5 employees. In general, the positive correlation is larger in magnitude for firm-quarters with 13 or more employees. This can be seen in Figure 1.25, which shows scatter plots with linear best fit lines for firm-quarters with the same number of employees.

Essentially, while complexity is correlated with both employee count and other market outcomes, and employee count is correlated with the other market outcomes, there seems to be a large, direct effect of complexity on market outcomes. Another way to see this is that when firm-size fixed effects are added to a regression of revenue on complexity, the point estimate for complexity decreases by around 60 percent, but remains economically and statistically significant. So while much of the effect of complexity on other outcomes seems to come through size, a sizable amount does not.

I do not interpret these correlations as causal. Rather, I take them as evidence that there is an organizational advantage that operates through channels beyond just firm size. For this reason, the model is built such that a common firm characteristic (γ_j) generates both larger firms and more complex organizations.

Figure 1.24: Organization Complexity for Similarly Sized Firms



Note: Each plot is a histogram of complexity among firm-quarters with the shown number of employees. I perform the analysis for salons with fewer than 25 employees, as these represent the bulk of the data. Salons with 1 employee are excluded because mechanically complexity is 0 for these salons.

1.11.19 Consumers Requesting Particular Staff

The stylized facts and the model treat the assignment of workers to tasks as a choice of the firm. In practice, some customers directly request particular stylists. The software allows salons to record when a staff member is requested for a task, and this information is captured in a variable titled ”Was Staff Requested.” This section establishes that although there is heterogeneity in how often staff are requested at different salons, this heterogeneity is not correlated with organization complexity.

I start by examining the variation in requests across salon-quarters in Figure 1.26. A large number of salon-quarters have no requests observed in a quarter (Panel A). Among those salon-quarters with at least one request, the request rate varies significantly, spanning close to 0 all the way to 1 with a mode around 0.8 (Panel B). Much of this heterogeneity comes form an aggregate increase in the request rate over time (Panel C). Therefore, I also run analyses excluding quarters before the first observed request for a salon. I call this sample “after adoption.”

The primary question is whether consumer requests are driving observed organization complexity. I test this using binned scatter plots in Figure 1.27. Both unconditionally (Panel A) and among salon-quarters with one request (Panel B), complexity does not appear to have a systematic relationship with the request rate.

Regressions with standard errors clustered at the salon level also reveal mixed results. In the full sample, the coefficient on the request rate is statistically insignificant and negative. In the after-adoption sample, the coefficient is statistically insignificant and positive. In both cases, the coefficients are economically insignificant: they imply that a 1-standard-deviation increase in the request rate is associated with less than a 0.08-standard-deviation change in complexity.

Further, Figure 1.28 shows there is no evidence of a positive relationship between firm size and the request rate (if anything, there may be a negative relationship), which suggests the positive relationship between complexity and firm size documented in the stylized facts is not driven by customer request.

In summary, the data suggest that customers often request specific stylists (many salons have request rates around 0.8) but that this varies significantly across salon-quarters. This is in line with the intuition that requests are common. However, the correlation between the request rate and complexity at salons is statistically and economically weak, evidence that while consumers do request staff, these requests are not first-order determinants of the complexity differences across salons.

At a theoretical level, there is a question of whether a strong positive correlation would matter. Whether a consumer requests a stylist or a firm assigns a stylist, the firm still must bare the associated organization cost. Further, if consumers prefer a particular stylist for a task, this likely reflects the stylist's quality at that task and a match effect. Since quality differences across tasks are already built into the model, if the match effect is small, this phenomenon is captured by the existing model.

1.11.20 Within-Visit Specialization

This section shows that many of the correlations between complexity and market outcomes persist when complexity is replaced with a simpler measure of within-visit specialization. I compute within-visit specialization as the number of customer visits[39] with two or more employees assigned divided by the number of customer visits with two or more services performed.

A histogram of this measure shows that it follows a similar power-law distribution as organization complexity, with observed values spanning the support and a long right tail. Like organization complexity, within-visit specialization is positively correlated with revenue, price and the share of repeat visits. However, unlike organization complexity, it has a non-monotone relationship with the number of employees.

These findings are further support that more internally specialized firms command a competitive advantage. To finish this section, I study the connection between com-

39. Visits are the number of unique customer-date pairs in a quarter.

plexity and within-visit specialization. A simple regression of complexity on within-visit specialization yields an R-squared of 0.50, suggesting that nearly half of the variation in complexity can be accounted for by specialization within-visit. Kendall's rank correlation coefficient between the two variables is 0.488. This can be interpreted as 25.6 percent of pairs of firm-quarters being discordant. Roughly, if two firm-quarters are drawn randomly, their ordering according to complexity and within-visit specialization will be the same 74.4 percent of the time. The strong connection between the two variables is visualized in a binned scatter plot in Figure 1.30.

1.11.21 Task Content Variance Decomposition

Using the estimated model, I can study the determinants of the task content of hair-salon jobs in Manhattan. As a first step, I decompose the variation in task content into a worker and firm component. Using the distribution of model generated jobs, I can write

$$b_j(i,k) = \bar{b}(i,k) + (b_j(i,k) - \bar{b}(i,k)).$$

I then adapt the method used by Song et al. (2019) to my setting. Fixing k and taking the distribution to be weighted by effective units of labor, I then decompose the variance into a worker-type component and a within-worker-type component, where recall ω_i is the share of total labor represented by workers of skill set i:

$$Var_{i,j}(b_j(i,k)) = Var_i(\bar{b}(i,k)) + \sum_i \omega_i Var_j(b_j(i,k)|i).$$

However, because $b_j(i,k)$ is generated by the structural model, and I am considering a single labor and product market, the within-type component comes entirely from variation in firm attributes. Therefore, I have decomposed the total variance in task content into a worker and firm component. Dividing through by $Var_{i,j}(b_j(i,k))$ gives the share of variance due to each component, as shown in Table 1.21. For the main tasks (cut, color, blow-dry), between 8 and 22 percent of variation in job task content is attributable to firms.

1.11.22 Bootstrap Procedure

During each bootstrap replication, the model is fully re-estimated. The estimation procedure has three loops, which are run with slightly looser tolerances than the primary estimation algorithm. The innermost loop, which is the Blahut–Arimoto algorithm, is run with a convergence tolerance of 10^{-10}. The middle loop, for which I choose $gamma_j$ to match each firm's complexity, is run with a tolerance of 10^{-8}. The outer loop, which finds the market parameters, uses the Nelder-Mead method. Relative and absolute tolerance are set to 10^{-8} with a maximum number of 4000 iterations.

Standard errors are computed as the sample standard deviation of the bootstrap distribution of each parameter. To check the stability of standard errors, I ran an additional 23 replications. The standard errors with these additional replications are within 4% of the reported standard errors.

1.11.23 The Full Distribution of Task Content

I can also go beyond the variance and compare the entire distribution of model-generated and observed job task content. This is a strong test of the model, because I observe 509 stylists in Manhattan during 2021Q2 (the estimation period), and I am asking the model to match their jobs using only firm-based moments. Figure 1.31 plots the two distributions for each of the six tasks. Although the match is not perfect, the model is able to replicate important features of the data. As an example, panel B shows that the fraction of time stylists spend on the coloring task is tri-modal in the data, with peaks at 0%, 40% and 90%. The model is able to approximate this pattern with a bimodal distribution, with a wide first peak that merges the 0% and 40% observed in the data.

1.11.24 Counterfactual Procedures

1.11.24.1 General Procedure

The general procedure used in all counterfactuals is as follows.

1. Weight each firm such that the observed total market share matches the share of people purchasing some amount of hair salon services in the Consumer Expenditure Survey. This means that each Manhattan salon in the data is assumed to represent 23 salons.[40]

2. Compute the implied total labor supply of each worker skill set by summing all labor demands at initial wages over all firms.

3. Make the relevant parameter changes that correspond to the counterfactual.

4. Guess wages.

5. Solve for organization structures: If allowing internal reorganization, use the Blahut–Arimoto algorithm described in the estimation procedure to solve for each firm's organization structure.

6. Compute optimal pricing: Given organization structures, qualities and costs of all firms are now known. Optimal prices are computed by iterating on each firm's best response pricing function until convergence.

7. Check labor market clearing by comparing the new labor demands to the total labor supply computed in step 3. If supply and demand for each type match, exit. Otherwise, return to step 4.

I assume that the exogenous quality (ν_j) and exogenous marginal cost (ϕ_j) remain the same in the counterfactual analyses. To solve for market clearing wages, I minimize the

40. It is necessary to weight firms for counterfactual analysis but not estimation. This is because during estimation I fix an equilibrium, but in counterfactuals I must find a new equilibrium, and firm pricing strategies depend on the number of other firms in the market.

sum of squared excess labor demand. I use the L-BFGS-B routine and stop only when the objective is less than 0.1. This corresponds to a very close match between labor supply and demand. I found it more efficient to use a minimization routine because the labor demands of each worker type depend in a complex manner on the entire vector of wages. That is, while each labor demand is monotone decreasing in own wage, firms can use other worker types as substitutes, so I cannot simply find each of the six market clearing wages sequentially.

Total welfare is defined as the sum of total wages, consumer welfare and total profit. Task specialization is defined as the total amount of labor spent on a worker's specialty tasks. Reported average prices, qualities and complexities are at the firm level, and are not weighted by market share. Wage statistics are weighted by the labor supply of each worker type.

1.11.24.2 Minimum Wage Technical Details

In general, a minimum wage can result in multiple equilibria. To ensure that there are not multiple equilibria, I solved the model under every possible permutation of binding minimum wages. That is, I assumed the minimum wage binds for worker types 1 and 2 only, 1, 2 and 3 only, etc. With six worker types, this amounts to solving the model $2^6 = 64$ times. Each time I solved the model, I fixed the wages of the binding types at $30, and then solved for the wages of the other types which clear the labor market for only those other types. Afterwards, I checked that:

1. Worker types with non-binding wages have wages greater than $30.

2. Worker types with binding minimum wages have excess labor supply.

Any solution which passed this check was considered a valid equilibrium. For example, for the case when the minimum wage is binding only for type 1, I set type 1's wage to $30 up-front, then solved for the other five wages which clear the market for the other five types. I then checked that types 1 through 6 have wages above $30 and type 1's excess labor supply is positive.

This process indicates there is a unique equilibrium in both counterfactuals. For both the full adjustment and no adjustment counterfactuals, only one of the 64 cases satisfied the checks as a valid equilibrium. One additional case in the full adjustment counterfactual never converged, meaning I could not find wages that cleared the labor market for the non-binding worker types. The wages, employment and task specialization in the initial, reallocation and full equilibrium are provided in Table 1.19.

1.11.25 Job-Level Heterogeneity

Table 1.15: Job Task Mix

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Share Haircut/Shave	62,671	0.387	0.344	0.000	0.008	0.669	1.000
Share Color/Highlight/Wash/Extensions	62,586	0.371	0.322	0.000	0.025	0.599	1.000
Share Blowdry/Style/Treatment	62,564	0.102	0.162	0.000	0.008	0.124	1.000
Share Administrative	62,702	0.061	0.168	0.000	0.000	0.027	1.000
Share Nail/Spa/Eye/Misc.	63,012	0.076	0.227	0.000	0.000	0.010	1.000

Note: This table displays summary statistics about the time spent on each task at the worker level. While worker averages correspond roughly to firm averages, there is greater heterogeneity across workers, supporting the idea that within firms there are distinct roles.

1.11.26 Flexible Labor-Labor Substitution

A key difference between the model in this paper and others in the task literature is that workers differ horizontally and firms differ in their organization costs and task-based production functions. These features allow for richer forms of labor-labor substitution, and as a result richer responses to policy change. Just as allowing for richer consumer substitution patterns is important for understanding the impact of policies on consumers, allowing for richer labor-labor substitution is important for understanding the impact of policies on workers.

In most models of task assignment, tasks can be ordered in a single dimension. For expositional purposes, let us call this dimension difficulty. Workers can also be ordered by their skill at completing difficult tasks. As articulated by Teulings (2000), this leads to distance-dependent complementarity, a strong assumption on substitution patterns. Across all firms, workers closer in education are substitutes, while those farther away in education are complements.

Distance-dependent complementarity restricts the effects of policy changes. As an example, consider a minimum wage increase in a model with competitive labor markets and only distance-dependent complementarity.[41] Beyond the direct effect on low-skill workers, the policy has two effects. First, it causes labor-labor substitution towards the closest available substitutes (medium-skill workers). Second, it increases costs, reducing prices and overall labor demand. Wage spillovers are decreasing in initial wage, potentially becoming negative near the top of the distribution. This is theoretically unambiguous and occurs regardless of model parameters.

In my model, distance-dependent complementarity does not hold. Instead, whether a worker is a substitute for or a complement to the binding worker type varies from firm to firm and depends on the firm-specific organization cost and firm-specific task-based production function. Aggregate employment/wage effects then depend on how market share is distributed across firms. If there are two firms, and worker type A is a substitute for minimum wage workers at firm 1 and a complement at firm 2, whether the minimum wage increases or decreases the workers' wage depends on the market share of firm 1 vs. firm 2. This is why, when I apply the model to Manhattan hair salons, I find a minimum wage increase generates non-monotonic wage spillovers. I visualize these in Figure 1.19. The fact that aggregate substitution is data-driven and ambiguous makes the model well suited to explore other questions. One notable example is the introduction of workers with new skill sets. In my model, it is ambiguous whether a new worker type will

Table 1.16: Regressions of Revenue on Complexity

Dependent Variable:	Revenue					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
(Intercept)	121578.1***					
	(14835.8)					
Organization Complexity	456571.3***	440904.1***	485026.4***	486995.5***	271694.6**	261697**
	(100394.8)	(108427.1)	(116918.9)	(125004.8)	(87031.1)	(80920.6)
Task Mix 2				-19070.7	-7609.7	14482.9
				(93817.4)	(78597)	(67354.5)
Task Mix 3				-8011.8	116011.4	98022
				(81014.1)	(106735)	(98077.1)
Task Mix 4				-24893.1	76296.2	67131.1
				(113959)	(96547)	(95768.9)
Task Mix 5				43954.8	14593.5	33562.4
				(50238.8)	(47813)	(56691.1)
Staff Request Rate						-94370.7
						(89112.9)
Fixed-effects						
Quarter-Year		Yes	Yes	Yes	Yes	Yes
County			Yes	Yes	Yes	Yes
Firm Size					Yes	Yes
Fit statistics						
Observations	5,116	5,116	5,116	5,116	5,116	5,116
R^2	0.01475	0.01915	0.3104	0.31047	0.34273	0.34365

Clustered standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05*

Note: This table reports the regressions of revenue on complexity under various specifications, including controlling for the rate of staff requested.

1.11.27 Supplementary Tables and Figures

Table 1.21: Model-Based Decomposition of Job Task-Content Variance

	Share of Task-Content Variance	
Task	Firm	Worker
Haircut/Shave	0.0761	0.9239
Color/Highlight/Wash	0.1194	0.8806
Blowdry/Style/Treatment/Extension	0.2180	0.7820
Administrative	0.0965	0.9035
Nail/Spa/Eye/Misc.	0.0865	0.9135

Note: The table displays a variance decomposition which uses the model to separate the variance of job task content into a worker and firm component.

41. For an empirical example of such a model, see Gregory and Zierahn (2022).

Table 1.17: Two Estimated Organization Structures

Specialist \ Task	Cut	Color	Blow Dry	Admin.	Nail/Misc.	Total
Cut	0.15	0.01	0.001	0.06	0	0.22
Color	0	0	0	0	0	0
Blow Dry	0	0	0	0	0	0
Admin.	0.31	0.03	0.003	0.45	0	0.784
Nail/Misc.	0	0	0	0	0	0
Tot.	0.455	0.036	0.004	0.505	0	1

(a) Salon 1, $I_j = 0.03$

Specialist \ Task	Cut	Color	Blow Dry	Admin.	Nail/Misc.	Total
Cut	0.180	0.003	0	0.006	0.003	0.193
Color	0.057	0.553	0	0.016	0.009	0.116
Blow Dry	0.012	0.002	0.097	0.003	0.002	0.636
Admin.	0	0	0	0	0	0
Nail/Misc.	0.004	0.001	0	0.001	0.050	0.055
Tot.	0.253	0.559	0.097	0.026	0.064	1

(b) Salon 2, $I_j = 0.70$

Note: These are estimated organization structures (B_j) for a high- and a low-complexity salon in New York in Quarter 2, 2021.

Table 1.18: Variance Decomposition: Without a Model

Across Firms

		Share of Variance	
Task	Share of Labor	Firm	Within-Firm
Haircut/Shave	0.4049	0.3744	0.6256
Color/Highlight/Wash	0.3902	0.2899	0.7101
Blowdry/Style/Treatment/Extension	0.0850	0.5056	0.4944
Administrative	0.0590	0.4900	0.5100
Nail/Spa/Eye/Misc.	0.0610	0.4124	0.5876

Across Quarters

		Share of Variance	
Task	Share of Labor	Quarter	Within-Quarter
Haircut/Shave	0.4049	0.0057	0.9943
Color/Highlight/Wash	0.3902	0.0062	0.9938
Blowdry/Style/Treatment/Extension	0.0850	0.0111	0.9889
Administrative	0.0590	0.0193	0.9807
Nail/Spa/Eye/Misc.	0.0610	0.0118	0.9882

Table 1.19: Minimum Wage Counterfactual Type-Specific Wages, Employment and Specialization

	Initial			Reallocation			Reorganization		
Worker Type	Hours	Wage	Task-Spec.	Hours	Wage	Task-Spec.	Hours	Wage	Task-Spec.
Haircut/Shave	537550	\$16.96	0.9463	506090	\$20.00	0.9459	502152	\$20.00	0.947
Color/Highlight/Wash	997053	\$37.75	0.7245	997053	\$37.33	0.7233	997053	\$37.52	0.7209
Blowdry/Style/Treatment/Extension	444040	\$20.91	0.4837	444040	\$21.88	0.4817	444040	\$21.64	0.4819
Administrative	41860	\$26.99	0.6801	41860	\$28.40	0.6807	41860	\$28.12	0.6809
Nail/Spa/Eye/Misc.	34844	\$81.16	0.8262	34844	\$81.63	0.826	34844	\$81.71	0.826

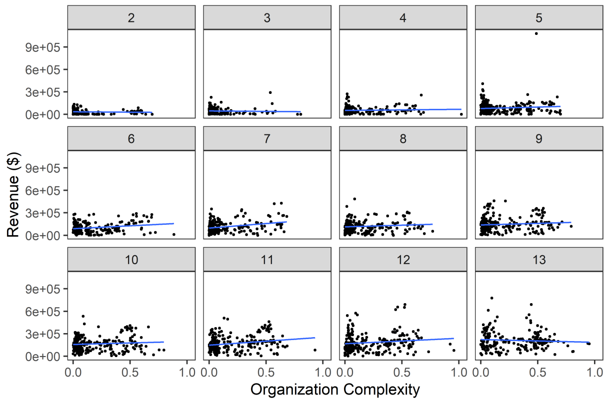
Note: This table displays employment and wage levels across the initial, reallocation and full equilibrium under a \$20 minimum wage. It provides context for the main counterfactual results, which are reported in percentages.

Table 1.20: Sales Tax Counterfactual Type-Specific Wages, Employment and Specialization

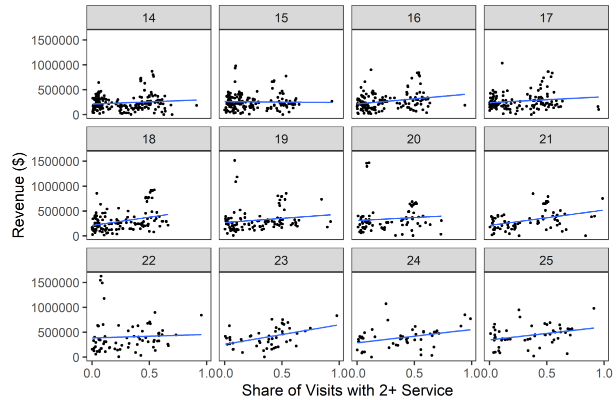
	Initial			Reallocation			Reorganization		
Worker Type	Hours	Wage	Task-Spec.	Hours	Wage	Task-Spec.	Hours	Wage	Task-Spec.
Haircut/Shave	537550	\$16.96	0.9463	537550	\$21.18	0.9471	537550	\$22.38	0.9491
Color/Highlight/Wash	997053	\$37.75	0.7245	997053	\$45.99	0.7326	997053	\$45.34	0.7432
Blowdry/Style/Treatment/Extension	444040	\$20.91	0.4837	444040	\$21.01	0.4946	444040	\$22.18	0.4982
Administrative	41860	\$26.99	0.6801	41860	\$30.15	0.6786	41860	\$31.85	0.6872
Nail/Spa/Eye/Misc.	34844	\$81.16	0.8262	34844	\$90.75	0.8351	34844	\$91.49	0.846

Note: This table displays employment and wage levels across the initial, reallocation and full equilibrium under the elimination of the service sales tax. It provides context for the main counterfactual results, which are reported in percentages.

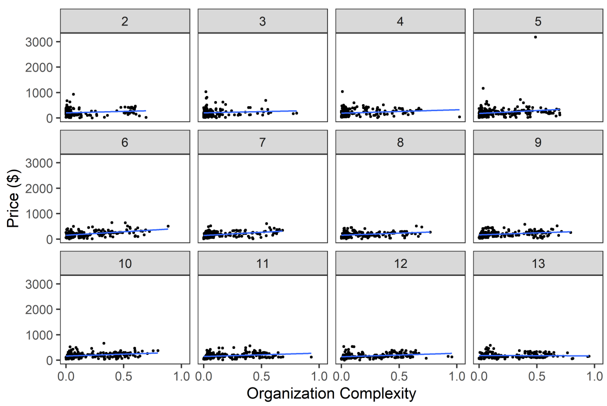
Figure 1.25: Complexity Relationships Among Similar-Size Firms: 2–13 Employees



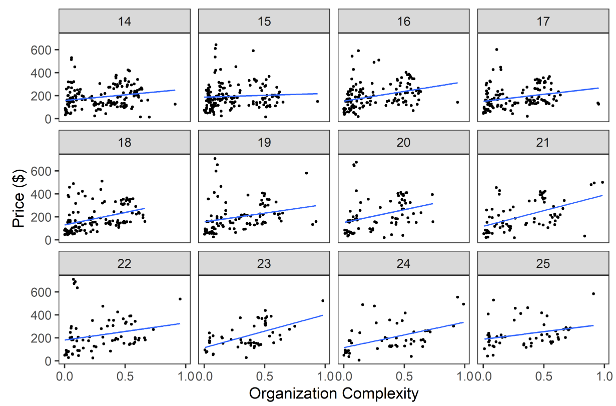
(a) Revenue (2–13 Employees)



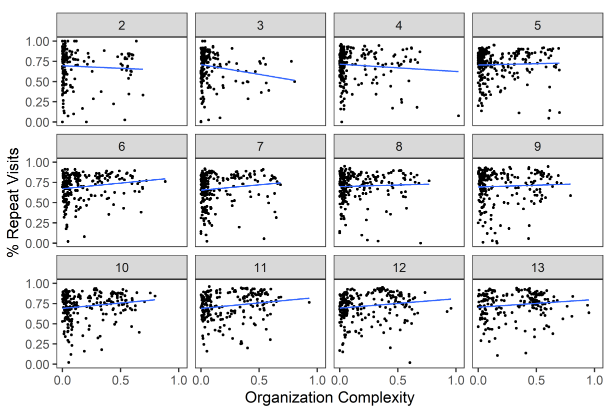
(b) Revenue (14–25 Employees)



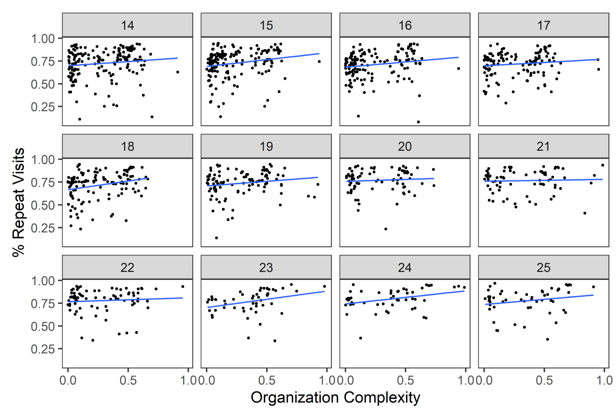
(c) Price (2–13 Employees)



(d) Price (14–25 Employees)

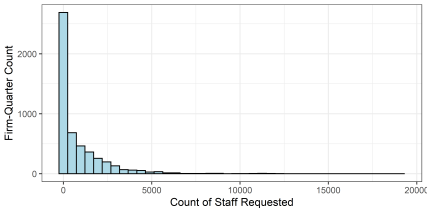


(e) Repeat Customers (2–13 Employees)

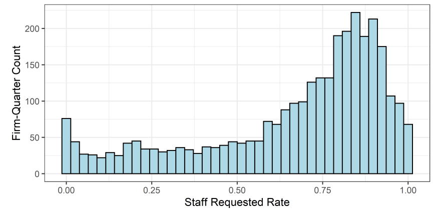


(f) Repeat Customers (14–25 Employees)

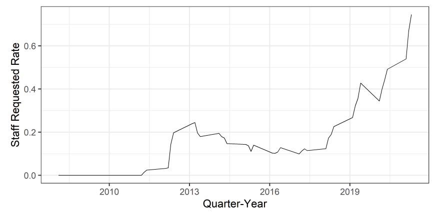
Figure 1.26: Was Staff Requested?



(a) Request Count Histogram

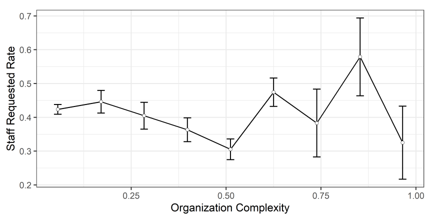


(b) Rate (After Adoption)

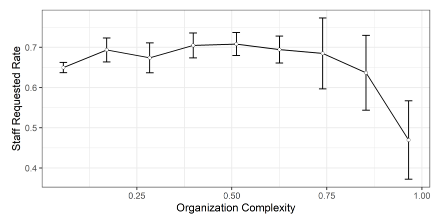


(c) Aggregate Rate Across Time

Figure 1.27: Request Rate and Organization Complexity

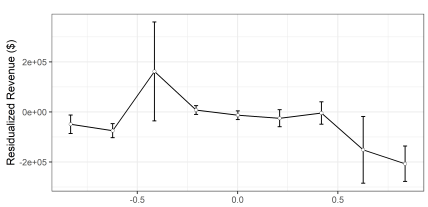


(a) All Salon-Quarters

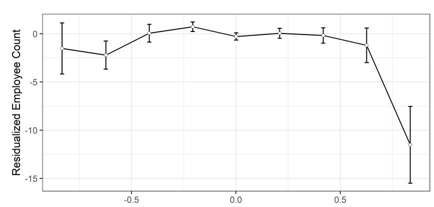


(b) After Adoption

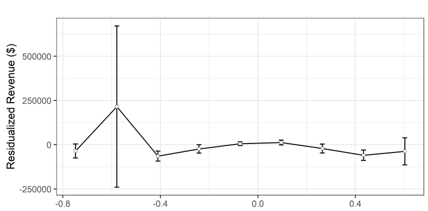
Figure 1.28: Request Rate and Firm Size



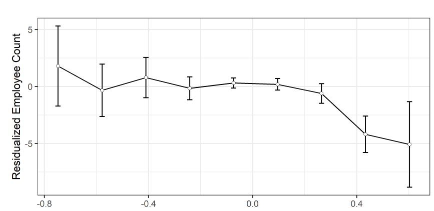
(a) All, Revenue



(b) All, Employees

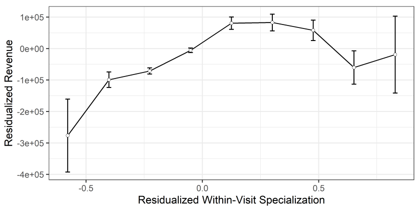


(c) After Adoption, Revenue

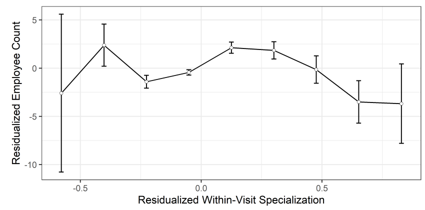


(d) After Adoption, Employees

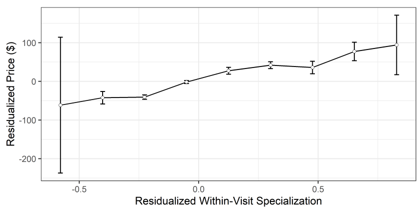
Figure 1.29: Within-Visit Specialization



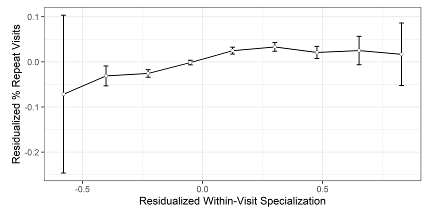
(a) Revenue



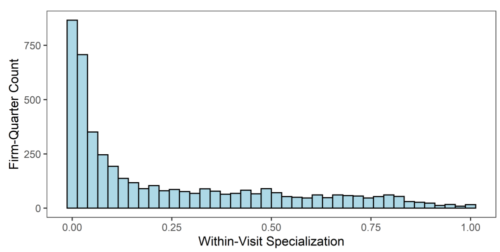
(b) Employees



(c) Price



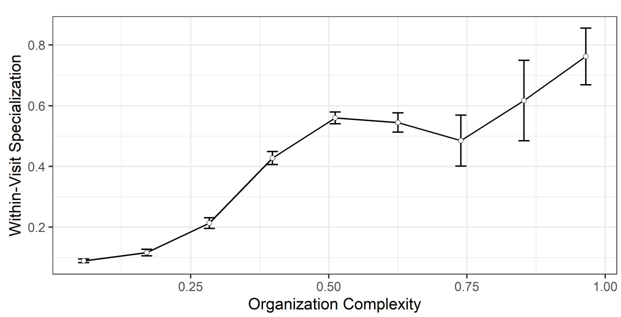
(d) Repeat Customers



(e) Histogram

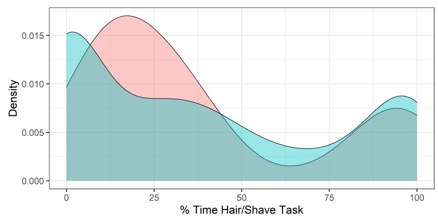
Note: Within-visit specialization is the share of visits with multiple services that are assigned to multiple employees.

Figure 1.30: Organization Complexity and Within-Visit Specialization

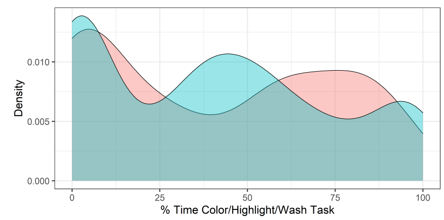


Note: There is a strong, positive correlation between complexity and within-visit specialization.

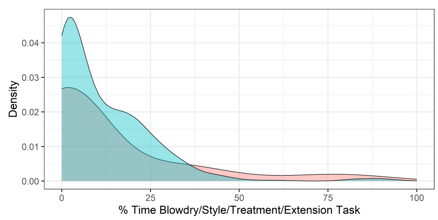
Figure 1.31: Model (Red) vs. Observed (Blue) Job Task Content in Manhattan



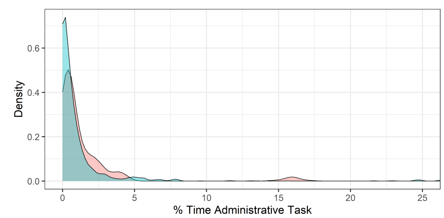
(a) Hair/Shave Task



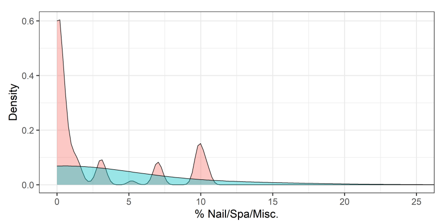
(b) Color/Highlight/Wash Task



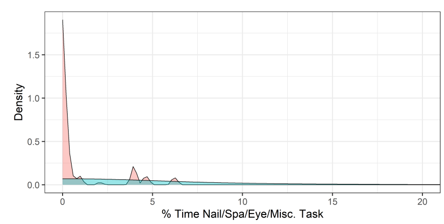
(c) Extension Task



(d) Blowdry/Style/Treatment Task

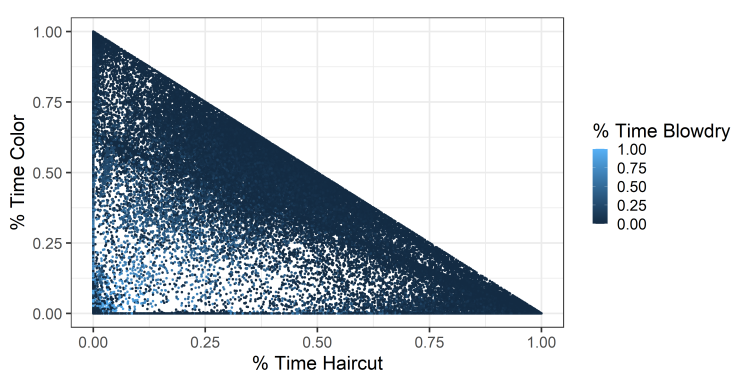


(e) Administrative Task



(f) Nail/Spa/Eye/Misc. Task

Figure 1.32: The Job Task-Mix Distribution



CHAPTER 2

Delegated Recruitment and Hiring Distortions

2.1 Introduction

Talent allocation has always been an economic force at the center of research about productivity, inequality, and discrimination. The internet and popular press are brimming with inspiring quotes, white papers and advice all conveying a similar message: people are everything. And yet, the actual search for talented people is frequently a delegated responsibility. Between 2010 and 2018, the fraction of job postings for recruiting roles more than doubled. As of 2020, 18 percent of employed American workers had found their current job through a recruiter or a headhunter (Black, Hasan, and Koning 2020). Delegated recruitment is now a major feature of the labor market landscape, but the impact of this new feature remains unclear. Are recruiters well-aligned agents of the firm, or does delegation introduce distortions into the hiring process?

To answer this question, we develop a model where a principal delegates sequential search over uncertain objects to an agent. In our model, the principal or the firm (she) employs the agent or the recruiter (he) to search for a worker. The recruiter does not know the exact value of a searched worker's productivity; rather, he holds a belief about worker productivity, which we will assume throughout the paper is characterized by an expectation and a variance. Contracts take a binary refund form, in which the recruiter is paid an amount upfront but must refund a portion of the payment if the hired worker is fired. Throughout the paper, we compare this model to a first-best benchmark, where the firm searches for a worker directly.

Our main theoretical contribution is a tractable analysis of delegated sequential search over uncertain objects under hidden actions when the principal is restricted to binary contracts. Our approach allows us to fully characterize the firm's optimal contract with a single first-order condition. The characterization shows that delegation via refund contracts is equivalent to making the search technology less accurate. We provide a necessary and sufficient condition under which there is less search effort exerted under delegation than there is in the first-best. Additionally, the expected productivity distribution among selected workers under delegation is lower than under first-best. Finally, we demonstrate that social surplus and search effort increase and converge to the first-best benchmark as heterogeneity in productivity variance decreases. We show that our results apply naturally to two common parametric distributions.

To fix ideas, consider two candidates for a data science position. Candidate A is traditional: they graduated from a four-year college with a degree in statistics and interned with a prominent firm. Candidate B is nontraditional: they only have a high school degree and are self-taught; however, B won a popular machine-learning competition. When comparing A and B prior to hire, B's productivity might have a higher expectation, but also higher variance than A's. We will show that delegation results in a bias against candidate B in favor of candidate A.

We restrict attention to refund contracts for two reasons. First, anecdotal and survey evidence suggests that this is the main contract form used in practice.[1] Second, search is private to the recruiter and productivity is private to the firm. Refund contracts recognize this reality, and only condition payment on *employment* which is public and easily measured. We take the termination threshold as given and independent of the contract. For many firms, this is reasonable. The person deciding to terminate an employee is generally not the same person who hires the recruiter. Additionally, large firms are generally required by anti-discrimination law to treat employees fairly or equally (Carlsson, Fumarco, and Rooth 2014). This makes it unlikely that a firm can terminate employees

1. We interviewed three recruiters, and they reported that such a contract was common and that the trial period is typically around 90 days. A survey by Top Echelon found that 96% of recruiters offer some sort of guarantee that a candidate will stay. Among those, 61% provided a replacement - but not money back - if the candidate failed to stay, while 26% offered a partial or full refund.

differently based on the channel through which they were hired.[2]

Our first result characterizes the general contracting problem through a single first-order condition. The characterization shows that refund contracts induce a special type of moral hazard which is equivalent to making the search technology less accurate in a Blackwell sense. It is as if the firm and the recruiter face a canonical multitasking problem, in which the task the firm cares about (maximizing productivity expectation) can only be encouraged together with a wasteful task (minimizing productivity variance). We compare delegation to a first-best benchmark where the firm searches directly for an applicant. Except in knife-edge cases, first-best social surplus and recruitment strategy are not achieved.

We then ask how delegation influences search strategy, and show that when the delegation-induced degradation of the search technology satisfies a weak condition, there is less search effort exerted under delegation. We further show that this implies the expected productivity distribution among selected workers will be lower (in a first-order stochastic dominance sense) under delegation than the first-best. Under an independence assumption, the productivity variance distribution among selected workers is also lower. Finally, we demonstrate social surplus and search effort increase and converge to the first-best benchmark as heterogeneity in productivity variance decreases. This implies that agency loss in this setting crucially depends on the magnitude of differences in productivity uncertainty across candidates. When some workers exhibit less uncertainty (higher variance after search) than others, then the recruiter will waste search effort seeking such workers at the expense of expected productivity.

We explore two testable implications of the model. First, the specific way in which the refund contract warps incentives implies that, even if two workers have the same productivity expectation, the one with lower unobserved productivity variance will be hired. Further, we show that statistical discrimination against a high-variance group is amplified by delegation: it is higher in a world with delegation than in a world with only

2. For a detailed review of the recruiting industry we refer the reader to Cowgill and Perkowski (2020).

direct search. Thus, our model provides a microfoundation for variance-based statistical discrimination. Second, our model predicts that firms are more likely to outsource recruitment during periods and in occupations where observable differences in productivity variance are small across workers. If variance cannot be well-predicted by observables, or productivity variance is indeed similar across worker types, then delegation is more likely.

Taken together, these two mechanisms suggest an interesting vicious cycle. Delegation induces bias against high-variance groups. If low labor market success causes these groups to exit the occupation, then the labor market in the future becomes more homogeneous, with respect to productivity variance. This reduction in heterogeneity makes delegation more appealing, thus increasing the share of firms utilizing recruiters as well as the average bias against high-variance groups.

Our modeling framework is general in that we do not specify an information structure. Instead, we take as primitives the posterior means and variances that result from some updating process. This is similar in spirit to the approach taken in the Bayesian persuasion literature, particularly Gentzkow and Kamenica (2016). A researcher can specify an information structure, derive the implied posterior mean and variance distributions, and apply our results. In our application section, we illustrate this process using a model of hiring which is popular in the discrimination literature that utilizes normal priors and normal signals. Our results can also be applied to models that do not have an information structure at all. Such models include those with match-specific effects and complementarities between types of firms and workers. We describe such an example in the Appendix.

The paper is structured as follows. First, we describe how our work contributes to the existing literature. Second, we introduce the model. Third, we characterize the general solution to the delegated and first-best problems. Fourth, we compare delegation to the first-best benchmark and present comparative statics. Fifth, we apply our results to specific parametric examples and economic situations. Finally, we discuss the implications and conclude.

2.2 Literature

Our main contribution is to the labor market intermediary literature. Within this literature, Cowgill and Perkowski (2020) is the only other paper explicitly focused on agency issues stemming from the firm-recruiter relationship. The authors investigate agency issues arising from balancing firm and worker preferences. Their paper grapples with the idea that recruiters maximize imperfectly aligned objective functions, and to that extent it shares many similarities with our paper. However the papers are otherwise complements: Cowgill and Perkowski (2020) investigates how worker and firm preferences are balanced in light of recruiter reputation while we explore how firms balance misalignment and effort provision when worker productivity is uncertain.

Our paper is most theoretically related to Ulbricht (2016), which explores a general delegated sequential search problem. Like in our setting, Ulbricht considers the case when search is unobserved by the principal, and shows that in an unrestricted contract space, the first-best can be achieved. Unlike Ulbricht (2016), we restrict the set of feasible contracts to what we call refund contracts. We consider the case when searched objects are uncertain and differ in their mean and variance. This combination of a contract restriction and two dimensions of heterogeneity prevents the firm from achieving the first-best.

Our paper is also related to the more general delegated choice literature. The models in this literature feature a principal who must trade-off the comparative advantage of the agent (the agent usually has better information) with the agent's bias. Within this literature, two relevant papers are Armstrong and Vickers (2010) and Alexander Frankel (2014). In both, the preference misalignment between the principal and agent are primitives of the economic model. The authors then focus on optimal delegation schemes given this misalignment. Our paper is different in that we are concerned with how these preferences are misaligned in the first place, and we show that contract restrictions can generate misalignment like that which is described in this literature. For our specific setting with sequential search and refund contracts, we find that the agent overvalues productivity variance and undervalues productivity expectation relative to the principal.

Because our paper extends the delegated choice literature to a specific context, it is similar in spirit to Alex Frankel (2021) and Che, Dessein, and Kartik (2013). Similar to us, these papers explore how a specific form of bias changes delegation. In Alex Frankel (2021), the principal and the agent value hard and soft information about a job candidate differently. In Che, Dessein, and Kartik (2013), the agent values the outside option differently than the principal; there is a sort of status quo misalignment. In our model, the recruiter prefers low-variance candidates more than the firm does, because these candidates are more likely to exceed a minimal level of competence and thus remain at the firm.

More broadly, our work is motivated by a desire to understand the "matching function" which is an important primitive in labor search and matching models. Following Shimer and Smith (2000) and Postel–Vinay and Robin (2002), our paper, like many other papers examining labor market sorting, considers workers with ex-ante productivity heterogeneity. Inspired by the individual specific and match specific productivity components in these models, we can think of the productivity expectation as individual ability, and the productivity variance as generated by match-specific effects which are not known until hire.

2.3 Model

Players and Actions: There is a single risk neutral firm that wants to fill a single job opening. To do this, she hires a recruiter to search for an ideal candidate. The recruiter is risk neutral and operates a sequential search technology over workers. We assume that workers are not players, and are either fired or quit exogenously when $a < 0$. For simplicity, we assume that the firm proposes the contract, and therefore extracts all surplus.

Search Technology, Information and Workers: The recruiter searches sequentially for a worker by paying a cost c per search. After each search, the recruiter observes a pair of attributes (μ, σ) describing the drawn worker's productivity. Specifically, productivity (a)

conditional on the two attributes (μ, σ) is given by:

$$a|(\mu, \sigma) = \mu + \sigma \cdot \varepsilon \tag{2.1}$$

The random variable ε represents the remaining uncertainty about worker productivity. We make the semi-parametric assumption that it has a symmetric distribution, a zero mean, and a variance normalized to 1.[3] We denote probability density function and cumulative density function of ε by f and F, respectively.[4] Because $E[a|\mu, \sigma] = \mu$ and $Var(a|\mu, \sigma) = \sigma^2$, we refer to μ as productivity expectation, and to σ^2 as productivity variance. These attributes are distributed in the population according to a joint CDF G, and each search is an independent draw from G. After observing (μ, σ), the recruiter can offer the current applicant to the firm or continue their search.[5]

Contracts: The firm is restricted to contracts of the form: $t(a) = \alpha - \beta\mathbb{I}\{a < 0\}$. We call these *refund contracts* because α is the recruiter's payment if the search is successful and β is the refund when the employee is fired or quits for any reason.

Payoffs: If the recruiter rejects the contract, his outside option is assumed to be 0. The firm's ex-post profit is realized productivity less any payments to the recruiter: $\pi(a) = a - t(a)$. The recruiter's ex-post utility consists of payments from the firm less total search cost, which is unknown ex-ante, but is c times the number of searches (N) ex-post: $u(a) = t(a) - N \cdot c$. We restrict attention to cases where some search is optimal when the firm operates the search technology directly (in the first-best).[6]

3. An example of such distribution could be $\varepsilon \sim N(0, 1)$. Then $a|(\mu, \sigma)$ would be distributed as $N(\mu, \sigma^2)$.

4. For technical reasons, we also assume that ε has a continuously differentiable positive PDF on $\mathbb{R}$.

5. As is well-known, it is without loss to ignore recall of previously searched workers.

6. We define some search as the expected number of searches is strictly greater than one. When this condition does not hold, the firm can create a degenerate contract with $\beta = 0$, and the problem becomes uninteresting.

2.3.1 Model Comments

Our framework is general, and is a reduced-form of several more structured models. Consider a model where the recruiter infers a worker's productivity based on observable characteristics, updating a prior over a. We do not specify the information structure, and instead focus on G, the joint distribution of posterior expectations and variances. We require only that posterior productivity can be written as in Equation 2.1. One such information structure is when all workers are part of a publicly observable group. Prior beliefs about each group are normal with potentially different variances and expectations. During search, the recruiter observes both group membership and a normal signal of ability, with precision that can be different for each group. Then signal precisions, prior variances and group proportions jointly determine the marginal distribution of σ. Likewise, these components and the prior expectation determine the marginal distribution of μ. μ, σ will not be independent in this case, because they both depend on group membership. We illustrate such an example in our application section and use it to explore how delegation influences statistical discrimination.

Our framework also nests other contexts. Consider a model where productivity has a firm, worker, and match-specific component. If the firm component is public knowledge and the recruiter can uncover the individual component through search, we can think of the match-specific component as a form of residual. Then, μ is the expected productivity of a firm and worker type, and σ^2 is the productivity variance of a firm and worker type. We describe how to map our results to an additive specification of such a matching model in the Appendix.

The firm payoff contains a rather than $\max\{a, 0\}$ because employment is an experience good: the firm must hire the employee in order to learn underlying productivity. One might also wonder why the firm does not receive a discounted sum of future profits when the employee is not fired. That is, why does the firm not receive a payoff of the form $a + \sum_t \delta^t \max\{a, 0\}$? This can be rationalized through symmetric learning and downwards rigid wages. Suppose a is productivity net of wages in an initial trial period. After this period, a is public. If $a > 0$, then the employee can negotiate higher wages so that future net productivity is 0. If $a < 0$, then the employee separates from the employer (due to downward rigid wages). Either way, the firm's profit is a less payments to the recruiter.

It should be noted that our qualitative results extend to these alternative specifications. To see this, note that both $\max\{a, 0\}$ and $a + \sum_t \delta^t \max\{a, 0\}$ are convex functions of a. Thus, when comparing workers with similar productivity expectations but different variances, the firm will tend to prefer higher variance candidates due to their option value. This increases misalignment between the firm and recruiter.

2.4 Analysis

In this section, we analyze the first-best benchmark and the actual equilibrium without imposing additional assumptions on G.

2.4.1 First-Best Benchmark

As a first-best benchmark, we consider the case when the firm can operate the search technology directly.[7] The firm is risk neutral, so it seeks to maximize expected profit. After searching an applicant, expected a is: $E[a|\mu, \sigma] = \mu$. As a result, the firm cares only about productivity expectation (μ). The first-best problem is thus a standard sequential search problem in the style of McCall (1970). The solution is a reservation rule, and the set of acceptable workers is given by an acceptance region of the form $\mathcal{D}_F = \{\mu, \sigma | \mu \geq \mu^*\}$, where μ^* solves:

$$c = \int_{\mu \geq \mu^*} (1 - G_\mu(\mu)) d\mu \tag{2.2}$$

These results are standard in the sequential search literature; in the interest of completeness, however, we provide derivations in the Appendix. To compare the first-best and

7. Equivalently, when there is no contract restriction. The firm would then optimally "sell-the-firm" by taking a fee from the recruiter and allowing the recruiter to be the residual claimant.

equilibrium, it is informative to rewrite Equation 2.2 in the following way:

$$(\mathbb{E}[\mu|\mu > \mu^*] - \mu^*) \cdot \Pr(\mu > \mu^*) = c \tag{2.3}$$

From this, we see that μ^* is equal to expected profit from search:

$$\mu^* = \mathbb{E}[\mu|\mu \geq \mu^*] - \frac{c}{Pr(\mu \geq \mu^*)} \tag{2.4}$$

Since we assume that first-best recruitment is profitable, the right-hand-side is positive and μ^* must also be positive.

2.4.2 Delegation Equilibrium

We now consider a Perfect Bayesian Equilibrium where the firm must delegate search to the recruiter. The firm does not observe the search strategy of the recruiter. The contract space is also restricted to refund contracts. Such contracts consist of an upfront payment (α) and a refund (β) which is returned to the firm if the candidate terminates. We require that any contract be both individually rational and incentive compatible.

Incentive compatibility requires that the search strategy the firm requests must be the recruiter's optimal sequential search strategy after the contract is accepted. The recruiter is concerned solely with avoiding a refund. Upon searching for a worker, expected utility from selecting that worker given μ, σ is:

$$\alpha - \beta\mathbb{E}[\mathbb{I}\{a < 0\}|(\mu, \sigma)] = \alpha - \beta\left(1 - F\left(\frac{\mu}{\sigma}\right)\right) = (\alpha - \beta) + \beta F\left(\frac{\mu}{\sigma}\right) \tag{2.5}$$

where F denotes the standard error ε cumulative density function (CDF). Equation 2.5 shows that the ratio μ/σ is a sufficient statistic for the recruiter. We will call this ratio *standardized productivity* throughout this paper, and denote it $\tilde{\mu}$. Intuitively, it indexes how likely the worker's productivity is to be satisfactory (above 0). The recruiter searches through workers, evaluating them based on Equation 2.5. Ignoring the constant part of her utility, the decision to terminate search and hire the current worker is given by the following Bellman equation.

$$U = -c + \int \max\{\beta F(\tilde{\mu}), U\} d\tilde{G}(\tilde{\mu}) \tag{2.6}$$

U is the continuation value and $\tilde{G}$ is the CDF of $\tilde{\mu}$ derived from the joint CDF of (μ, σ). Equation 2.6 emphasizes that the recruiter's problem again reduces to standard sequential search, after we note that $\tilde{\mu}$ is a sufficient statistic and the function $\beta F(x)$ is strictly increasing for positive β. Similar to the first-best benchmark, we know from well-known properties of sequential search that the solution is a reservation rule in $\tilde{\mu}$. We formalize this in the following lemma.

Lemma 3 *In any incentive compatible contract, the recruiter's acceptance region takes the form:*

$$\mathcal{D}_R = \{\tilde{\mu} | \tilde{\mu} \geq \tilde{\mu}^*\}$$

where $\tilde{\mu}^*$ solves:[8]

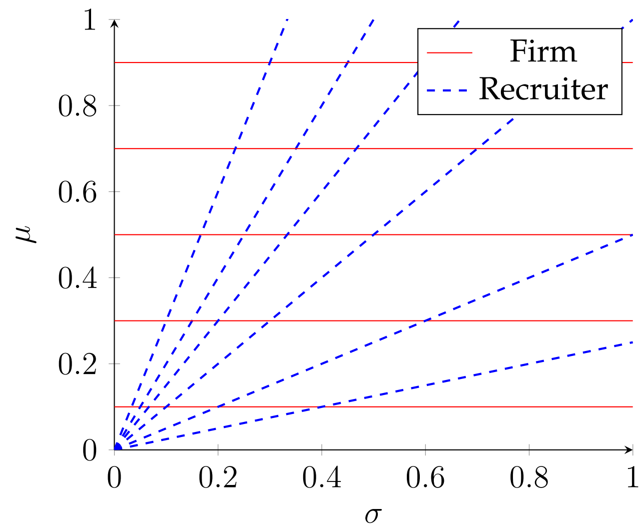
$$c = \int_{\tilde{\mu} \geq \tilde{\mu}^*} \beta F(\tilde{\mu}) - \beta F(\tilde{\mu}^*) d\tilde{G}(\tilde{\mu}) \tag{IC}$$

The proof of Lemma 3 is in the Appendix. The economic intuition is as follows. The refund contract encourages the recruiter to care about standardized productivity, rather than expected productivity. There will be a fundamental misalignment between the firm and recruiter, which can be visualized by graphing the firm's isoprofit curves and the recruiter's indifference curves over the space of worker types. We focus on the case when $\beta > 0$, which we will later show is the most relevant case.

The recruiter's indifference curves all emanate from the origin, with higher indifference curves being more steeply sloped. In any incentive compatible contract, the recruiter

8. This formulation is true for continuously distributed $\tilde{\mu}$ and a proper interior solution (non-degenerate search), but can be easily generalized to a system of inequalities otherwise.

Figure 2.1: Indifference and Isoprofit Curves Over Worker Types



Note: The figure demonstrates the fundamental misalignment between the recruiter and the firm induced by the two dimensions of worker heterogeneity and the refund contract.

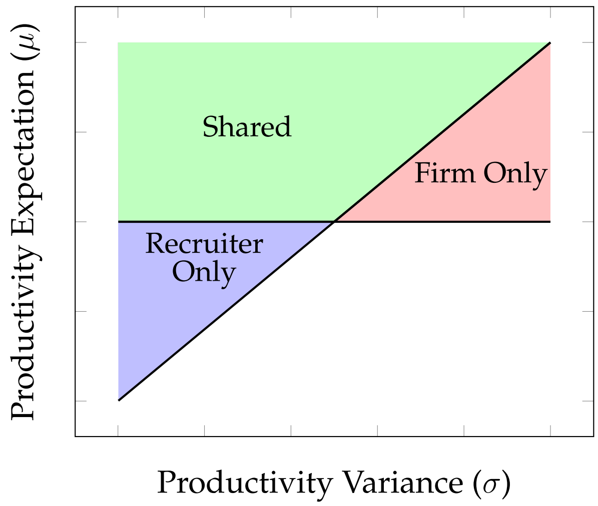
will attempt to minimize productivity variance more than the firm would like, in order to climb to a higher indifference curve. An implication of Figure 2.1 and Lemma 3 is that the recruiter's acceptance region will be triangular[9], whereas the firm's will be rectangular. Figure 2.2 illustrates this.

The figure displays three important partitions of the worker type space. In the recruiter only region are the low-variance, low-expectation workers the recruiter hires but the firm would prefer excluded. We refer to these as "safe bets," and they are inefficiently hired in equilibrium. In the firm only region are high-variance and high-expectation workers that the firm would like to hire, but the recruiter excludes. We call these workers "diamonds in the rough," and they are inefficiently excluded in equilibrium. In the Appendix, we show that under general conditions the acceptance regions will not be subsets of each other. A direct consequence of this is that there will always be a positive measure of diamonds in the rough and safe bets.

Figure 2.2 also provides intuition about how the equilibrium contract is determined. The firm effectively chooses the slope of the diagonal line that defines the three regions.

9. or trapezoidal if the support of expected ability does not include 0.

Figure 2.2: Recruiter vs. Firm Acceptance Regions Over Applicant Types



Note: In a delegated equilibrium the optimal contract will induce an acceptance region which is a distinctly different shape then the first-best acceptance region. Inefficiency is driven by the sub-optimal inclusion of the blue region and the sub-optimal exclusion of the red region.

Examining Equation IC reveals β controls the slope.[10] More powerful incentives (higher β) increase the slope. A steeper slope increases the share of workers that are inefficiently excluded, but decreases the share that are inefficiently included.

We now begin to characterize equilibrium. We have just shown that $\tilde{\mu}$ is a sufficient statistic for the recruiter. In order to focus on non-trivial cases, we introduce a weak assumption about the relationship between standardized productivity and expected productivity.

Assumption 1 $\mathbb{E}[\mu|\tilde{\mu} = x]$ *is weakly increasing in* x.[11]

For the rest of the paper, it is assumed to be satisfied. Intuitively, Assumption 1 means that larger standardized productivity implies larger expected productivity of the candidate. This assumption is quite natural, given that $\tilde{\mu} = \mu/\sigma$.[12] If the conditional expectation is

10. Applying the implicit function theorem reveals that $\tilde{\mu}^*$ is increasing in β.

11. This condition is often referred to in the statistics literature as positive quadrant dependence in expectation, which is slightly weaker than positive quadrant dependence, and much weaker than positive affiliation.

12. To break this assumption, the association between μ, σ needs to be so strong that the expectation grows faster than linearly. For example, $\sigma = \gamma\mu^2$ will cause the conditional expectation to be decreasing.

flat, then the problem becomes uninteresting. There is no way to encourage the recruiter to search strategically, and thus the firm will either not hire or offer a degenerate contract, where $\beta = 0$ and the recruiter returns the first applicant they searched.

We are now ready to characterize the delegation equilibrium. An equilibrium contract consists of an upfront payment α, contingent refund β and acceptance region $\mathcal{D}_R$, such that:

1. The firm maximizes profit.

2. The recruiter accepts the contract (individual rationality).

$$\mathbb{E}[t(a)|D_R] - \frac{c}{Pr((\mu, \sigma) \in D_R)} \geq 0 \tag{IR}$$

3. The acceptance region $\mathcal{D}_R$ is the optimal sequential search strategy of the recruiter, given the contract details (incentive compatibility).

This defines a quite general problem, with two dimensional sequential search and moral hazard. However, our prior discussion of the recruiter's problem, particularly Lemma 3, allows us to characterize the solution in a simple way. Lemma 3 shows that the acceptance region in the delegated search problem is defined by a reservation rule $\tilde{\mu}^*$. Moreover, any threshold $\tilde{\mu}^*$ can be induced by an incentive compatible and individually rational contract. As all parties are risk neutral and the firm pays the initial transfer α only to keep the recruiter indifferent between accepting and rejecting the contract (IR binds), then the firm extracts all social surplus and only cares about the acceptance region induced by the contract.

Theorem 2 *The delegated search equilibrium is given by the solution to a standard sequential search problem over $\mathbb{E}[\mu|\tilde{\mu}]$. The solution is determined by a reservation rule $\tilde{\mu}^*$, which solves:*

$$(\mathbb{E}[\mu|\tilde{\mu} \geq \tilde{\mu}^*] - \mathbb{E}[\mu|\tilde{\mu} = \tilde{\mu}^*]) \cdot \Pr(\tilde{\mu} \geq \tilde{\mu}^*) = c \tag{2.7}$$

and by (IC) and (IR) refund contract which induces the recruiter to select workers with $\tilde{\mu} \geq \tilde{\mu}^*$

Corollary 2.1 *The firm's profit under delegation is positive and equal to $\mathbb{E}[\mu|\tilde{\mu} = \tilde{\mu}^*]$. As a result, the optimal threshold $\tilde{\mu}^*$ is also positive.*

Theorem 2 has practical significance, proving that the general contracting problem is characterized by the solution to a much simpler problem. Indeed, the entire search strategy under moral hazard is defined by a threshold rule, which is uniquely pinned down by a single first-order condition. The entire problem essentially collapses into standard one-dimensional sequential search over a transformed distribution.

Theorem 2 also holds deeper economic insight. Comparing equations 2.7 and 2.3, we see that the equation characterizing equilibrium is identical to the one characterizing the first-best if we replace μ with $\tilde{\mu}$. Imposing the binary refund contract is equivalent to allowing the firm to search itself over $E[\mu|\tilde{\mu}]$ rather than μ. Thus, moral hazard in this setting has the effect of making the search technology more blunt or less accurate. The contract restriction requires the firm to resort to using the imperfect signal $\tilde{\mu}$ as a proxy for μ. In this way, search is noisy under delegation. This realization forms the foundation for the rest of our results.

2.5 Results

In the prior section, we provide a general characterization of the delegation equilibrium. In this section, we present first-best comparisons and comparative statics under minimal restrictions on the joint distribution of μ, σ. As shown in Section 4, first-best search is over μ directly while delegated search is over $\mathbb{E}[\mu|\tilde{\mu}]$. The latter has a distribution which is a mean preserving contraction of the former one as it is based on observing noisy "signal" $\tilde{\mu} = \mu/\sigma$ instead of observing μ explicitly.

In some special cases, the two distributions could be identical (for instance, if σ has a degenerate distribution). Thus, we begin by showing exactly when first-best profit and search strategy are not achieved under delegation. Throughout, we refer to strictly lower profit and a different search strategy under delegation as "not achieving the first-best."

Definition 3 *Let two random variables with CDFs F and G have compact supports S_F and S_G. F is a **strict mean preserving spread** of G if their expectations are the same and $\forall x \in$ int S_F*

$$\int_x^{\bar{x}} (1 - F(s))ds > \int_x^{\bar{x}} (1 - G(s))ds$$

This definition is a slightly stronger notion of a mean-preserving spread. For the distributions of μ and $\mathbb{E}[\mu|\tilde{\mu}]$, the difference is as follows. If for some x, y it is true that thresholds on μ and $\tilde{\mu}$ select the same set of workers $\{(\mu,\sigma)|\mu \geq x\} = \{(\mu,\sigma)|\tilde{\mu} \geq y\} \notin \{\emptyset, \{(\mu,\sigma)\}\}$, then the inequality in the definition is violated (it becomes an equality) and for a specific search cost c, the first-best and the second-best acceptance regions are the same. This, however, only happens for very specific distributions of (μ,σ). This result is formalized in the next proposition.

Proposition 8 *First-best is not achieved in equilibrium for any search cost if and only if μ is a strict mean-preserving spread of $E[\mu|\tilde{\mu}]$.*

The formal proof consists of analyzing the firm's profits in two cases and the corresponding first-order conditions. The first order conditions equalize the excess wealth order – integrals from the definition – to the search cost c. As they are never equal to each other, the optimal thresholds can never be the same in the two cases. Moreover, the threshold must be strictly lower for the distribution of $\mathbb{E}[\mu|\tilde{\mu}]$, and thus, the profit must also be lower than it is in the first-best.

Expected productivity (μ) is always a mean-preserving spread of $E[\mu|\tilde{\mu}]$ using the standard definition. It will also be a strict mean-preserving spread, except for in a few knife-edge cases. Therefore, the proposition implies that the first-best is generally not achieved. We call the special cases when first-best is achieved knife-edge because they require a specific type of degeneracy in the joint distribution of (μ,σ). Our intuition suggests the first-best should not be achievable, because incentive compatibility requires the acceptance regions to be fundamentally different shapes. This is visualized in Figure 2.2. The proposition simply spells out when our intuition is violated.

As we showed earlier, the contract restriction makes the delegation problem equivalent to direct search with a less accurate search technology. Productivity variance is what garbles the search technology: when productivity variance is degenerate, standardized productivity ($\tilde{\mu}$) becomes a perfect signal of expected productivity, and the first-best is achieved.

Thinking about $\tilde{\mu}$ as a signal for μ allows us to view Proposition 8 through the lens of Blackwell (1953). First-best is not achievable whenever μ is statistically sufficient for $\tilde{\mu}$ in a strict sense, or put another way when μ dominates $\tilde{\mu}$ in the Blackwell order in a strict sense. Incentive compatibility and the refund contract convolute μ and σ in a specific way, such that the joint distribution of μ, σ will impact the equilibrium outcome. We will explore the mechanics of this relationship in the next few results.

With this established, we next wish to understand how the characteristics of accepted workers compare under delegation and the first-best. Particularly, we would like to see how the distribution of the workers' productivity expectation and variance differ in the delegated search and the first-best search benchmarks. We know that the firm searching directly cares more about the candidate's productivity expectation, and does not care about productivity variance at all. However, the recruiter prefers lower productivity variance. Intuitively, we should anticipate higher productivity expectation as well as higher productivity variance in the first-best than under delegation.

We begin with the productivity variance of accepted workers. When variance and expectation are independent, the distribution of variance among workers accepted in the first-best should be equal to the population distribution. The firm wishes to ignore productivity variance. However, the recruiter cares about productivity variance, and even when the two attributes are independent, the acceptable workers under delegation will tend to be lower variance than the general population. This is formalized in the following proposition.

Proposition 9 *If μ and σ are independent, then the productivity variance of accepted workers in the first-best is higher (in a first-order stochastic dominance sense) than under delegation.*

One way to understand this result is that refund contracts induce a special type of moral hazard, which biases the recruiter in favor of low-variance candidates, even when both the firm and the recruiter are risk neutral. A crucial question here is whether this bias in favor of low-variance candidates comes at the expense of high-expectation candidates. We turn to this next.

It follows directly from Proposition 1 that, because social surplus is higher in the first-best, expected productivity is higher on average than under delegation. However, under a broad set of circumstances, we can say that the entire distribution of productivity expectations of selected workers is higher in the first-best than under delegation, in a first-order stochastic dominance sense. A sufficient condition for this relies on search effort. We will define **search effort** as the percentage of candidates that are unacceptable, and denote it q. This object maps one-to-one with the expected number of searches, which is $1/(1-q)$. Thus search effort is itself an equilibrium outcome worth understanding. We anticipate less search effort in the delegated problem: if the search technology is less accurate, then the expected benefit of another search should be lower under delegation. Later in the section, we characterize necessary and sufficient conditions when this anticipated result holds. We first illustrate the connection between search effort and the distribution of productivity expectations in Lemma 4.

Lemma 4 *If search effort is lower under delegation than the first-best, then the productivity expectation of accepted workers in the first-best is higher (in a first-order stochastic dominance sense) than under delegation.*

To analyze search effort, we first observe that the strict mean-preserving spread condition from Proposition 8 does not generate a clear comparison of search effort under first-best and delegation. We need a stronger concept. As Chateauneuf, Cohen, and Meilijson (2004) and Zhou (2020) prove, one such concept is excess wealth order. We introduce this concept next, and prove that it is necessary and sufficient for generating clear comparative statics for all search costs.

Definition 4 *Let random variable X have a smooth CDF F. The excess wealth at threshold x^* is the expected benefit of one additional search if we already have x^*. It can be expressed as:*

$$EW_X(x^*) = E[(X - x^*)^+] = \int_{x^*}^{\infty} (1 - F(x))dx$$

To derive the necessary and sufficient condition for comparative statics in search effort, recall that the first-order condition of a sequential search problem over a random variable X with CDF F is:

$$(\mathbb{E}[X|X \geq x^*] - x^*)Pr(X^* \geq x) = EW_X(x^*) = c$$

Rewrite the equation in terms of quantiles of X:

$$EW_X(F^{-1}(q)) = c$$

The left-hand side is decreasing in q, implying that for a variable with greater excess wealth, the q that solves the equation will be also higher. To formalize this result, we define excess wealth order and state Theorem 3.

Definition 5 *A variable X_1 with CDF F_1 dominates a variable X_2 with CDF F_2 in the excess wealth order if:*

$$EW_{X_1}(F_1^{-1}(q)) \geq EW_{X_2}(F_2^{-1}(q)) \; \forall \, q \in (0, 1)$$

Theorem 3 *For any cost c search effort is greater for X_1 than for X_2 if and only if X_1 excess wealth order dominates X_2.*

Excess wealth order is a well-known variability order, and is discussed at length in Shaked and Shanthikumar (2007). It is sometimes referred to as the right-spread order. Excess wealth order implies that X_1 is a mean-preserving spread of X_2 if they have the same expectations, but the converse is not true. In this sense, it is stronger than the concept of a mean-preserving spread.[13] With this in hand, we present the following result.

Corollary 3.1 *If μ excess wealth order dominates $\mathbb{E}[\mu|\tilde{\mu}]$, then the search effort is greater in the first-best than in the delegated search benchmark. Therefore, the productivity expectation of accepted workers in the first-best is higher (in a first-order stochastic dominance sense) than under delegation.*

In the parametric examples section, we illustrate that for all joint lognormal or Pareto distributed (μ, σ), μ dominates $\mathbb{E}[\mu|\tilde{\mu}]$ in the excess wealth order. This is striking in the lognormal case: for arbitrarily high positive correlation between (μ, σ) dominance is maintained. In this sense, excess wealth order is not overly restrictive, and we have indeed shown that it is the least restrictive ordering needed for comparing equilibrium and first-best search effort.

We now reconsider the example discussed in the introduction. Candidate A, with a traditional resume, can be thought of as a safe-bet: low-variance, low-expectation. Candidate B, with a non-traditional resume, can be thought of as a diamond in the rough: high-variance, high-expectation. If the recruiter shared the firm's preferences, then he would select solely based on expected productivity and choose B. However, the refund contract introduces misalignment. The recruiter does not care about expected productivity, but rather how likely each candidate is to be above the firing threshold. This induces the recruiter to care about productivity variance, which will therefore cause the recruiter to inefficiently hire A over B.

Corollary 3.1 allows us to formalize the intuition from the example. While we already know that first-best social surplus is not achieved in equilibrium, we can say more: part of the lost social surplus is because the recruiter focuses search effort on finding low-variance candidates. The corollary proves that search is less efficient precisely because the recruiter is wasting his search effort minimizing variance instead of maximizing expectations. This manifests in a pool of acceptable workers which contains too few

13. However, it is weaker than the more widely-used dispersive order which, in past work, has been used to derive comparative statics in search intensity or duration.

high-expectation candidates and too many low-variance candidates. As we discuss in the application section, this proposition implies that intermediation results in more variance-based statistical discrimination. The refund contract restriction results in an inefficient bias against high-variance candidates, despite risk neutrality of all actors.

Finally, we investigate how search effort, social surplus, and equilibrium profit change as the underlying distribution of workers in the labor market changes.

Proposition 10 *Consider a parameter θ of the joint distribution of μ, σ such that:*

1. *The marginal distribution of μ does not depend on θ*
2. *$\mathbb{E}_{\theta_1}[\mu|\tilde{\mu}]$ dominates $\mathbb{E}_{\theta_2}[\mu|\tilde{\mu}]$ in the excess wealth order for all $\theta_1 > \theta_2$.*
3. *$\mathbb{E}[\mu|\tilde{\mu}]$ converges to μ in distribution as θ converges to some value $\bar{\theta}$*

Then social surplus, equilibrium profit, and search effort are increasing in θ, and converge to first-best values as $\theta \to \bar{\theta}$

Restated verbally, the distortions and inefficiencies caused by delegation depend on the level of heterogeneity in the labor market. The proposition also illuminates a natural connection between our model and canonical multitasking models (Holmstrom and Milgrom 1991). To see this connection, suppose we define two "tasks": search along the expectation (μ) dimension and search along the variance (σ) dimension. Like in many multitasking models, the firm cannot provide incentives for each task individually, and can only encourage search over an aggregate metric. In our case, this metric is standardized productivity ($\tilde{\mu}$). If we take logs we have:

$$log(\tilde{\mu}) = log(\mu) - log(\sigma)$$

How do we interpret this expression? Well, it implies that the firm can only "buy" an increase in μ if it is willing to also "buy" a reduction in σ. Thus, we are in a situation where total search effort is rewarded, but there is a wasteful task which cannot be properly distinguished from the productive task. In equilibrium, this manifests in workers that are first-order stochastically dominated in terms of productivity expectation, but also first-order stochastically dominated in terms of variance compared to the first-best. In the proposition, raising θ effectively removes heterogeneity in productivity variance, the wasteful dimension of search. Removing this heterogeneity makes the recruiter focus on maximizing productivity expectation. In the extreme case when almost all heterogeneity in σ is removed, and $\mathbb{E}[\mu|\tilde{\mu}]$ converges to μ, first-best is achieved.

2.6 Parametric Examples

In the last section, we illustrate the main economic forces qualitatively, without imposing parametric assumptions on G, the distribution of productivity variance and expectation. In this section, we show more explicit results using specific parametric joint distributions.

2.6.1 Lognormal

Assumption 2 ***(Lognormal Productivity)*** *μ, σ are distributed joint lognormal. That is:*

$$\begin{pmatrix} log(\mu) \\ log(\sigma) \end{pmatrix} \sim N\left(\begin{bmatrix} m_\mu \\ m_\mu \end{bmatrix}, \begin{bmatrix} s_\mu^2 & s_{\sigma,\mu} \\ s_{\sigma,\mu} & s_\sigma^2 \end{bmatrix} \right)$$

Lognormal distributions are a common way to model positive-valued economic objects. The ability to easily incorporate correlation between μ, σ makes the parameterization more flexible. This also makes it a more credible way to model the posteriors generated by a proper information structure as posterior means and variances will generally not be independent.

In the Appendix, we provide and prove a series of lemmas that allow us to link our general results to the lognormal family of distributions. Two aspects of lognormal random variables make the analysis tractable. First, the multiplication of two lognormal random variables is, again, lognormal. This, taken together with the normal projection formula, implies that $E[\mu|\tilde{\mu}]$ will remain lognormal. Thus, first-best and equilibrium search will be over distributions within the same family, making the resulting problems easier to compare. Second, we show that two lognormal random variables with the same shift parameter m can be ranked in the excess wealth order by their shape parameters, s. Because $E[\mu|\tilde{\mu}]$ is a mean-preserving contraction of μ and both are lognormal, this makes comparison in the excess wealth order straightforward.

The last step needed to apply our general results is to observe that the shape parameter s_σ satisfies the conditions for θ laid out in Proposition 10. In particular, because s_σ is exactly the variance of $log(\sigma)$, it is clear to see how degeneracy of σ is achieved as it approaches 0.

Proposition 11 *If productivity variance and expectation are distributed according to Assumption 2 and further, there is not perfect positive dependence, then:*

1. *The first-best profit and acceptance region are not achieved.*
2. *Search effort is strictly greater in the first-best than under delegated search.*
3. *The productivity expectation of hired workers is higher in the first-best than delegated search.*
4. *Profit and search effort increase as s_σ decreases, and converge to first-best levels as $s_\sigma \to s^2_{\sigma,\mu}/s^2_\mu$.*

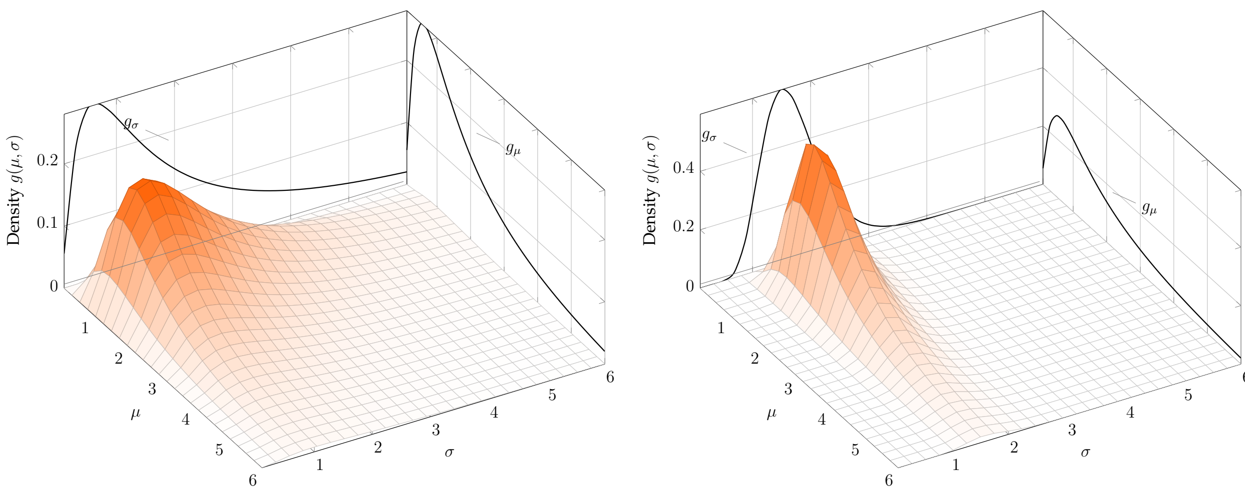
If additionally $s_{\sigma,\mu} = 0$, the productivity variance of hired workers is higher in the first-best than delegated search.

A nice feature of the result is that it is true for virtually all lognormal joint distributions, with arbitrarily negative or positive correlations. This illustrates that the necessary and sufficient excess wealth order condition is a quite weak requirement for some distributional families, further implying that our results regarding the under-provision of search effort and failure to achieve first-best are robust.

The fourth part of the proposition has a useful interpretation in terms of multiplicative noise. Consider a zero mean lognormal random variable Z that is independent of productivity variance and expectation. Multiplying productivity variance by Z is equal in distribution to increasing s_σ while holding all other parameters fixed. Therefore, we can use the proposition to conclude that additional multiplicative noise, or additional multiplicative heterogeneity in the variance dimension, makes search less efficient and reduces search effort.

We visualize what occurs when s_σ decreases by plotting two bivariate lognormal distributions with different values of s_σ in Figure 2.3. Decreasing s_σ has the effect of collapsing the distribution along the σ dimension, which makes the recruiter less tempted to waste search effort.

Figure 2.3: Lognormal Joint Distribution



(a) Example with $s_\sigma = 1$ **(b)** Example with $s_\sigma = 0.3$

Note: The panels plot an example distribution with $m_\mu = 0.5, m_\sigma = 0.5, s_\mu = 1, \rho = 0$. The only difference is that s_σ is smaller in panel b than a. This represents a reduction in heterogeneity in productivity variance. The marginal distributions are projected.

2.6.2 Pareto

In this section, we assume the productivity attributes follow Pareto distributions. This assumption is convenient because it yields closed-form solutions for the threshold rules. In turn, these closed-form solutions allow straightforward comparative statics.

Assumption 3 ***(Pareto Productivity)*** *μ, σ are distributed independently with marginal Pareto* *distributions. That is, their joint probability density function is given by:*

$$g(\mu, \sigma) = \frac{\theta_\mu x_\mu^{\theta_\mu}}{\mu^{\theta_\mu+1}} \frac{\theta_\sigma x_\sigma^{\theta_\sigma}}{\sigma^{\theta_\sigma+1}} \mathbb{I}\{\mu \geq x_\mu\}\mathbb{I}\{\sigma \geq x_\sigma\}$$

where both variables have finite expectations ($\theta_\mu > 1, \theta_\sigma > 1$).

Under a Pareto assumption, the random variable $\mathbb{E}[\mu|\tilde{\mu}]$ is almost Pareto: it has an atom at the beginning of the distribution, and is Pareto conditional on being above the atom. This is convenient for analysis because it implies that the equilibrium reservation rule will have a closed-form solution. We provide these closed form solutions in the Appendix. Additionally, we focus on the case where delegated search has an interior solution.[14]

The most important parameter for our analysis is the shape parameter θ_σ. In terms of interpretation, θ_σ represents the level of heterogeneity with respect to productivity variance in the worker pool. We show this visually in Figure 2.4, which plots three different Pareto densities with different values of θ_σ and $x_\sigma = 1$. As the notation implies, θ_σ satisfies all of the conditions laid out in Proposition 10. In particular, as θ rises, $E[\mu|\tilde{\mu}]$ rises in the excess wealth order.[15] As $\theta_\sigma \to \infty$ we approach a perfectly homogeneous population with respect to productivity variance. Under this distributional assumption, we can apply our nonparameter results to show the following result.

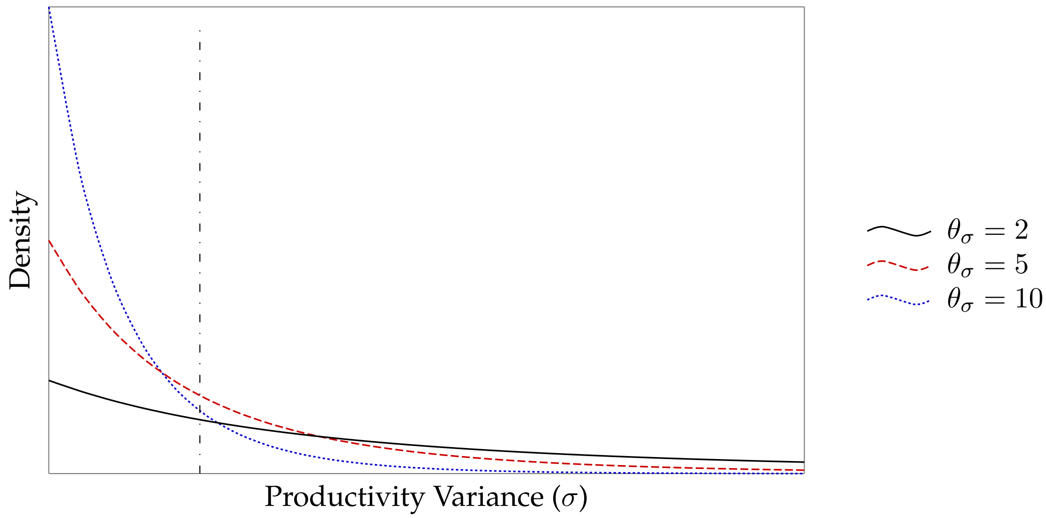
Proposition 12 *If productivity variance and expectation are distributed according to Assumption 3, then:*

1. *The first-best profit and acceptance region are not achieved.*
2. *Search effort is greater in the first-best than under delegated search.*
3. *The productivity variance and productivity expectation of hired workers is higher in the first-best than delegated search.*

14. In the Appendix we show an interior solution is guaranteed if $\frac{x_\mu \theta_\sigma}{(\theta_\mu+\theta_\sigma-1)(\theta_\mu-1)} \geq c$.

15. See the Appendix for a full proof.

Figure 2.4: Densities of σ for Different Values of θ_σ



Note: The figure displays different marginal distributions of σ when the shape parameter is changed. It captures the idea that as θ_σ rises heterogeneity declines.

4. *Profit and search effort increase as θ_σ increases, and converge to first-best levels as $\theta_\sigma \to \infty$.*

The proof is provided in the Appendix, and is a direct consequence of the results presented in the general analysis section. The Pareto parameterization makes clear the connection between Proposition 10 and labor force heterogeneity. Increasing the θ_σ parameter reduces heterogeneity in the variance dimension. Since σ is a wasteful dimension of search in terms of social surplus, reducing heterogeneity in that dimension lowers the returns to searching along that dimension, and thus raises efficiency.

2.7 Applications

2.7.1 The Choice to Delegate

In the first-best benchmark, the firm operates the search technology directly. Therefore, the difference between first-best profit and profit under delegation is exactly the agency loss incurred by the firm. We can use this observation and Proposition 10 to understand the decision to delegate.

A firm must perform a cost-benefit analysis when deciding whether to delegate the recruiting function. The benefit of delegation is the comparative advantage of the recruiter. The cost is agency loss: as we have shown, there are fundamental differences between how the recruiter and the firm order potential workers. We can model this by supposing that, prior to designing the contract with the recruiter, the firm must decide whether it will delegate at all. If it chooses not to delegate, it can perform search directly at a different search cost c_F, which is strictly larger than the recruiter's search cost c.

The firm will decide to delegate when the comparative advantage of the recruiter outweighs the agency loss of delegation. We can apply Proposition 10 to understand how changes in the labor force impact the decision to delegate.

Proposition 13 *As heterogeneity in productivity variance decreases, the firm is more likely to delegate. When workers are homogeneous with respect to productivity variance, the firm will always delegate.*[16]

When the recruiter and the firm face the same search cost, we know from prior results that direct search is always more profitable than delegated search. However, when the recruiter has a lower search cost, there is the possibility that delegated search is more profitable. As heterogeneity in the wasteful dimension vanishes, profit under delegation rises, while profit from direct search stays constant. Eventually, when all agency loss is gone, only the comparative advantage effect remains, and delegated search is optimal.

If we think about different occupations as having different labor market pools, recruiter utilization in each occupation will depend on the amount of heterogeneity in productivity variance across workers. Occupations where workers have similar productivity variance will feature higher recruiter utilization, and occupations with more heterogeneous productivity variance will feature a larger share of firms recruiting directly.

One crucial distinction to make here is that all results in this paper are concerned with differences in variance, not the overall level of variance. Some occupations may have higher average levels of productivity variance. This does not, however, imply that those

16. In our model, a decrease in heterogeneity corresponds with a decrease in θ, in the same sense as Proposition 10.

occupations will have higher recruiter usage. What matter are the differences in variances among workers in the same occupation. As differences in variance become larger within an occupation, recruiter utilization becomes less likely.

2.7.2 Statistical Discrimination

In this section, we show delegation can amplify statistical discrimination. Along the way, we illustrate how to map an explicit information structure with a common prior to our primitives (μ, σ). We believe this section may be of interest to researchers who want to apply our framework to data.

Consider an economy in which all workers can be divided into two groups. The recruiter and firm have a common prior over productivity which is normal and the same for both groups: $a \sim N(\mu_0, \sigma_0^2)$. After searching a worker, the recruiter observes the group membership of the worker and a signal about productivity. The recruiter does not observe the group prior to the draw (there is no directed search). The signal takes a normal form: $Y = a + \xi_i^n$, $\xi_i^n \sim N(0, \tau_i^{-2})$, where i indexes the groups A and B. Notice that the only difference between the two groups is that they have different signal precision or information quality. Without loss, suppose that Group A has better signal precision: $\tau_A > \tau_B$. We can interpret this as the recruiter *better understanding*, the work histories of individuals from Group A.

First, we map this situation to our primitives. The recruiter Bayesian updates their normal prior with the normal signal, generating a normal posterior belief about productivity given the signal:

$$a|Y = y \sim N\left(\frac{\tau_i^2}{1/\sigma_0^2} y + (1 - \frac{\tau_i^2}{1/\sigma_0^2})\mu_0, \frac{\sigma_0^2 \tau_i^{-2}}{\sigma_0^2 + \tau_i^{-2}}\right)$$

From this expression, we can see that σ will take two values with equal probability. Because the precision of the signal for Group A is higher ($\tau_A > \tau_B$) we have that $\sigma_B > \sigma_A$. When the recruiter is faced with two candidates with the same expected productivity but from different groups, it will prefer the candidate from Group A. Additionally, $\mu_i = E[a|\xi_i]$ (that is expected productivity unconditional on the signal realization but conditional on group) will be normal.

We say there is statistical discrimination if the probability a hired worker is from Group A is higher than the probability that a hired worker is from Group B. Notice that even the first-best can feature statistical discrimination. Indeed, this is precisely what was pointed out in Heckman (1998): even if a decision maker uses the same expected productivity threshold for hiring all groups (meaning there is no taste-based discrimination) differences in productivity variance can cause arbitrary statistical discrimination in either direction. Whether the first-best features statistical discrimination against the high-variance group depends on whether the optimal threshold lies above or below the prior mean, $\mu^* > \mu_0$.

However, we wish to understand how delegation impacts statistical discrimination. Does delegation tend to reduce or increase statistical discrimination against group B? More formally, is the probability a hired worker is from Group B higher or lower than the first-best? If it is higher, Group B workers will have more labor market success under delegation. If it is lower Group B workers will more success under direct search.

Proposition 14 *The probability the hired worker is from Group B is lower under delegation than the first-best. Therefore, variance-based statistical discrimination is greater under delegation than the first-best.*

The proof is provided in the Appendix. Prior to this result we established that individuals with high productivity variance face a disadvantage under delegation. The proposition makes a stronger claim. If we think about the probability a hired worker is from Group B as the employment rate of Group B (this would be true under monopsony), then the result implies Group B will have a lower employment rate when search is delegated to a recruiter than when search is conducted directly. Put another way, Group B is unambiguously worse off under delegation and Group A is unambiguously better off. Recall that this occurs despite the fact that the two groups have identical underlying productivity distributions.

The proposition also highlights the potential power of homophily. One reason $\tau_A > \tau_B$ might be because the recruiter is a member of Group A. They better understand resumes or work histories from people who are similar to them. Without delegation, it is unclear which group fairs better. It will depend on the level of the hiring threshold μ^*. But when there is informational homophily delegation reduces the labor market success of Group B (the out group).

Suppose Group A workers are members of a racial majority, while Group B workers are members of a racial minority. Our model suggests that delegation can lead to discrimination against a minority group even in the absence of explicit bias. It is sufficient that people understand those of the same racial group better. This might explain racial hiring gaps persists in many occupations despite explicit verbal commitments by leaders to improve diversity. For example, black representation in computer and math occupations fell between 2002 and 2016 (Muro, Berube, and Whiton 2018). Overall, this section illustrates how delegation can form a microfoundation for variance-based statistical discrimination.

2.7.3 A Vicious Cycle

Under our model, we have established that recruiter utilization should be higher in occupations where the labor market is more homogeneous with respect to productivity variance. On the other hand, we have also established that delegation exacerbates statistical discrimination against groups or individuals with higher productivity variance. These two mechanisms dynamically interact in an interesting way.

Consider an occupation with an initially heterogeneous labor pool. Suppose firms decide whether to use a recruiter. In period 1, only the firms with the highest opportunity cost will choose to use a recruiter, meaning that the majority of hiring will take place directly. However, applicants with high productivity variance will still face a slight disadvantage, and be hired at slightly lower rates. Suppose workers become discouraged after lack of hiring success. They leave the occupation before the next period. In period 2, the workers who leave are replaced by workers with the same characteristics as the initial labor force. Even though new workers enter, the labor market will be more homogeneous with respect to productivity variance due to the disproportionate outflow of high-variance workers. This will increase the share of firms using recruiters. In turn, this increases the outflow of discouraged high-variance workers. As time progresses, a vicious cycle unfolds: recruiter utilization rises and the labor market becomes more homogeneous.

If we think productivity variance is connected to factors like socioeconomic status, this vicious cycle will tend to work against efforts to make the occupation more diverse. Crucially, this argument relies on discouraged workers leaving the occupation. It is not clear how powerful a force this cycle represents, and how it would be impacted by a fully specified search and matching model with congestion effects and other aspects. However, it is an observable implication of our model and one worth mentioning for future work.

2.8 Discussion

2.8.1 Beyond Recruiters

In our framework, the recruiter is induced to act risk averse. The specific mechanism we provide that generates this behavior is the refund contract, which is prolific among recruiters that are hired externally. However, there are many reasons why employees of the firm may also exhibit induced risk aversion. For example, an article from a leading human resource association mentions that many internal recruiters and human resource staff have bonuses based on the cost of hiring and the time to fill a position (Hirshman 2018). Such cost-based metrics ignore productivity, and can bias an agent in favor of candidates who do not have outside offers. The same article also suggests turnover as a possible metric for gauging recruiter and human resource performance. This is consistent with our interviews: the internal recruiter we interviewed, who also holds a dual human resources role, said that turnover can be an important part of human resource performance evaluation. Since a bonus based on turnover is similar in spirit to a refund contract, it could generate the same sort of misalignment that we discuss in this paper.

More generally, human resource employees are responsible for reducing legal risks arising from people. As Dufrane et al. (2021) put it:

> "There are numerous laws and regulations governing the employment relationship that HR professionals must understand and navigate in order to help ensure their organizations avoid costly fines and other penalties, including the potential harm to the organization's reputation."

Behaviors from employees that can give rise to "costly fines and penalties" include serious crimes committed by employees such as sexual harassment and fraud. Thus, human resource departments internalize the downside risks of hiring a malicious or negligent employee, but do not internalize the upside benefits of hiring a star.

These arguments highlight the possibility that our results extend to human resource employees. This greatly broadens the impact of our analysis, because while only some firms use recruiters, a much larger share have some form of human resource department. As of 2018, it is estimated that there are 671,140 human resource workers (Labor Statistics 2019). In 2018, the U.S. Census Bureau estimated that there were 664,757 businesses with 20 or more employees (Bureau 2021). Assuming that most businesses with less than 20 employees do not hire a human resource specialist, this therefore implies that there is more than one human resource worker for every business.

2.8.2 Heterogeneity in Productivity Variance

Many of our results examine how heterogeneity in the variance dimension of beliefs impacts search behavior. We have also shown that delegation results in differential treatment of two workers with the same productivity expectation but different productivity variance. These results beg the question: what generates differences in productivity variance across workers?

There are several sources worth discussing. First, job experience can generate variation in information quality, and thus generate differences in productivity variance for people of different ages. Many theoretical papers, starting with Jovanovic (1979), are based on the idea that work experience provides information about productivity. This idea is supported by empirical work. Fredriksson, Hensvik, and Skans (2018) show that match quality appears to be better among older workers. As a result, age and work experience can generate differences in productivity variance.

Continuing along this same vein, credit constraints can make it such that differences in parental income generate heterogeneity in productivity variance. High quality signals of productivity are expensive. For example, the cost of data science boot camps can be around \$2,000-\$17,000 for just a small period of instruction (Williams 2020). Prestigious universities are usually either extremely expensive (a year's tuition can be in excess of U.S. median annual earnings) or extremely selective. Even with financial aid, individuals from disadvantaged backgrounds often do not have the resources to invest in the preparatory work needed to be admitted.[17]

Unequal access to information can also contribute to inequality of opportunity. For example, currently, only 71% of eligible college applicants file the Free Application for Federal Student Aid (*How America Pays for College 2020*). As a result, job seekers will often need to pay for productivity signals using family support. Children of wealthy parents will tend to have lower productivity variance, meaning that if we compared two workers with the same expected ability but different parental wealth, we would expect the child of wealthier parents to be approached more by recruiters, even if recruiters have no intrinsic bias towards wealthy workers. This will tend to reinforce existing socioeconomic inequality.

Finally, a recruiter might be better able to interpret the resume or life experience of a worker from the same socioeconomic group. Factors like religion, nationality, language, and cultural background play large roles in processing signals of productivity. For example, Bencharit et al. (2019) find that 86 percent of European Americans want to convey

17. SAT preparation classes, tutoring, college admissions counseling, AP testing, etc.

excitement in a job interview, compared to 72 percent of Asian Americans and 48 percent of Hong Kong Chinese. In the same study, it was found European Americans rated their ideal job candidate as excited, while Hong Kong Chinese rated their ideal candidate as calm. This reflects a form of homophily which, in the language of our model, would make a recruiter of European descent have a higher productivity variance about a candidate who was of Asian descent than a candidate of European descent.

Taken together, these examples and our results suggest a form of two-way causality. On the one hand, our application section suggests that delegation is *more likely* in industries that are socioeconomically homogeneous. On the other hand, Proposition 9 and our other results suggest that delegation *helps homogenize* industries.

2.9 Conclusion

We develop a theoretical model of delegated recruitment with uncertain productivity, and show that the general, nonparametric version of the model can be reduced to classic sequential search over modified objects. Our characterization reveals insights into how contract restrictions and delegation translate into moral hazard with a multitasking flavor: a risk neutral recruiter will over-select low-variance candidates at the expense of high-expectation candidates. Using this insight, we prove that the efficiency of delegation critically relies on the level of heterogeneity in the population with respect to productivity variance.

Our framework provides mechanisms for why firms outsource recruitment, why firms might statistically discriminate in the absence of bias, and why occupations can become homogeneous over time. Additionally, we argue that our results extend more broadly to many agents who are involved in hiring processes. We also discuss how socioeconomic inequality can be a source of productivity variance heterogeneity.

We believe that our approach is generally useful for analyzing economic situations in which search over objects of uncertain quality is delegated to a third party (recruiters, mortgage brokers, venture capitalists). The refund contracts is common in many settings;

however we want to emphasize that our general approach is valid generally. Any type of contract shapes a recruiter's indifference curves over the distribution of posteriors. Then, solving the delegated search problem is equivalent to solving a standard sequential search problem for the distribution of expected productivity conditional on those indifference curves. Moreover, comparing the results to the first-best benchmark completely relies on the comparison of this conditional expectation distribution and unconditional expected productivity distribution. In other words, our procedure can be used for more complicated contracts and families of posterior distributions.

One limitation of our analysis also presents an opportunity for further work. Although we have a theory of delegated recruitment, it is unclear whether our theory and the testable implications we derive are consistent with recruiter and firm behavior. This is mainly because data on recruiter behavior is scant, and when it does exist, it is scattered across many platforms. Further empirical work will require a combination of innovative data sources, such as text analysis of recruiter LinkedIn messages and careful field experiments.[18] A particular challenge when evaluating our theory (and indeed, any theory of recruiter behavior) is mapping observed characteristics of workers to beliefs. Despite these challenges, we believe that this research frontier is worth pushing forward. Delegation is now a common practice across many industries, and our paper offers theoretical evidence that it can generate distortions in the hiring process. Verifying the existence and the magnitude of these distortions will require taking ideas to data.

18. Field experiments that elicit beliefs will require new techniques, like the two-sided audit. For an excellent example see Cowgill and Perkowski (2020).

Bibliography

Adenbaum, Jacob. 2021. “Endogenous Firm Structure and Worker Specialization.”

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

Aravindakshan, Ashwin, and Brian Ratchford. 2011. “Solving share equations in logit models using the LambertW function.” *Review of Marketing Science* 9.

Argyres, Nicholas S., Teppo Felin, Nicolai Foss, and Todd Zenger. 2012. “Organizational Economics of Capability and Heterogeneity.” *https://doi.org/10.1287/orsc.1120.0746* 23, no. 5 (October): 1213–1226.

Armstrong, Mark, and John Vickers. 2010. “A model of delegated project choice.” *Econometrica* 78 (1): 213–244.

Ashton, Adam, and Phillip Reese. *Soaring overtime fattens paychecks of California cops and firefighters. But at a cost.*

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. “The Skill Content of Recent Technological Change: An Empirical Exploration.” *The Quarterly Journal of Economics* 118, no. 4 (November): 1279–1333.

Baker, George, Robert Gibbons, and Kevin J. Murphy. 2002. “Relational Contracts and the Theory of the Firm.” *The Quarterly Journal of Economics* 117, no. 1 (February): 39–84.

Baker, George P., and Thomas N. Hubbard. 2003. “Make Versus Buy in Trucking: Asset Ownership, Job Design, and Information.” *American Economic Review* 93, no. 3 (June): 551–572.

Bargain, Olivier, Kristian Orsini, and Andreas Peichl. 2014. “Comparing labor supply elasticities in europe and the united states new results.” *Journal of Human Resources* 49 (3): 723–838.

Becker, Gary S., and Kevin M. Murphy. 1992. "The division of labor, coordination costs, and knowledge." *Quarterly Journal of Economics* 107 (4): 1137–1160.

Belzunce, Felix, Carolina Martinez-Riquelme, Jose M Ruiz, and Miguel A Sordo. 2016. "On sufficient conditions for the comparison in the excess wealth order and spacings." *Journal of Applied Probability* 53 (1): 33–46.

Bencharit, Lucy Zhang, Yuen Wan Ho, Helene H Fung, Dannii Y Yeung, Nicole M Stephens, Rainer Romero-Canyas, and Jeanne L Tsai. 2019. "Should job applicants be excited or calm? The role of culture and ideal affect in employment settings." *Emotion* 19 (3): 377.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841–890.

Black, Ines, Sharique Hasan, and Rembrand Koning. 2020. "Hunting for talent: Firm-driven labor market search in America." *Available at SSRN.*

Blackwell, David. 1953. "Equivalent comparisons of experiments." *The annals of mathematical statistics,* 265–272.

Blahut, Richard E. 1972. "Computation of Channel Capacity and Rate-Distortion Functions." *IEEE Transactions on Information Theory,* no. 4.

Bloedel, Alexander, and Weijie Zhong. 2021. "The Cost of Optimally Acquired Information*." June.

Blundell, Richard, Antoine Bozio, and Guy Laroque. 2011. "Labor supply and the extensive margin." *American Economic Review* 101 (3): 482–486.

Bureau, US Census. 2021. *2018 SUSB Annual Data Tables by Establishment Industry,* May.

Caliendo, Lorenzo, Giordano Mion, Luca David Opromolla, and Esteban Rossi-Hansberg. 2020. “Productivity and organization in portuguese firms.” *Journal of Political Economy* 128, no. 11 (November): 4211–4257.

Caliendo, Lorenzo, Ferdinando Monte, and Esteban Rossi-Hansberg. 2015. “The Anatomy of French Production Hierarchies.” *The Journal of Political Economy* 123, no. 4 (July).

Caliendo, Lorenzo, Esteban Rossi-Hansberg, Lorenzo Caliendo, and Esteban Rossi-Hansberg. 2012. “The Impact of Trade on Organization and Productivity.” *The Quarterly Journal of Economics* 127, no. 3 (August): 1393–1467.

Caplin, Andrew, and Barry Nalebuff. 1991. “Aggregation and Imperfect Competition: On the Existence of Equilibrium.” *Econometrica* 59, no. 1 (January): 25.

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. 2018. “Firms and labor market inequality: Evidence and some theory.” *Journal of Labor Economics* 36, no. S1 (January): S13–S70.

Carlsson, Magnus, Luca Fumarco, and Dan-Olof Rooth. 2014. “Does the design of correspondence studies influence the measurement of discrimination?” *IZA Journal of Migration* 3 (1): 1–17.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. “The Effect of Minimum Wages on Low-Wage Jobs.” *The Quarterly Journal of Economics* 134, no. 3 (August): 1405–1454.

Charles, Kerwin K, Matthew S Johnson, Melvin Stephens Jr, and Do Q Lee. 2019. “Demand conditions and worker safety: Evidence from price shocks in mining.”

Chateauneuf, Alain, Michele Cohen, and Isaac Meilijson. 2004. “Four notions of mean-preserving increase in risk, risk attitudes and applications to the rank-dependent expected utility model.” *Journal of Mathematical Economics* 40 (5): 547–571.

Chausse, Pierre. 2021. *Package 'gmm'.*

Che, Yeon-Koo, Wouter Dessein, and Navin Kartik. 2013. “Pandering to persuade.” *American Economic Review* 103 (1): 47–79.

Chen, Po-Ning. n.d. “Chapter 6 Lossy Data Compression and Transmission.”

Chen, Songnian, Yahong Zhou, and Yuanyuan Ji. 2018. “Nonparametric identification and estimation of sample selection models under symmetry.” *Journal of Econometrics* 202 (2): 148–160.

Chetty, Raj. 2012. “Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply.” *Econometrica* 80 (3): 969–1018.

Conway, Sadie H, Lisa A Pompeii, David de Porras, Jack L Follis, and Robert E Roberts. 2017. “The identification of a threshold of long work hours for predicting elevated risks of adverse health outcomes.” *American Journal of Epidemiology* 186 (2): 173–183.

Cowgill, Bo, and Patryk Perkowski. 2020. “Delegation in Hiring: Evidence from a Two-Sided Audit.” *Columbia Business School Research Paper,* no. 898.

Crawford, Gregory S, Oleksandr Shcherbakov, and Matthew Shum. 1997. “Quality Over-provision in Cable Television Markets †.” *American Economic Review* 2019 (3).

Csaba, Dániel. 2021. “Attention Elasticities and Invariant Information Costs *.”

d’Aspremont, C., J. Jaskold Gabszewicz, and J.-F. Thisse. 1979. “On Hotelling’s ”Stability in Competition”.” *Econometrica* 47, no. 5 (September): 1145.

Dai, Tianjiao, and Juuso Toikka. 2022. “Robust Incentives for Teams.” *Econometrica* 90, no. 4 (July): 1583–1613.

Dembe, Allard E, J Bianca Erickson, Rachel G Delbos, and Steven M Banks. 2005. “The impact of overtime and long work hours on occupational injuries and illnesses: new evidence from the United States.” *Occupational and environmental medicine* 62 (9): 588–597.

Denti, Tommaso, Massimo Marinacci, and Luigi Montrucchio. 2020. "A note on rational inattention and rate distortion theory." *Decisions in Economics and Finance* 43, no. 1 (June): 75–89.

Dessein, Wouter, and Tano Santos. 2006. "Adaptive organizations." *Journal of Political Economy* 114, no. 5 (October): 956–995.

Du, Yu, and Ravi Varadhan. 2020. "SQUAREM: An R package for off-the-shelf acceleration of EM, MM and other EM-like monotone algorithms." *Journal of Statistical Software* 92:1–41.

Dufrane, Amy, Michael Bonarti, Jim DeLoach, and Rebecca DeCook. 2021. *The HR Function's Compliance Role,* June.

Ellickson, Paul B. 2007. "Does Sutton apply to supermarkets?" *The RAND Journal of Economics* 38, no. 1 (March): 43–59.

Frankel, Alex. 2021. "Selecting Applicants." *Econometrica* 89 (2): 615–645.

Frankel, Alexander. 2014. "Aligned delegation." *American Economic Review* 104 (1): 66–83.

Fredriksson, Peter, Lena Hensvik, and Oskar Nordström Skans. 2018. "Mismatch of talent: Evidence on match quality, entry wages, and job mobility." *American Economic Review* 108 (11): 3303–3338.

Freund, Lukas. 2022. "Superstar Teams: The Micro Origins and Macro Implications of Coworker Complementarities." *SSRN Electronic Journal* (December).

Galperin, Ron. 2015. *DOT Traffic Control for Special Events.*

Garicano, L. 2000. "Hierarchies and the organization of knowledge in production." *Journal of Political Economy* 108 (5): 874–904.

Garicano, Luis, and Thomas N. Hubbard. 2016. "The Returns to Knowledge Hierarchies." *The Journal of Law, Economics, and Organization* 32, no. 4 (November): 653–684.

Garicano, Luis, and Esteban Rossi-Hansberg. 2006. "Organization and Inequality in a Knowledge Economy."

Garicano, Luis, and Yanhui Wu. 2012. "Knowledge, communication, and organizational capabilities." *Organization Science* 23, no. 5 (February): 1382–1397.

Gentzkow, Matthew, and Emir Kamenica. 2016. "A Rothschild-Stiglitz approach to Bayesian persuasion." *American Economic Review* 106 (5): 597–601.

Gibbons, Robert. 2020. "March-ing toward organizational economics." *Industrial and Corporate Change* 29, no. 1 (February): 89–94.

Gregory, Terry, and Ulrich Zierahn. 2022. "When the minimum wage really bites hard: The negative spillover effect on high-skilled workers." *Journal of Public Economics* 206 (February): 104582.

Guardado, José R, and Nicolas R Ziebarth. 2019. "Worker investments in safety, workplace accidents, and compensating wage differentials." *International Economic Review* 60 (1): 133–155.

Haanwinckel, Daniel. 2020. "Supply, Demand, Institutions, and Firms: A Theory of Labor Market Sorting and the Wage Distribution."

Heart Disease Facts. 2020.

Heckman, James J. 1998. "Detecting discrimination." *Journal of economic perspectives* 12 (2): 101–116.

Heckman, James J, and Edward J Vytlacil. 1999. "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the national Academy of Sciences* 96 (8): 4730–4734.

Hirshman, Carolyn. 2018. *Incentives for Recruiters?,* April.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." *JL Econ. \& Org.* 7:24.

———. 1994. "The Firm as an Incentive System." *American Economic Review* 84 (4): 972–991.

Horrace, William C, and Ronald L Oaxaca. 2006. "Results on the bias and inconsistency of ordinary least squares for the linear probability model." *Economics Letters* 90 (3): 321–327.

How America Pays for College 2020.

Incidence rates for nonfatal occupational injuries and illnesses. 2020.

Johnson, Matthew S., and Michael Lipsitz. 2022. "Why Are Low-Wage Workers Signing Noncompete Agreements?" *Journal of Human Resources* 57, no. 3 (May): 689–724.

Jovanovic, Boyan. 1979. "Job matching and the theory of turnover." *Journal of political economy* 87 (5, Part 1): 972–990.

Jung, Junehyuk, Jeong Ho (John) Kim, Filip Matějka, and Christopher A. Sims. 2019. "Discrete Actions in Information-Constrained Decision Problems." *The Review of Economic Studies* 86, no. 6 (November): 2643–2667.

Kim, Woorim, Eun-Cheol Park, Tae-Hoon Lee, and Tae Hyun Kim. 2016. "Effect of working hours and precarious employment on depressive symptoms in South Korean employees: a longitudinal study." *Occupational and environmental medicine* 73 (12): 816–822.

Kniesner, Thomas J, and W Kip Viscusi. 2019. "The value of a statistical life." *Forthcoming, Oxford Research Encyclopedia of Economics and Finance,* 15–19.

Kremer, Michael. 1993. "The o-ring theory of economic development." *Quarterly Journal of Economics* 108 (3): 551–575.

Kugler, Maurice, and Eric Verhoogen. 2012. "Prices, Plant Size, and Product Quality." *The Review of Economic Studies* 79, no. 1 (January): 307–339.

Kuhn, Andreas, and Oliver Ruf. 2013. "The value of a statistical injury: New evidence from the Swiss labor market." *Swiss Journal of Economics and Statistics* 149 (1): 57–86.

Kuhn, Moritz, Jinfeng Luo, Iourii Manovskii, and Xincheng Qiu. 2022. "Coordinated Firm-Level Work Processes and Macroeconomic Resilience *."

Labor Statistics, U.S. Bureau of. 2019. *May 2018 National Occupational Employment and Wage Estimates,* April.

Lavetti, Kurt. 2020. "The estimation of compensating wage differentials: Lessons from the Deadliest Catch." *Journal of Business & Economic Statistics* 38 (1): 165–182.

Lazear, Edward P. 2009. "Firm-specific human capital: A skill-weights approach." *Journal of Political Economy* 117, no. 5 (October): 914–940.

Lee, David L, Justin McCrary, Marcelo J Moreira, and Jack Porter. 2020. "Valid t-ratio Inference for IV." *arXiv preprint arXiv:2010.05058.*

Lee, Jonathan M, and Laura O Taylor. 2019. "Randomized safety inspections and risk exposure on the job: Quasi-experimental estimates of the value of a statistical life." *American Economic Journal: Economic Policy* 11 (4): 350–74.

Li, Fei, and Can Tian. 2013. "Directed search and job rotation." *Journal of Economic Theory* 148, no. 3 (May): 1268–1281.

Liebman, Jeffrey B, Erzo F P Luttmer, and David G Seif. 2009. "Labor supply responses to marginal Social Security benefits: Evidence from discontinuities." *Journal of Public Economics* 93 (11-12): 1208–1223.

Lindenlaub, Ilse. 2017. "Sorting Multidimensional Types: Theory and Application." *The Review of Economic Studies* 84, no. 2 (April): 718–789.

Lipnowski, Elliot, and Doron Ravid. 2022. "Predicting Choice from Information Costs" (May).

Lucas, Robert E. 1978. "On the Size Distribution of Business Firms." *The Bell Journal of Economics* 9 (2): 508.

Martinez, Elizabeth A., Nancy Beaulieu, Robert Gibbons, Peter Pronovost, and Thomas Wang. 2015. "Organizational Culture and Performance." *American Economic Review* 105, no. 5 (May): 331–35.

Matêjka, Filip, and Alisdair McKay. 2015. "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." *American Economic Review* 105, no. 1 (January): 272–98.

McCall, John Joseph. 1970. "Economics of information and job search." *The Quarterly Journal of Economics,* 113–126.

McFadden, D. 1973. "Conditional logit analysis of qualitative choice behavior."

Meier, Stephan, Matthew Stephenson, and Patryk Perkowski. 2019. "Culture of trust and division of labor in nonhierarchical teams." *Strategic Management Journal* 40, no. 8 (August): 1171–1193.

Moran, Molly J, and Carlos Monje. 2016. "Guidance on treatment of the economic value of a statistical life (vsl) in us department of transportation analyses-2016 adjustment." *US Department of Transportation.*

Muro, Mark, Alan Berube, and Jacob Whiton. 2018. *Black and Hispanic underrepresentation in tech: It's time to change the equation,* April.

New York City Department of Finance. 2022. *New York State Sales and Use Tax,* May.

Nocke, Volker, and Nicolas Schutz. 2018. "Multiproduct-Firm Oligopoly: An Aggregative Games Approach." *Econometrica* 86, no. 2 (March): 523–557.

Ocampo, Sergio. 2022. "A Task-Based Theory of Occupations with Multidimensional Heterogeneity *."

Parada-Contzen, Marcela, Andrés Riquelme-Won, and Felipe Vasquez-Lavin. 2013. "The value of a statistical life in Chile." *Empirical Economics* 45 (3): 1073–1087.

Pomatto, Luciano, Philipp Strack, and Omer Tamuz. 2022. "The Cost of Information."

Postel–Vinay, Fabien, and Jean–Marc Robin. 2002. "Equilibrium wage dispersion with worker and employer heterogeneity." *Econometrica* 70 (6): 2295–2350.

Rosen, Sherwin. 1982. "Authority, Control, and the Distribution of Earnings." *Bell Journal of Economics* 13 (2): 311–323.

Sattinger, Michael. 1975. "Comparative Advantage and the Distributions of Earnings and Abilities." *Econometrica* 43, no. 3 (May): 455.

Semykina, Anastasia, and Jeffrey M Wooldridge. 2018. "Binary response panel data models with sample selection and self-selection." *Journal of Applied Econometrics* 33 (2): 179–197.

Shaked, Avner, and John Sutton. 1987. "Product Differentiation and Industrial Structure." *The Journal of Industrial Economics* 36, no. 2 (December): 131.

Shaked, Moshe, and J George Shanthikumar. 2007. *Stochastic orders.* Springer Science & Business Media.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423.

Shimer, Robert, and Lones Smith. 2000. "Assortative matching and search." *Econometrica* 68 (2): 343–369.

Shkel, Yanina, and Sergio Verdú. 2018. "A Coding Theorem for f-Separable Distortion Measures." *Entropy* 20, no. 2 (February).

Song, Jae, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter. 2019. "Firming Up Inequality." *The Quarterly Journal of Economics* 134, no. 1 (February): 1–50.

Steinbach, Alison. 2019. *Police, firefighters earn most overtime among public employees, at times doubling salaries.*

Stock, James H, and Motohiro Yogo. 2002. *Testing for weak instruments in linear IV regression.* Technical report. National Bureau of Economic Research.

Teulings, Coen N. 2000. "Aggregation Bias in Elasticities of Substitution and the Minimum Wage Paradox." *International Economic Review* 41, no. 2 (May): 359–398.

Tian, Jianrong. 2019. "Attention Costs without Distraction and Entropy."

Tirole, Jean. 2009. "Cognition and Incomplete Contracts." *American Economic Review* 99, no. 1 (March): 265–94.

Tishby, Naftali, Fernando C Pereira, and William Bialek. 2000. "The information bottleneck method."

Ulbricht, Robert. 2016. "Optimal delegated search with adverse selection and moral hazard." *Theoretical Economics* 11 (1): 253–278.

Viscusi, W Kip, and Joseph E Aldy. 2003. "The value of a statistical life: a critical review of market estimates throughout the world." *Journal of risk and uncertainty* 27 (1): 5–76.

Viscusi, W Kip, and Joni Hersch. 2001. "Cigarette smokers as job risk takers." *Review of Economics and Statistics* 83 (2): 269–280.

Williams, Alex. 2020. *Our Ultimate Guide to the Best Data Science Bootcamps.*

Williamson, Ann, Rena Friswell, Jake Olivier, and Raphael Grzebieta. 2014. "Are drivers aware of sleepiness and increasing crash risk while driving?" *Accident Analysis & Prevention* 70:225–234.

Williamson, Oliver. 1984. "The Incentive Limits of Firms: A Comparative Institutional Assessment of Bureaucracy," 736–763.

Zhou, Jidong. 2020. "Improved Information in Search Markets."

Zhou, Xiang, and Yu Xie. 2019. "Marginal treatment effects from a propensity score perspective." *Journal of Political Economy* 127 (6): 3070–3084.

2.10 Appendix

2.10.1 Match-Specific Productivity

We have interpreted productivity expectation and variance (μ, σ) as characterizing the beliefs of the recruiter about a worker prior to hire. However, our results extend naturally to models where productivity of worker i at firm j can be decomposed into a worker effect, a firm effect, and a match-specific effect, like so:

$$a_{i,j} = \gamma_i + \kappa_j + \epsilon_{i,j}$$

For simplicity, assume that all three components are independent. It is reasonable to assume that the firm effect, κ_j, is known. Indeed, since a recruiter is searching for the same firm, it is constant across all workers, and thus only shifts the mean of the productivity distribution. We can think of the match-specific effect as being mean 0 with known variance σ_ϵ^2 and unpredictable prior to hire. The expertise of the recruiter lies in their ability to predict γ_i. After searching a worker, reading their resume, and conducting a preliminary interview, the recruiter forms an estimate of expected γ_i, denoted $\bar{\gamma}_i$, with a corresponding variance estimate σ_i.

Assuming normality of all components, the recruiter views $a_{i,j}$ after an interview as normal, with expectation $\kappa_j + \bar{\gamma}_i$ and variance $\sigma_\epsilon^2 + \sigma_i^2$. These two components correspond to our productivity expectation and variance, (μ, σ). If the recruiter can perfectly predict

γ_i, $\sigma_i^2 = 0$ for all i and the productivity variance distribution is degenerate. First-best is achieved, and there are no distortions from delegation.

Consider the simple case when the recruiter can only distinguish between two groups of workers. The groups have the same expected worker effect $\bar{\gamma}$ but different worker effect variances σ_i. In this situation, our model predicts that the recruiter will always prefer the lower variance group, because this group has a higher standardized productivity. In any incentive compatible contract, the lower-variance group is always hired. However, the firm is indifferent between the groups, and will hire the first worker searched.

2.10.2 Proof of First-Best Search

The proof of the optimal sequential search strategy (without delegation) is well-known, but we include it for completeness. Denote V as the value function of the firm. Denote the marginal distribution of μ as F. The dynamic programming problem of the firm is given by:

$$V = -c + \int \max\{E[a|\mu = u], V\} dF(\mu)$$

Note that if there were recall (so that the highest previously viewed μ could be carried as a state variable), then the firm would never exercise the option.

$$V = -c + \int \max\{\mu, V\} dF(\mu)$$

Re-writing, this yields:

$$0 = -c + \int \max\{\mu - V, 0\} dF(\mu)$$

So, the optimal strategy is a reservation rule characterized by μ^*, where $V = \mu^*$. Thus:

$$c = \int \max\{\mu - V, 0\} dF(\mu) \leftrightarrow c = \int_{\mu > \mu^*} \mu - \mu^* dF(\mu)$$

Integration by parts gives:

$$c = -[(1 - F(\mu))(\mu - \mu^*)]_{\mu^*}^{\bar{\mu}} + \int_{\mu^*}^{\bar{\mu}} (1 - F(\mu))d\mu$$

Since the first term is 0, this simplifies to:

$$c = \int_{\mu^*}^{\bar{\mu}} (1 - F(\mu))d\mu$$

As an aside, note that we can re-arrange the intermediate equation this way:

$$c = \int_{\mu > \mu^*} \mu dF(\mu) - (1 - F(\mu^*))\mu^* \leftrightarrow \mu^* = \frac{1}{1 - F(\mu^*)} \left(\int_{\mu > \mu^*} \mu dF(\mu) - c \right)$$

which can be compactly re-written as:

$$\mu^* = E[\mu | \mu \geq \mu^*] - \frac{c}{Pr(\mu \geq \mu^*)}$$

2.10.3 Proof of Lemma 3

The dynamic programming problem of the recruiter is given by:

$$U = -c + \int \max\{-\beta E_a[\mathbb{I}\{a \leq 0\} | (u, s)], U\} dG(\mu, \sigma)$$

$$U = -c + \int \max\{-\beta F(-\mu/\sigma), U\} dG(\mu, \sigma)$$

Re-writing, this yields:

$$0 = -c + \int \max\{\beta F(\mu/\sigma) - U, 0\} dG(\mu, \sigma)$$

Observe that utility only depends on μ/σ, so we can reduce the problem to one-dimensional search. As long as β is negative, utility will be increasing in μ/σ. The firm will always set $\beta \leq 0$ when $E[a|\tilde{\mu} = x]$ is increasing in x (which is what we assumed). Thus, we will have a reservation rule strategy in the ratio μ/σ. Denote this reservation rule x^*. Returning to the recruiter's problem, we can re-write using standardized productivity ($\tilde{\mu}$):

$$c = \int \max\{\beta F(\tilde{\mu}), U\} d\tilde{G}(\tilde{\mu}) \leftrightarrow c = \int_{\tilde{\mu} \geq x^*} \beta F(\tilde{\mu}) - \beta F(x^*) d\tilde{G}(\tilde{\mu})$$

2.10.4 Proof of Theorem 2

The next lemma is implicitly used in the Proof of Theorem 2 while showing that the search over $\tilde{\mu}$ is equivalent to search over $\mathbb{E}[\mu|\tilde{\mu}]$, regardless of whether or not the first one contains more information about the worker.

Lemma 5 ***No-atom optimal search.*** *Let one search over a pool of uniformly distributed x with a payoff $f(x)$, ($f'(x) \geq 0$)) and a cost $c > 0$ per search. Let $x^* \in (0, 1)$ be a unique optimal search threshold. Then, $\forall \varepsilon > 0 : f(x^* - \varepsilon) < f(x^* + \varepsilon)$.*

Proof. The intuition of the statement is that one being able to set a threshold on the CDF of the search variable (rather than the variable itself) would never strictly prefer to set it within an atom than anywhere else. The problem described in the lemma can be stated as

$$\max_{x'}\{\mathbb{E}[f(x)|x \geq x'] - \frac{c}{1 - x'}\}$$

The derivative with respect to x' is

$$(\mathbb{E}[f(x)|x \geq x'] - f(x')) * (1 - x') - c = (*)$$

Let us suppose that x^* is the unique maximizer, and that $\exists\ \varepsilon > 0 :$ s.t. $f(x)$ is flat on $(x^* - \varepsilon; x^* + \varepsilon)$. Let $\bar{x} = x^* + \varepsilon$. Locally for $x' \in (x^* - \varepsilon; x^* + \varepsilon)$

$$\mathbb{E}[f(x)|x \geq x'] = \frac{(1 - \bar{x}) * \mathbb{E}[f(x)|x \geq \bar{x}] + (\bar{x} - x') * f(x^*)}{1 - x'}$$

Then, simplifying the derivative of the outcome with respect to x' gives

$$(*) = (1-\bar{x}) * \mathbb{E}[f(x)|x \geq \bar{x}] + (\bar{x} - x') * f(x^*) - f(x^*) * (1-x') - c$$

$$= (1-\bar{x}) * (\mathbb{E}[f(x)|x \geq \bar{x}] - f(x^*)) - c$$

which apparently does not depend on x' and is constant for $x' \in (x^* - \varepsilon; x^* + \varepsilon)$. Then, x^* cannot be a unique maximizer, since – depending on the sign of the derivative – one should either increase or decrease the threshold, or is indifferent in some small neighborhood around x^*. ■

We apply Theorem 2 to the firm's problem, which is given by Equations OBJ, IR, IC, and VAL:

$$\max_{\alpha,\beta,\mathcal{D}_R} E[a - \beta\mathbb{I}\{a > 0\}|(\mu,\sigma) \in D_R] - \alpha$$

s.t.

$$\alpha + u^* \geq 0 \tag{IR}$$

$$c = \int_{u \geq u^*} (1 - M(u))du \tag{IC}$$

$$\mathcal{D}_R = \{\mu, \sigma|\mu/\sigma \geq F^{-1}\left(\frac{u^*}{\beta}\right)\} \tag{REGION}$$

First, we prove that the IR constraint must bind. Suppose it does not. Then, the firm could lower α by ϵ and increase maximized profit without violating any other constraints. This contradicts optimality. Thus, IR binds at the optimum. From the end of the proof of Lemma 3, we have that:

$$u^* = E[u|u \geq u^*] - \frac{c}{Pr(u \geq u^*)}$$

Plugging this into binding IR and solving for α:

$$\alpha = -E[u|u \geq u^*] + \frac{c}{Pr(u \geq u^*)}$$

Substituting the result into the objective obtains:

$$\max_{\beta, \mathcal{D}_R} E[a|(\mu, \sigma) \in D_R] - \frac{c}{Pr((\mu, \sigma) \in D_R}$$

which is the desired form of the objective. Using Lemma 2, the modified problem becomes:

$$\max_{\beta, u^*} E[a|\mu/\sigma \geq F^{-1}(u^*/\beta)] - \frac{c}{Pr(\mu/\sigma \geq F^{-1}(u^*/\beta))}$$

$$c = \int_{u \geq u^*} (1 - M(u))du \tag{IC}$$

This makes apparent the fact that the objective is no longer constrained by the constraints (since we have an extra degree of freedom), and in fact only depends on $x := F^{-1}(u^*/\beta)$.

The firm's choice of the contract creates the incentives over $\tilde{\mu}$ in the recruiter's optimal stopping problem. This, along with the binding IR constraint in the delegated problem, means that the firm implicitly searches over $\tilde{\mu}$. Given the monotonicity Assumption 1 and Lemma 3, this is equivalent to searching over $\mathbb{E}[\mu|\tilde{\mu}]$, which is the firm's outcome. This optimal search is characterized by the first-order condition stated in the theorem.

Therefore, we can maximize the objective without constraints to derive x, then use the definition of x and the IC constraint to derive β, u^*. Finally, α can be retrieved from the binding IR constraint, and thus, the problem reduces in the way stated in the proposition.■

2.10.5 Proof of Proposition 9

Proof. Note that under independence, $\sigma|\mathcal{D}_F$ is the same as the unconditional distribution of σ. Then:

$$\begin{aligned}
Pr(\sigma \leq y|(\mu,\sigma) \in \mathcal{D}_R) &= Pr(\mu \leq y\tilde{\mu}^*|(\mu,\sigma) \in \mathcal{D}_R) * Pr(\sigma \leq y|\mu \leq y\tilde{\mu}^* \; \& \; (\mu,\sigma) \in \mathcal{D}_R) \\
&\quad + Pr(\mu > y\tilde{\mu}^*|(\mu,\sigma) \in \mathcal{D}_R) * Pr(\sigma \leq y|\mu > y\tilde{\mu}^* \; \& \; (\mu,\sigma) \in \mathcal{D}_R) \\
&= Pr(\mu \leq y\tilde{\mu}^*|(\mu,\sigma) \in \mathcal{D}_R) * 1 + (1 - Pr(\mu \leq y\tilde{\mu}^*|(\mu,\sigma) \in \mathcal{D}_R)) * G_\sigma(y) \\
&> G_\sigma(y) = Pr(\sigma \leq y|(\mu,\sigma) \in \mathcal{D}_F)
\end{aligned}$$

Notice that the first quantity is the conditional CDF in the recruiter acceptance region. The second-to-last line shows that the this CDF is essentially a weighted average of 1 and $G_\sigma(y)$, which is always weakly greater than $G_\sigma(y)$. This proves first-order stochastic dominance of σ by $\sigma|\mathcal{D}_F$. ■

2.10.6 Proof of Lemma 4

We will use the notations from Figure 2.2.

$$p = \frac{Pr(A)}{Pr(A) + \Pr(B)}$$

$$q = \frac{Pr(C)}{Pr(C) + \Pr(B)}$$

We can conclude that $p > q$ since search effort is lower in the second-best ($Pr(\text{FB}) < Pr(\text{SB})$).

$$\mu|\text{SB} \sim (\mu|A)p(\mu|B)$$

$$\mu|\text{FB} \sim (\mu|C)q(\mu|B)$$

where notations on the RHS are used for mixture distribution. In other words, one could write each of them as a three-component mixture:

$$\mu|\text{SB} \sim (\mu|A)(\text{w/p } q) + (\mu|A)(\text{w/p } p-q) + (\mu|B)(\text{w/p } 1-p)$$

$$\mu|\text{FB} \sim (\mu|C)(\text{w/p } q) + (\mu|B)(\text{w/p } p-q) + (\mu|B)(\text{w/p } 1-p)$$

Given the support of $\mu|A,\ \mu|B,\ \mu|C$, it is trivial to conclude that $\mu|B \succ_{\text{FOSD}} \mu|A$ and $\mu|C \succ_{\text{FOSD}} \mu|A$. Thus, each of the components in the first-best μ mixture first order stochastically dominates the components in the second-best μ mixture. Given that the mixture probabilities are identical, this implies that the whole FB mixture dominates the SB mixture

$$\mu|\text{FB} \succ_{\text{FOSD}} \mu|\text{SB}$$

(this simply follows from the formula of a mixture CDF).

2.10.7 Proof of Proposition 10

The firm extracts all surplus from the recruiter, so profit and social surplus are the same. Consider any two values of the parameter $\theta' > \theta$. Then $E_{\theta'}[\mu|\tilde{\mu}]$ dominates $E_{\theta}[\mu|\tilde{\mu}]$ in the excess wealth order. Thus we have from Theorem 3 that search effort increases in θ. Excess wealth order dominance implies $E_{\theta'}[\mu|\tilde{\mu}]$ is a mean-preserving spread of $E_{\theta}[\mu|\tilde{\mu}]$. By well-known properties of sequential search, this implies profit is higher under θ' so profit and social surplus are increasing in θ. Finally, because the marginal distribution of μ does not depend on θ first-best profit, social surplus and search effort are fixed as θ rises. When $E_{\theta}[\mu|\tilde{\mu}] \overset{d}{=} \mu$, it must be that first-best values are archieved. Because $E_{\theta}[\mu|\tilde{\mu}]$ converges to μ in distribution as $\theta \to \bar{\theta}$, we then have that profit, social surplus and search effort converge to their first-best values from below.

2.10.8 Pareto Productivity Distribution

With a Pareto productivity distribution we can characterize μ^* this way:

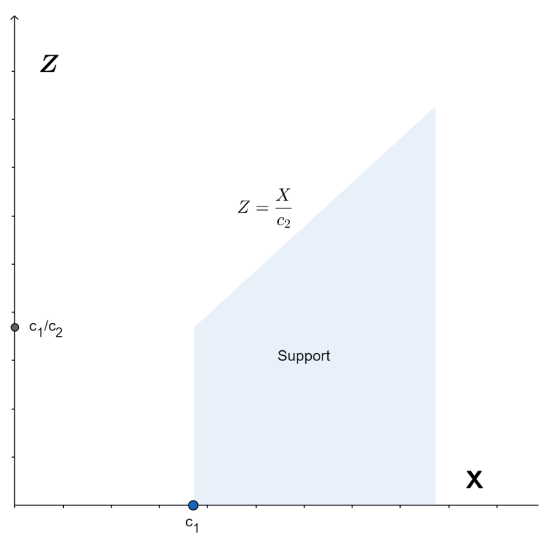
$$c = \int_{\mu^*}^{\infty} 1 - G_\mu(x)dx = \int_{\mu^*}^{\infty} \left(\frac{x_\mu}{x}\right)^{\theta_\mu} dx$$

Integration and solving for μ^* yields the optimal first-best threshold. Note that the first-best solution does not depend on the distribution of σ. We now derive several useful aspects of the joint distribution. Suppose that μ, σ is jointly distributed according to the density g from Assumption 3. We now derive the joint density of $\mu, \tilde{\mu} := \mu/\sigma$, which we denote f. By the transformation theorem, this is given by:

$$f(\mu, \tilde{\mu}) = g(\mu, \mu/\tilde{\mu}) \cdot \frac{\mu}{\tilde{\mu}^2}$$

$$f(\mu, \tilde{\mu}) = \frac{\theta_\mu \theta_\sigma x_\mu^{\theta_\mu} x_\sigma^{\theta_\sigma}}{x^{\theta_\mu+\theta_\sigma+1}} z^{-1+\theta_\sigma} \mathbb{I}\{\mu \geq x_\mu\}\mathbb{I}\{\mu/\tilde{\mu} \geq x_\sigma\}$$

Figure 2.5: Support for (X=μ,Z=$\tilde{\mu}$)



Now, we derive the marginal distribution of $\tilde{\mu}$. Consider first when $z \leq x_\mu/x_\sigma$. Then, the first indicator implies that the second is satisfied, and we can get the marginal:

$$f_{\tilde{\mu}}(\tilde{\mu}) = \int_{x_\mu}^{\infty} g(x,z)dx = \frac{\theta_\mu \theta_\sigma}{(\theta_\sigma + \theta_\mu)} z^{-1+\theta_\sigma} \left(\frac{x_\sigma}{x_\mu}\right)^{\theta_\sigma}$$

In the other case, the second indicator implies the first, so:

$$f_{\tilde{\mu}}(\tilde{\mu}) = \int_{x_\sigma \tilde{\mu}}^{\infty} f(\mu, \tilde{\mu})d\mu = \frac{\theta_\mu \theta_\sigma}{\theta_\mu + \theta_\sigma} z^{-1-\theta_\mu} \left(\frac{x_\mu}{x_\sigma}\right)^{\theta_\mu}$$

Now, we get the marginal CDF by cases:

$$F(\tilde{\mu}) = \begin{cases} \frac{\theta_\mu}{\theta_\mu + \theta_\sigma} \left(\frac{x_\sigma}{x_\mu}\right)^{\theta_\sigma} \tilde{\mu}^{\theta_\sigma} \text{ if } \tilde{\mu} \leq x_\mu / x_\sigma \\ 1 - \frac{\theta_\sigma}{\theta_\mu + \theta_\sigma} \tilde{\mu}^{-\theta_\mu} \left(\frac{x_\mu}{x_\sigma}\right)^{\theta_\mu} \text{ else} \end{cases}$$

The conditional distribution is then:

$$f(\mu|\tilde{\mu}) = \frac{f(\mu, \tilde{\mu})}{f_{\tilde{\mu}}(\tilde{\mu})} = \begin{cases} \frac{x_\mu^{\theta_\mu + \theta_\sigma}(\theta_\mu + \theta_\sigma)}{\mu^{\theta_\mu + \theta_\sigma + 1}} \mathbb{I}\{\mu \geq x_\mu\} \text{ if } \tilde{\mu} \leq x_\mu / x_\sigma \\ \frac{(x_\sigma \tilde{\mu})^{\theta_\mu + \theta_\sigma}(\theta_\mu + \theta_\sigma)}{\mu^{\theta_\mu + \theta_\sigma + 1}} \mathbb{I}\{\mu \geq x_\sigma \tilde{\mu}\} \text{ else} \end{cases}$$

$$E[\mu|\tilde{\mu} = z] = \begin{cases} \frac{(\theta_\mu + \theta_\sigma)}{(\theta_\mu + \theta_\sigma - 1)} x_\mu \text{ if } z \leq x_\mu / x_\sigma \\ \frac{(\theta_\mu + \theta_\sigma)}{(\theta_\mu + \theta_\sigma - 1)} x_\sigma z \text{ else} \end{cases}$$

For $z > x_\mu / x_\sigma$

$$\mathbb{E}[X|Z > z] = \frac{\theta_\mu + \theta_\sigma}{\theta_\mu + \theta_\sigma - 1} \cdot \frac{\theta_\mu}{\theta_\mu - 1} x_\sigma z$$

$$\mathbb{E}[X|Z > z] - \mathbb{E}[X|Z = z] = \frac{\theta_\mu + \theta_\sigma}{\theta_\mu + \theta_\sigma - 1} \cdot \frac{1}{\theta_\mu - 1} x_\sigma z$$

Thus, the First Order Condition determining SB search threshold z^* – $([\mu|\tilde{\mu} > z^*] - [\mu|\tilde{\mu} = z^*]) * \Pr(\tilde{\mu} > z^*) = c$ – for independently Pareto distributed μ and σ with parameters (x_μ, θ_μ) and (x_μ, θ_μ) can be re-written as

$$\frac{\theta_\mu + \theta_\sigma}{\theta_\mu + \theta_\sigma - 1} \cdot \frac{1}{\theta_\mu - 1} x_\sigma z^* \cdot \frac{\theta_\sigma}{\theta_\mu + \theta_\sigma} {z^*}^{-\theta_\mu} \left(\frac{x_\mu}{x_\sigma}\right)^{\theta_\mu} = c$$

Figure 2.6: Conditional Expectation Function

or

$$\frac{\theta_\sigma}{(\theta_\mu+\theta_\sigma-1)(\theta_\mu-1)}\cdot\frac{x_\mu^{\theta_\mu}}{x_\sigma^{\theta_\mu-1}}\cdot\frac{1}{c}={z^*}^{\theta_\mu-1}$$

$$z^*=\left(\frac{\theta_\sigma}{(\theta_\mu+\theta_\sigma-1)(\theta_\mu-1)}\right)^{\dfrac{1}{\theta_\mu-1}}\cdot\frac{(x_\mu^{\theta_\mu}/c)^{\dfrac{1}{\theta_\mu-1}}}{x_\sigma}$$

Re-arrange:

$$z^*=\left(\frac{x_\mu^{\theta_\mu}\theta_\sigma}{c(\theta_\mu+\theta_\sigma-1)(\theta_\mu-1)}\right)^{\dfrac{1}{\theta_\mu-1}}\cdot\frac{1}{x_\sigma}$$

is increasing in x_μ, θ_σ and decreasing in x_σ, c (and probably increasing in θ_μ - not clear). Note that the firm will select $z \geq x_\mu/x_\sigma$ if, and only if:

$$\frac{x_\mu\theta_\sigma}{(\theta_\mu+\theta_\sigma-1)(\theta_\mu-1)}\geq c$$

That is, so long as costs are not too large. If we plug in $z = x_\mu/x_\sigma$ (the knife-edge case), we have that:

$$\frac{\theta_\mu+\theta_\sigma}{\theta_\mu+\theta_\sigma-1}\cdot\frac{1}{\theta_\mu-1}\cdot\frac{\theta_\sigma}{\theta_\mu+\theta_\sigma}<\frac{c}{\theta_\mu x_\mu/(\theta_\mu-1)}*\frac{\theta_\mu}{\theta_\mu-1}$$

$$\frac{\theta_\sigma}{(\theta_\mu + \theta_\sigma - 1)\theta_\mu} < \frac{c}{\mathbb{E}[\mu]} < 1$$

It is possible that this is satisfied. We restrict attention to when it is not, which generates the assumption:

$$\frac{\theta_\sigma}{(\theta_\mu + \theta_\sigma - 1)\theta_\mu} < \frac{c}{\mathbb{E}[\mu]} \leftrightarrow \frac{\theta_\sigma x_\mu}{(\theta_\mu + \theta_\sigma - 1)(\theta_\mu - 1)} \geq c$$

For completeness, we consider the other FOC (when $z < x_\mu/x_\sigma$):

$$((p * \mathbb{E}[\mu|\tilde{\mu} = x_\mu/x_\sigma] + (1-p) * \mathbb{E}[\mu|\tilde{\mu} > x_\mu/x_\sigma]) - \mathbb{E}[\mu|\tilde{\mu} = x_\mu/x_\sigma]) * Pr(\tilde{\mu} > z^*) = c$$

Where $p = Pr(\tilde{\mu} < x_\mu/x_\sigma|\tilde{\mu} > z^*) \;\Rightarrow\; (1-p) * Pr(\tilde{\mu} < x_\mu/x_\sigma|\tilde{\mu} > z^*) = \dfrac{\theta_\sigma}{\theta_\mu + \theta_\sigma}$. Thus, the FOC is equivalent to

$$(1-p)Pr(\tilde{\mu} < x_\mu/x_\sigma|\tilde{\mu} > z^*)(\mathbb{E}[\mu|\tilde{\mu} > \frac{x_\mu}{x_\sigma}] - \mathbb{E}[\mu|\tilde{\mu} = \frac{x_\mu}{x_\sigma}]) = \frac{\theta_\sigma}{\theta_\mu + \theta_\sigma} \frac{\theta_\mu + \theta_\sigma}{\theta_\mu + \theta_\sigma - 1} \frac{x_\mu}{\theta_\mu - 1} = c$$

$$\frac{x_\mu \theta_\sigma}{(\theta_\mu + \theta_\sigma - 1)(\theta_\mu - 1)} = c$$

Thus, the closed form solutions assuming an interior solution are:

$$\tilde{\mu}^* = \frac{1}{x_\sigma}\left(\frac{x_\mu^{\theta_\mu}\theta_\sigma}{c(\theta_\mu + \theta_\sigma - 1)(\theta_\mu - 1)}\right)^{\frac{1}{\theta_\mu - 1}} \qquad \mu^* := \left(\frac{x_\mu^{\theta_\mu}}{c(\theta_\mu - 1)}\right)^{\frac{1}{\theta_\mu - 1}}$$

$$\frac{Pr_2}{Pr_1} = \frac{\theta_\sigma}{\theta_\mu + \theta_\sigma} \cdot \left(\frac{\theta_\mu + \theta_\sigma - 1}{\theta_\sigma}\right)^{\frac{\theta_\mu}{\theta_\mu - 1}}$$

$$\frac{\partial \log(Pr_2/Pr_1)}{\partial \theta_\sigma} = -\frac{\theta_\mu}{(\theta_\mu + \theta_\sigma - 1)(\theta_\mu + \theta_\sigma)\theta_\sigma} < 0$$

$$\lim_{\theta_\sigma \to \infty} \frac{Pr_2}{Pr_1} = 1$$

2.10.8.1 Proof of Proposition 12

All we need to apply the general results to this case is to establish an excess wealth order between μ and $\mathbb{E}[\mu|\tilde{\mu}]$ and how the latter one changes in θ_σ.

$$\mu \sim \text{Pareto}(x_\mu, \theta_\mu)$$

$$\mathbb{E}[\mu|\tilde{\mu}] \sim \begin{cases} \dfrac{\theta_\mu + \theta_\sigma}{\theta_\mu + \theta_\sigma - 1} * x_\mu, & \text{w/p } \dfrac{\theta_\mu}{\theta_\mu + \theta_\sigma}, \\ \text{Pareto}\left(\dfrac{\theta_\mu + \theta_\sigma}{\theta_\mu + \theta_\sigma - 1} * x_\mu, \theta_\mu\right), & \text{w/p } \dfrac{\theta_\sigma}{\theta_\mu + \theta_\sigma} \end{cases}$$

One could notice that the latter one is in fact a mean preserving contraction of the former one.

We can first derive excess wealth of μ for the quantile q threshold:

$$EW_\mu(q) = \int_{F_\mu^{-1}(1-q)}^{+\infty} (1 - F_\mu(x))dx = \frac{1}{\theta_\mu - 1} * x_\mu^{\theta_\mu} * \mu_q^{-(\theta_\mu - 1)} = \frac{1}{\theta_\mu - 1} * x_\mu * (1-q)^{\frac{\theta_\mu - 1}{\theta_\mu}}$$

Now we can derive it for the delegated search. Let us denote $p \equiv \frac{\theta_\sigma}{\theta_\mu + \theta_\sigma}$ and $A \equiv \frac{\theta_\mu + \theta_\sigma}{\theta_\mu + \theta_\sigma - 1}$.

$$EW_{\mathbb{E}[\mu|\tilde{\mu}]} = \int_{F_{\mathbb{E}[\mu|\tilde{\mu}]}^{-1}(1-q)}^{+\infty} (1 - F_{\mathbb{E}[\mu|\tilde{\mu}]}(x))dx = \frac{1}{\theta_\mu - 1} * p(x_\mu * A)^{\theta_\mu} (\mu_q')^{-(\theta_\mu - 1)}$$

where

$$\mu_q' = A x_\mu * \left(\frac{p}{1-q}\right)^{\frac{1}{\theta_\mu}}$$

Therefore

$$EW_{\mathbb{E}[\mu|\tilde{\mu}]} = EW_\mu(q) * A * p^{\frac{1}{\theta_\mu}} = EW_\mu(q) * \frac{\theta_\mu + \theta_\sigma}{\theta_\mu + \theta_\sigma - 1} * \left(\frac{\theta_\sigma}{\theta_\mu + \theta_\sigma}\right)^{\frac{1}{\theta_\mu}} = EW_\mu(q) * \left(\frac{\Pr_2}{\Pr_1}\right)^{-\frac{\theta_\mu - 1}{\theta_\mu}}$$

Given the proof before we know that $EW_{\mathbb{E}[\mu|\tilde{\mu}]}$ is increasing in θ_σ and converges to 1 if $\theta_\sigma \to +\infty$. Thus μ EW order dominates $\mathbb{E}[\mu|\tilde{\mu}]$. Also, $\mathbb{E}[\mu|\tilde{\mu}]$ is EW increasing in θ_σ and converges to μ as $\theta_\sigma \to +\infty$. Theorem 3 and Proposition 10 conditions are satisfied which finishes the proof of Proposition 12.

2.10.9 Lognormal Productivity Distribution

Lemma 6 *Given two lognormal random variables $X_1 \sim LogNormal(u_1, s_1^2)$ and $X_2 \sim LogNormal(u_2, s_2^2)$ with the same mean, $X_1 \succ_{EW} X_2$ if $s_1 > s_2$.*

Proof. As shown by Belzunce et al. (2016) Corollary 2.1, a sufficient condition for $X_1 \succ_{EW} X_2$ is (1) $\lim_{p\to 0^+} F_{X_1}^{-1}(p) - F_{X_2}^{-1}(p)) \leq E[X_1] - E[X_2]$ and (2) that there exists a value p_0 such that:

1. $F_{X_1}^{-1}(p) - F_{X_2}^{-1}(p)$ is increasing on $[p_0, 1)$.
2. $F_{X_1}^{-1}(p) - F_{X_2}^{-1}(p)$ is decreasing on $(0, p_0)$.

Condition 1 is easy to check because in our case, X_1, X_2 have the same mean so the condition amounts to the $\lim_{p\to 0^+} F_{X_2}^{-1}(p) - F_{X_1}^{-1}(p)) \leq 0$. Since lognormal distribution are bounded below by 0, this condition holds because the quantile function evaluated at 0 for any lognormal is 0. To check the second condition, we write out the difference in quantile functions explicitly:

$$F_{X_1}^{-1}(p) - F_{X_2}^{-1}(p) = exp\Big(u_1 + s_1(\sqrt{2}erf^{-1}(2p-1))\Big) - exp\Big(u_2 + s_2(\sqrt{2}erf^{-1}(2p-1))\Big)$$

Taking the derivative wrt p gives:

$$s_1\sqrt{2}\sqrt{\pi}exp(-erf^{-1}(2p-1)^2)F_{X_1}^{-1}(p) - s_2\sqrt{2}\sqrt{\pi}exp(-erf^{-1}(2p-1)^2)F_{X_2}^{-1}(p)$$

We can factor out portions because they are positive:

$$\sqrt{2}\sqrt{\pi}exp(-erf^{-1}(2p-1)^2)\Big(s_1F^{-1}_{X_1}(p)-s_2F^{-1}_{X_2}(p)\Big)$$

Since the outside is positive we can focus just on the inside:

$$s_1F^{-1}_{X_1}(p)-s_2F^{-1}_{X_2}(p)=s_1exp\Big(u_1+s_1(\sqrt{2}erf^{-1}(2p-1))\Big)-s_2exp\Big(u_2+s_2(\sqrt{2}erf^{-1}(2p-1))\Big)$$

This is positive when:

$$log(s_1/s_2)+u_1+s_1(\sqrt{2}erf^{-1}(2p-1)))\geq u_2+s_2(\sqrt{2}erf^{-1}(2p-1)$$

Since $s_1 > s_2$, we can re-write again as:

$$\frac{log(s_1/s_2)+u_1-u_2}{\sqrt{2}(s_2-s_1)}\leq erf^{-1}(2p-1)$$

The left is constant in p. The right however is increasing in p, and as p approaches 0 the right approaches negative infinity meaning that the original function $F^{-1}_{X_1}(p)-F^{-1}_{X_2}(p)$ is decreasing over an initial range. As p approaches 1 the right approaches positive infinity, meaning that eventually the function is increasing. Thus there exists a value p_0 before which the function is decreasing and after which it is increasing if $s_2 < s_1$. Therefore we can apply the corollary from Belzunce et al. (2016) and $X_1 \succ_{EW} X_2$ if $s_1 > s_2$. The lemma implies that we can rank lognormal random variables with the same expectation using the shape parameter s. Higher s implies a higher excess wealth order.

Lemma 7 *Suppose two random variables X_1, X_2 are distributed according to Assumption 2. Then $V := E[X_1|X_1X_2^{-1}]$ is lognormal. Also $s_V < s_1$ as long as they are not perfectly positively dependent.*[19]

Proof.

19. As long as the underlying normal random variables have a correlation coefficient that is not 1.

$$(X_1, X_2) = \exp(Z_1, Z_2)$$

$$(Z_1, Z_2) \sim N(\cdot)$$

The distribution of $Z_1|Z_1 - Z_2 = y$ is:

$$N\left(u_1 + \frac{s_1^2 - s_{1,2}}{s_1^2 + s_2^2 - 2s_{1,2}}\left[y - u_1 + u_2\right], (1 - \frac{(s_1^2 + s_{1,2})^2}{(s_1^2 + s_2^2 - 2s_{1,2})(s_1^2 - s_{1,2})})s_1^2\right)$$

Thus $W := exp(Z_1)|Z_1 - Z_2 = y$ is log-normal with u, s parameters corresponding to the mean and variance of the underlying normal random variable $Z_1|Z_1 - Z_2 = y$. Then $E[exp(Z_1)|Z_1 - Z_2 = y]$ is $exp(u_W + s_W^2/2)$ which is:

$$exp\left(u_1 + \frac{s_1^2 - s_{1,2}}{s_1^2 + s_2^2 - 2s_{1,2}}\left[y - u_1 + u_2\right] + (1 - \frac{(s_1^2 + s_{1,2})^2}{(s_1^2 + s_2^2 - 2s_{1,2})(s_1^2 - s_{1,2})})s_1^2/2\right)$$

Notice that if we denote:

$$\gamma = exp\left(u_1 + \frac{s_1^2 - s_{1,2}}{s_1^2 + s_2^2 - 2s_{1,2}}\left[- u_1 + u_2\right] + (1 - \frac{(s_1^2 + s_{1,2})^2}{(s_1^2 + s_2^2 - 2s_{1,2})(s_1^2 - s_{1,2})})s_1^2/2\right)$$

$$\zeta = \frac{s_1^2 - s_{1,2}}{s_1^2 + s_2^2 - 2s_{1,2}}$$

we can write:

$$E[exp(Z_1)|Z_1 - Z_2 = y] = \gamma exp(\zeta y)$$

This is the conditional expectation function. To get the distirbution of $E[exp(Z_1)|Z_1 - Z_2]$ we apply this function to the random variable $Z_1 - Z_2$. This variable is equivalent to the log of X_1/X_2 so: $V = E[X_1|X_1X_2^{-1}] = E[exp(Z_1)|exp(Z_1 - Z_2)] \stackrel{d}{=} \gamma exp(\zeta log(X_1/X_2)) = \gamma(X_1/X_2)^{\zeta}$

By known properties of lognormal random variables, we also have that the power and multiplication of a lognormal remains lognormal, so:

$$\gamma(X_1/X_2)^\zeta \sim LogNormal(\zeta u_{X_1/X_2} + log(\gamma), \zeta^2 s^2_{X_1/X_2})$$

We now employ the last lemma to show that this distribution is excess wealth order dominated by the original distribution X_1. Because $E[X_1|X_1X_2^{-1}]$ is a mean-preserving contraction of X_1, we know the mean condition is satisfied. We then need only show the shape paramater is less than the original shape parameter of X_1, that is we wish to know if:

$$s_V^2 = \zeta^2 s^2_{X_1/X_2} < s_1^2$$

Plugging in $s^2_{X_1/X_2}$ and ζ:

$$\frac{(s_1^2 - s_{1,2})^2}{(s_1^2 + s_2^2 - 2s_{1,2})^2}(s_1^2 + s_2^2 - 2s_{1,2}^2) = \frac{(s_1^2 - s_{1,2})^2}{s_1^2 + s_2^2 - 2s_{1,2}} < s_1^2$$
$$\implies \frac{s_{1,2}}{s_1 s_2} = corr(X_1, X_2) \leq 1$$

Thus $s_V < s_1$ as long as there is no perfect positive dependence.

Lemma 8 *Suppose two random variables X_1, X_2 are distributed according to Assumption 2. Suppose X_1', X_2' are generated with all the same parameter values except with a higher value for $s_\sigma' > s_\sigma$. Then $s_V' < s_V$.*

Proof. In the last lemma we showed that $V := E[X_1|X_1X_2^{-1}]$ is lognormal when X_1, X_2 are lognormal. Further, the shape parameter s_V can be expressed as:

$$\frac{(s_1^2 - s_{1,2})^2}{s_1^2 + s_2^2 - 2s_{1,2}}$$

This expression is decreasing in s_2, therefore since σ corresponds to X_2, $s_\sigma' > s_\sigma$ implies $s_V' < s_V$. Because the expectation is invariant to changes in this parameter, we have that increases in this parameter result in decreases in the excess wealth order.

Lemma 9 *Suppose $Z \sim Lognormal(0, s_z^2)$ and is independent of X_1, X_2. $V := E[X_1|X_1X_2^{-1} \cdot Z^{-1}]$ is equal in distribution to $V' := E[X_1'|X_1'X_2'^{-1}]$ for some $s_V' < s_V$.*

Proof. Multiplying X_2 by an independent lognormal random variable with shift parameter of 0 is equivalent to a lognormal random variable with location parameter u_2 and shape parameter $s_z^2 + s_2^2$. The unlderying normal random variable has the same covariance with the normal random variable underlying X_1, thus $s_{1,2}$ is unchanged. Therefore it is equivalent to increasing s_2^2 directly. This implies that it is equivalent to adding mean-zero independent normal noise to the underlying normal random variables.

2.10.10 Proof of Proposition 11

Under the lognormal assumption without positive dependence, we can apply Lemma 7 with $X_1 = \mu, X_2 = \sigma$ and conclude that $E[\mu|\tilde{\mu}]$ is dominated by μ in the excess wealth order in a strict sense. Since excess wealth order implies mean preserving spread we have by Proposition 8 that first-best is not achieved. We also have directly by corollary 3.1 that search effort is lower under delegation as is the productivity expectation of hired workers in an FOSD sense. Finally, s_σ satisfies all requirements for θ so we can apply Proposition 10 to conclude profit and search effort increase and converge as s_σ decreases to 0.

If additional $s_{\sigma,\mu} = 0$ we have independence of μ, σ and we can apply Proposition 9 and say that productivity variance of hired workers is higher in the first-best.

2.10.11 Proof of Proposition 13

By the envelope theorem, the derivative of profit in the second-best with respect to search cost is:

$$\frac{\partial \pi^{SB}}{\partial c} = \frac{1}{Pr(\tilde{\mu} \geq \tilde{\mu}^*)} \geq 0$$

And in the first-best it looks similarly:

$$\frac{\partial \pi^{FB}}{\partial c} = \frac{1}{Pr(\mu \geq \mu^*)} \geq 0$$

Thus in both cases optimal profit is monotone decreasing in search cost. First-best profit is always weakly larger than delegated equilibrium profit for the same search cost. Thus there exists a value c^* such that all for all direct search costs above this value the firm delegates, and for all less it searches directly. As heterogeneity in σ decreases in the sense of Proposition 10, profit under delegation rises and it converges to first-best when the search costs are the same. Thus as heterogeneity decreases c^* falls and the firm delegates under a wider range of direct search costs. When σ is degenerate and workers are homogeneous in productivity variance, first-best profit is achieved. Then for any direct search cost that is weakly greater than the delegated search cost the firm delegates. Since we always assume there is some comparative advantage ($c_F > c$) the firm always delegates.

2.10.12 Inefficiency results

Proposition 15 *As soon as the first-best and delegated accepted regions are not identical (almost surely), then neither of them is a subset of the other one.*

Corollary 3.2 *In the delegated equilibrium, there is a set of excluded efficient candidates and a set of included inefficient candidates compared to the First Best.*

Proof.

Lemma 10 *There is a search over objects x in a set A with outcome $h(x)$. Let there be a total order on a finite set of possible search acceptance regions $\{A_k\}_{k=1}^n$:*

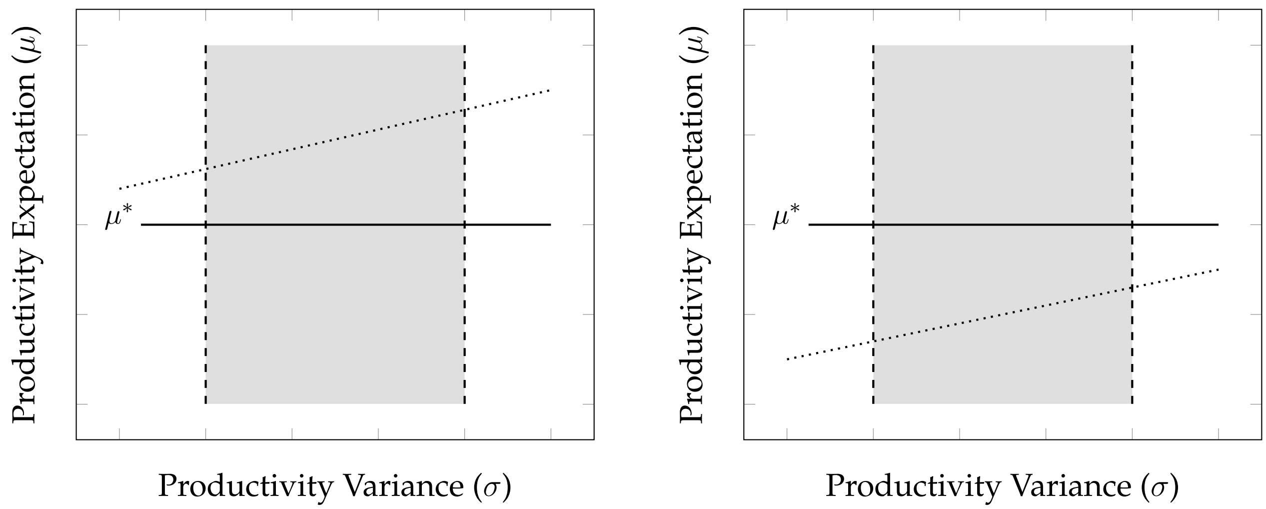
$$A \supset A_1 \supset A_2 \dots \supset A_n$$

such that $\mathbb{E}[h(x)|A_k \backslash A_{k+1}]$ is increasing in k. The excess wealth is defined as

$$EW(A_k) = \Pr(A_{k+1}) * (\mathbb{E}[A_{k+1}] - \mathbb{E}[A_k \backslash A_{k+1}])$$

Then $EW(A_k)$ is decreasing in k.

Proposition 15 is obviously true in the case of full infinite support $S_{\mu,\sigma} = (-\infty, +\infty) \times (0, +\infty)$. In this case, the first-best and the delegated acceptance regions non-trivially intersect and define excluded efficient and included inefficient regions of positive measures. The cases of bounded or discreet support violating Proposition 15 statement are stylized in the two graphs below.



(a) Lower Boundary of Delegation Fully Above μ^*

(b) Lower Boundary of Delegation Fully Below μ^*

Figure 2.7: Acceptance Regions with Bounded Support

Graph (a) can never be the case as the expected ability on the delegated acceptance line is higher than in the first best, which also means that the delegated firm's profit is higher than the first best which cannot be the case.

Using Lemma 10, we want to show that excess wealth for the first-best and delegated acceptance regions are not the same, which would violate one of the sequential search FOCs. All we need to do to apply Lemma 10 is to construct an ordered set of possible acceptance regions incorporating the first-best and the delegated excess wealth.

$$\begin{cases} A_1 = \{(\mu,\sigma) | \tilde{\mu} \geq \tilde{\mu}^*\} \\ A_2 = \{(\mu,\sigma) | \tilde{\mu} > \tilde{\mu}^*\} \\ A_3 = \{(\mu,\sigma) | \mu \geq \mu^*\} \\ A_4 = \{(\mu,\sigma) | \mu > \mu^*\} \end{cases}$$

The only technical condition to verify for applying Lemma 10 is that

$$\mathbb{E}[\mu | A_1 \backslash A_2] \leq \mathbb{E}[\mu | A_2 \backslash A_3] \leq \mathbb{E}[\mu | A_3 \backslash A_4]$$

which follows from staring at the graph and Assumption 1.

Now A_1 is the delegated acceptance region and A_3 is the first-best one and Lemma 10 together with search FOCs yields a contradiction

$$c = EW(A_1) > EW(A_3) = c$$

Therefore A_3 cannot be a subset of A_1. These two must non-trivially intersect creating non-empty excluded efficient and included inefficient regions. ■

2.10.13 Proof of Proposition 14

We can directly apply Proposition 15 to prove the statistical discrimination result as soon as we verify that the conditional expectation $E[\mu|\tilde{\mu} = x]$ is increasing, continuously differentiable and goes from negative to positive infinity. To do this we can write the conditional expectation as:

$$E[\mu|\tilde{\mu} = x] = pE[\mu|\tilde{\mu} = x, \sigma = \sigma_A] + (1-p)E[\mu|\tilde{\mu} = x, \sigma = \sigma_B]$$

which given the definition of standardized productivity evaluates to:

$$E[\mu|\tilde{\mu} = x] = p\sigma_A x + (1-p)x\sigma_B$$

This expression is clearly increasing because all coefficients are positive. It is also linear, meaning it spans the full real line for values of x that span the whole real line. Therefore the conditions for Proposition 15 are satisfied and we have that the delegated and first-best regions are not subsets. Because of the shapes of the regions, this implies that the proportion of hired workers from Group B must be smaller under delegation.

CHAPTER 3

Workplace Injury and Labor Supply within an Organization

3.1 Introduction

Workplace injury is a large economic burden. In the United States, injuries on the job cost $170.8 billion in 2018 alone. Such a cost is comparable to that of more well-known medical issues like heart disease.[1] However, much of the risk-relevant information is known only by the worker. A worker knows if they slept enough the night before. They know if they are feeling sick. They know if they drank too much alcohol at a party yesterday. They understand best their own physical capacity to safely work. At the same time, labor supply varies greatly across people (Blundell, Bozio, and Laroque 2011). Among Los Angeles traffic officers, a city report documents that in a single year, one worker earned $15 in overtime while another earned over $100,000 (Galperin 2015). How do such voluntary labor supply decisions impact workplace injury?

To answer this question, I develop a framework for understanding the connection between labor supply and workplace injury within an organization. I apply the framework to novel high-frequency panel data which details Los Angeles traffic officer work patterns, pay and workers' compensation claims. I use variation in the leave of coworkers to identify how labor supply varies with injury risk. I find daily labor supply is downward sloping in injury risk: officers are less likely to work when they are more likely to be injured. This self-selection generates an observed injury rate among Los Angeles

1. The CDC estimates that in 2014-2015, the annual cost of heart disease was around $219 billion (*Heart Disease Facts* 2020).

traffic officers that is at least 48 percent lower than the underlying average injury rate. My framework allows me to decompose selection against injury into a part that could be deduced by an analyst (predictable component) and part known only by the worker (the private component). For traffic officers, 96 percent of selection is attributable to the private component. The vast majority of injury mitigation comes through unobservable selection that could not be replicated by a manager assigning shifts directly.

This paper has an important practical implication: carefully designed overtime assignment mechanisms can reduce injuries within organizations. Because so much of selection is due to private factors, mechanisms which encourage workers to act on risk-relevant private information will lower injury rates. An auction which awards extra shifts to the workers who bid the lowest wage is one such mechanism. I compare such shift auctions to a system where workers are put in a random list and given the option to accept or reject a shift. I show via simulation that shift auctions result in 11% fewer injuries than the list mechanism.

This paper has three primary contributions: one methodological and two substantive. Methodologically, it provides a framework which links intensive margin labor supply with workplace injury risk. This framework formalizes the connection between the willingness to work and the propensity to be injured. Given a sufficiently strong instrument, it also allows the researcher to identify the average underlying injury rate: the probability of injury of a randomly drawn worker who is forced to work on a random day. I show the approach is numerically equivalent to a marginal treatment effects (MTE) strategy. MTE equivalence allows the researcher to leverage recent advances in the literature to decompose the link between injury and labor supply into a portion due to predictable willingness to work and private willingness to work. The relative importance of these components has an economic interpretation. When the predictable portion is large, a central planner or manager can greatly reduce the injury rate by assigning work using only observables. When the unpredictable portion is large, reducing the injury rate requires eliciting the private information of workers using an appropriate mechanism like an auction.

Substantively, this paper contributes to the large literature across economics, public health and epidemiology which studies the relationship between overtime and health. These papers use data covering a large number of diverse individuals to estimate the association between workplace injury and overtime (Dembe et al. 2005, Kim et al. 2016, Conway et al. 2017). My paper is complimentary: while it is less externally valid, it is more internally valid. I account for unobserved selection into overtime, and in so doing uncover the causal effect of an additional day of work on workplace injury. Importantly, my approach distinguishes between the observed injury rate and the counterfactual average underlying injury rate. I show that for Los Angeles traffic officers, the two are very different quantities. Most of the prior literature focuses on the observed injury rate. Although I cannot claim that my estimates hold for the general population, my results show it is dangerous to equate the observed injury rate with the average underlying injury rate. Estimating the average underlying injury rate requires accounting for labor supply-induced selection. Failing to do so biases estimates towards zero.

The second substantive contribution is to the labor supply literature in economics. Estimates of the intensive margin of labor supply abound in the labor economics literature (Liebman, Luttmer, and Seif 2009, Bargain, Orsini, and Peichl 2014, Blundell, Bozio, and Laroque 2011 , Chetty 2012). I complement this literature by demonstrating how injury risk can be an important unobserved confounder when estimating the elasticity of labor supply with respect to the wage. For a particular occupation, I show that labor supply is more elastic when injury is more likely. Because injury risk varies across jobs and also across the life cycle, this can help researchers interpret differences in labor supply elasticities by age and occupation. It is likely that some of the documented heterogeneity in elasticities is due not just to differences in preferences but also differences in injury risk.

Third, this paper contributes to the literature on compensating differentials (Moran and Monje 2016, Parada-Contzen, Riquelme-Won, and Vasquez-Lavin 2013, Kuhn and Ruf 2013, Viscusi and Aldy 2003). Only recently has the literature begun focusing on workplace safety as a firm specific amenity that can be adjusted (Lavetti 2020, Lee and Taylor 2019, Charles et al. 2019). I reinforce this finding by showing that the way overtime is assigned can greatly change the injury rate at a specific firm. On the worker side, Viscusi and Hersch 2001 and Guardado and Ziebarth 2019 make the important point that workers have some control of their own workplace safety. I affirm this, and suggest a specific pathway: workers can reduce risk by only working shifts when their injury risk is low.

The paper proceeds as follows. I begin by introducing a framework that links high-frequency labor supply decisions and injury risk. I then introduce the data and institutional details. Third, I present the main results of estimation. Fourth, I discuss the implications the results hold for shift assignment mechanisms and labor supply elasticities. For those interested, the Appendix documents how my results can be used to estimate the value of a statistical injury.

3.2 Conceptual Framework

In this section I develop a framework that links high-frequency labor supply decisions and individual injury risk. The framework allows the researcher to estimate whether labor supply decisions mitigate or propagate injury risk within an organization. Throughout, I refer to workers as "officers" because I will utilize my framework to study Los Angeles traffic officers. However, the framework is general and can be applied to other settings.

There are N officers indexed by i who make daily decisions to work on dates $t = 1, 2, .., T$. Denote the binary work decision W_{it} and the binary injury outcome Y^*_{it}. Y^*_{it} is the true underlying injury outcome, which is only observed when an individual works. When an officer does not work, Y^*_{it} is counterfactual. I specify that Y^*_{it} is determined by the following equation:

$$Y^*_{it} = \begin{cases} 1 \text{ if } X'_{it}\beta + C_{i2} + U_{it2} \geq 0 \\ 0 \qquad \text{otherwise} \end{cases} \tag{3.1}$$

X_{it} represents time-varying controls including date fixed effects. The sum $C_{i2} + U_{it2}$ represents what I call *private* injury risk. It is private because it is unknown to the analyst or the organization but may be partially known by the officer. C_{i2} represents time-invariant, person-specific injury risk. It captures factors like chronic health conditions (obesity, heart disease, diet, etc) and demographics. U_{it2} represents factors that make a particular officer more likely to be injured on a particular day. $X_{it}'\beta$ is predictable injury risk, because an organization can predict it given sufficient data.

If an officer does not work then Y_{it}^* is not observed (it is counterfactual). This induces a selection problem. The analyst only observes injury outcomes among individuals who work. Denote Y_{it} as the injury outcome that is observed. Then I have that observed injury is the product of the work decision and underlying injury outcome. Formally:

$$Y_{it} = Y_{it}^* \cdot W_{it} \tag{3.2}$$

Each officer decides to work if the expected utility of work is greater than not working. The utility of work relative to not working takes the linear form $Z_{it}'\alpha + C_{i1} + U_{it1}$. Thus the decision to work is given by:

$$W_{it} = \begin{cases} 1 \text{ if } Z_{it}'\alpha + C_{i1} + U_{it1} \geq 0 \\ 0 \qquad \text{otherwise} \end{cases} \tag{3.3}$$

Z_{it} includes all factors in X_{it} as well as at least one time-varying instrument. The sum $C_{i1} + U_{it1}$ represents *private* willingness to work. It is private because it is unknown to the analyst or the organization but is known by the officer. Similar to the injury unobservables, C_{i1} represents unobserved time invariant taste for work, due to things like a greater enjoyment from the job, or a lower value of leisure. U_{it1} represents unobserved time varying taste for work, driven by factors like wealth shocks, family events, or insufficient sleep the night before. $Z_{it}'\alpha$ is predictable willingness to work, because an organization can predict it given sufficient data.

In order to model the private component of selection in a way that is both simple and flexible. Thus, I specify that private willingness to work and private injury risk are jointly normally distributed, and are independent of all other variables conditional on person-specific means of all time-varying observables (denoted $\bar{Z}_i$).

Assumption 4 *Conditional on Z_i, X_i:*

$$\begin{pmatrix} C_{i1} + U_{it1} \\ C_{i2} + U_{it2} \end{pmatrix} \sim N\left(\begin{bmatrix} \bar{Z}_i\gamma_1 \\ \bar{Z}_i\gamma_2 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

where $-1 \leq \rho \leq 1$ and throughout $\Phi(\cdot)$ is the standard normal CDF.

This approach allows within-worker mean dependence between the unobserved components and the components of Z_{it}. For the rest of the paper I implicitly include $\bar{Z}_i$ in Z_{it} and X_{it}.[2] I will often plot and refer to the demeaned quantiles of the two private components as V_{i1}, V_{i2}:

$$V_{i1} := \phi^{-1}(C_{i1} + U_{it1} - \bar{Z}_i\gamma_1) \qquad V_{i2} := \phi^{-1}(C_{i2} + U_{it2} - \bar{Z}_i\gamma_2)$$

These have a more natural scale and they connect directly to percentiles and to the well-known idea of treatment resistance.

3.2.1 Parameters of Interest

This framework is useful because it allows me to formally define and estimate the following three important quantities:

1. **Observed Injury Rate:** This is the probability of injury conditional on working. It is defined as:

$$\mathbb{E}[Y|W = 1] = \mathbb{E}[Y^*|W = 1]$$

2. This vastly simplifies notation.

It can be estimated as the number of injuries divided by the number of shifts worked.

2. **Average Underlying Injury Rate:** This is the expected probability of injury of a random officer forced to work on a random date. It is defined as:

$$\mathbb{E}[Y^*] = \mathbb{E}_X[\Phi(X'\beta)]$$

3. **Labor supply as a function of private injury risk:** This quantifies how labor supply varies with private injury risk. I denote this function as $L(v)$ throughout and it is defined and can be estimated as:

$$L(v) = \Phi\bigg(\frac{z'\alpha + \rho\Phi^{-1}(v)}{(1-\rho^2)^{1/2}}\bigg)$$

The framework allows me to speak precisely about selection and the connection between labor supply and injury risk. If there is selection against injury the average underlying injury rate should be higher than the observed injury rate, that is:

$$\mathbb{E}[Y^*] > \mathbb{E}[Y|W=1] \tag{3.4}$$

To understand how private injury risk enters the labor supply function, we can analyze $L(v)$.

Lemma 11 *$L(v)$ is strictly decreasing if and only if $\rho < 0$.*

Lemma 11, which is proved in Appendix Section 3.9.1, establishes that whether $L(v)$ is upward or downward sloping depends only on the sign of ρ, an estimated parameter. If $\rho < 0$, labor supply is downward sloping in private injury risk, and the data is consistent with officers using labor supply decisions to avoid injury. If $\rho > 0$, labor supply is upward sloping in private injury risk, and the data is consistent with officers using labor supply decisions to induce injury.

Sleep is a good way to illustrate these ideas. Recent work suggests that people are aware they are too sleepy to drive (Williamson et al. 2014), and driving is a major part of traffic officer's jobs. However, most supervisors do not know how well-rested a given employee is on a given day. Therefore the number of hours of sleep the night prior to a shift is private information about an employee's ability to work safely. If officers avoid working shifts on days when they are sleep deprived, we would expect ρ to be negative.

The direction of the inequality in equation 3.4 determines the overall direction of selection. However, we can decompose overall selection into predictable and private components. The private component is captured by the correlation between private injury risk ($C_{i2} + U_{it2}$) and private willingness to work ($C_{i1} + U_{it1}$). Under the normal specification, this correlation is fully captured by ρ. A negative correlation between these components is consistent with an officer possessing private risk-relevant information and using this information to *mitigate injury.* A positive correlation is consistent with officers possessing private information and using it to *exacerbate injury.* Predictable selection is captured by the correlation between $Z_{it}'\alpha$ and $X_{it}'\beta$.

3.2.2 A Connection to the Marginal Treatment Effect

My framework connects naturally to the marginal treatment effect, as introduced in Heckman and Vytlacil 1999. Work is the treatment, and officers are induced to take the treatment by an instrument, in my case leave of coworkers. The outcome of interest is workplace injury. Because a worker cannot be injured if they do not work, we have that $Y_{it}(0) = 0$. Thus the treatment effect is exactly $Y_{it}(1) - Y_{it}(0) = Y_{it}(1)$, that is the probability of injury conditional on work. The *marginal treatment effect* of work on workplace injury, is then given by:

$$MTE(\tilde{u}, x) = \Phi\left(\frac{X_{it}'\beta - \rho\Phi^{-1}(\tilde{u})}{(1-\rho^2)^{1/2}}\right) \tag{3.5}$$

where $\tilde{u}$ is unobserved resistance to treatment (work). It connects directly to the quantiles introduced earlier: $\tilde{u} = 1 - V_{i2}$. It follows directly from Lemma 11 that the MTE is increasing in $\tilde{u}$ if ρ is positive, just like labor supply as a function of unobserved injury propensity. In this way, the MTE approach is the dual of the labor supply approach. One perspective asks how additional injury risk impacts average labor supply. The other asks how inducing additional labor supply changes the injury rate among marginal workers.

The duality clarifies that my empirical strategy is a marginal treatment effects approach in a panel data setting. The main differences from typical applications of the marginal treatment effect are that I explicitly account for the binary outcome and I relax the usual exclusion restriction. These two adjustments are crucial because workplace injury is quite rare and very little demographic information is available.

Given this equivalence, I can leverage recent developments in the marginal treatment effects literature. In particular, I follow Zhou and Xie 2019 and express the marginal treatment effect as a function of the propensity to be treated rather than covariates (X_{it}):

$$M\tilde{T}E(\tilde{u}, p) = \mathbb{E}_{X_{it}}[MT\mathbb{E}(\tilde{u}, X_{it})|\Phi(Z'_{it}\alpha) = p] = \mathbb{E}_{X_{it}}[\Phi\left(\frac{X'_{it}\beta - \rho\Phi^{-1}(\tilde{u})}{(1-\rho^2)^{1/2}}\right)|\Phi(Z'_{it}\alpha) = p]$$

Now the marginal treatment effect is a function of two scalars with straightforward interpretations. $\tilde{u}$ is unobserved resistance to work. This maps directly to private willingness to work:, specifically $\tilde{u} = \Phi^{-1}(-C_{i1} - U_{it1})$. p is propensity to work. This maps directly to predictable willingness to work, specifically $p = \Phi(Z'_{it}\alpha)$. Thus we can project selection into a private and predictable dimension.

Intuitively, a savvy manager could use historical data and institutional knowledge to derive the predictable component, p. This manager could then use these predictions to assign work to minimize injury. Not so with the private component. Even the most savvy manager can only derive the average relationship, and will never know the exact $\tilde{u}$ for a particular officer on a particular day. This private component captures many things, most prominently private health information, like how much an officer slept or drank the night before. A key element of this paper is estimating the relative importance of the private and public components. If the private component dominates, then an organization which wants to reduce its injury rate will need to design mechanisms which essentially elicit private information from workers. In the language of the mechanism design literature, incentive compatibility will be important.

3.3 Data and Institutional Details

In this section I present an overview of the population being studied: Los Angeles traffic officers. I first review the details of the traffic officer job, overtime assignment, and pay structure. I then present some descriptive statistics and associations observed in their pay and workers' compensation data.

3.3.1 Institutional Details

The population of workers studied in this analysis are Los Angeles traffic officers. The city of Los Angeles is divided into 18 divisions, and work assignments, including overtime, are controlled at the division level. Throughout this study, I will refer to officers who work in the same division (work location) as "coworkers."

Los Angeles traffic officers control their labor supply mainly by working additional overtime shifts. Traffic officers are union employees covered by the Memorandum of Understanding 18 (MOU) between the City of Los Angeles and Service Employees International Union Local 721.[3] According to the MOU, traffic officers are non-exempt employees eligible for overtime pay under the Fair Labor Standards Act (Department of Labor 2017). The MOU describes the manner in which officers are paid for regular as well as overtime and "early report" hours. The city is required to pay a minimum of four hours of premium pay if an employee is required to return to work "following the termination of their shift and their departure from the work location" (MOU, 30). If an officer is required to come into work earlier than their regularly scheduled time, they must be paid one and a half times their hourly rate for the amount of time worked prior to the regularly scheduled time (MOU, 32).

Over 150,867 hours were billed to overtime pay codes in calendar year 2015. This overtime comes from three sources. First, there is overtime arising from excess demand for traffic control due to something like an emergency (i.e. a pipe burst or a broken traffic

3. The version reviewed is available online: cao.lacity.org/MOUs/MOU18-18.pdf

light). Second, there is overtime generated by the absence of a scheduled officer during a normal shift. The data reveals that officers take leave for all sorts of reasons, including bereavement, sickness, vacation, jury duty, etc. For a full list of the various types of leave see Appendix Table 3.16. Finally, there is *special events* overtime, which based on city reports is likely the main source of overtime. Special events include the Los Angeles Marathon, Dodger games, the Oscars, parades, and protests.

The MOU describes the general policy for the assignment of overtime amongst traffic officers. "Management will attempt to assign overtime work as equitably as possible among all qualified employees in the same classification, in the same organizational unit and work location" (MOU, 27). Employees must also be notified 48 hours in advance for non-emergency overtime and unofficial overtime that is not sanctioned by a supervisor is "absolutely prohibited" (MOU, 28). Workers cannot add additional hours to their shift unless authorized. For this reason my paper focuses on the decision to work additional shifts rather than the decision to work additional hours.

The specifics on overtime is assigned to individual officers are not spelled out in the MOU. However, a report by the City Controller's office (Galperin 2015) gives more details about special events overtime. Special events overtime is assigned using a mechanism officers call "spinning the wheel." The generation and assignment of overtime is summarized in Figure 3.1.

Under the wheel spin system, officers first volunteer to be on an overtime list. Each month, the list is sorted according to seniority. As special events become available they are offered sequentially to officers in the order they appear on the list. Once offered a shift, an officer may work it or find a substitute. As more shifts become available it is necessary to request officers from further down the list.

This institutional setting allows me to use leave of coworkers as an instrumental variable to achieve identification. Consider how a coworker (j) going on leave impacts an officer (i). If j goes on leave, two things occur. First, when j goes out sick, the department must "find replacements to perform the individual's regular job duties." Thus the

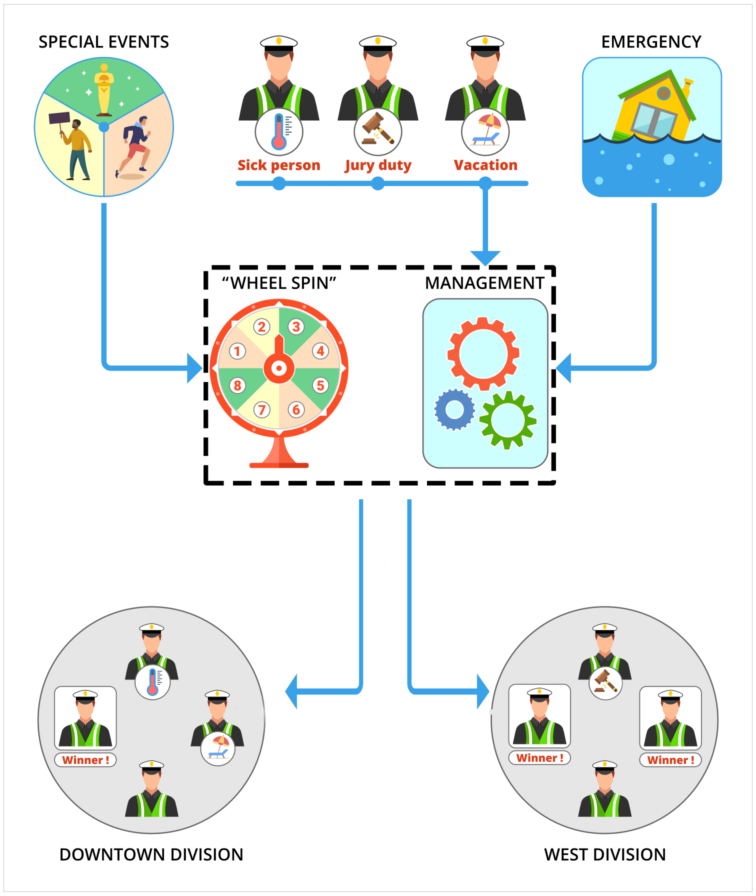


Figure 3.1: The Overtime Assignment Process

absence generates overtime. Second, j's absence means there is one less person in the division (work location) who can take on additional overtime. Both forces lead to an increase in the probability that i will work. It is this variation that I will exploit in order to identify the underlying injury rate.

A reader may be concerned that officers which do not volunteer never have to work overtime. If there are always sufficient volunteers to fill any need, this issue could threaten identification. However, Galperin 2015 states that while only 192 officers signed up to volunteer in FY 2013-2014[4], 471 officers worked overtime. This is evidence that management occasionally exhausts the volunteer list and has to force force non-volunteers to work overtime.

4. This is the year prior to my analysis period which spans part of FY 2014-2015 and FY 2015-2016.

Finally, traffic officers are an ideal population for exploring how injury risk affects labor supply decisions. They receive frequent opportunities to choose to work additional shifts. At the same time, traffic officers represent a middle ground among public safety occupations. The closest occupation with statistics on the BLS website for 2019 was crossing guards and flaggers.[5] In 2019, the nonfatal injury incidence rate was 128.6 injuries per 10,000 workers (*Incidence rates for nonfatal occupational injuries and illnesses* 2020). This was above the incidence rate for firefighters (56.2) and below the incidence rate for police officers (733.8). Traffic officers are representative of occupations where hazards are present (e.g. fast-moving traffic, hot weather) but not pervasive (e.g. carrying a gun, investigating violent crimes).

3.3.2 Data

The analysis population is limited to full-time officers with at least one work-related pay record between January 1, 2015 and September 1, 2016. Additional details regarding how the sample is constructed are listed in the Appendix 3.9.4. The result of the data construction process is an unbalanced daily panel of 553 traffic officers. Table 3.1 reports descriptive statistics at the officer and officer-date level. The typical officer is around 45 years old and is observed working 332 days.

Table 3.1 also includes summary statistics on injuries. I define an "injury" as the submission of a workers' compensation claim. The vast majority of claims list medical expenses paid out, implying that the claim was approved and a real injury occurred. The probability that an injury will occur on any given day is quite low. However, 34 percent of officers are injured at least once in the period studied and 10 percent of officers are injured multiple times. The cause and nature of injuries are tabulated in Appendix Table 3.7. Most injuries are related to the fact that traffic officers work outside in heavy traffic: officers can be sideswiped, get into car accidents, or suffer heat-induced injuries. Injuries span the gamut from superficial to serious.

5. Traffic officers are not exactly crossing guards but are also not exactly police officers.

Table 3.1: Descriptive Statistics

Panel A: Officer Statistics					
	Mean	Std. Dev.	p10	p50	p90
Days Observed	469.91	155.31	194.00	544.00	573.00
Days Worked	332.11	127.88	100.00	379.00	447.00
Injuries Observed	0.46	0.76	0.00	0.00	1.00
Divisions Worked	1.25	0.46	1.00	1.00	2.00
Age	45.14	9.73	32.03	44.65	58.31
Observations	553				

Panel B: Officer-Date Statistics					
	Mean	Std. Dev.	p10	p50	p90
Worked	0.71	0.46	0.00	1.00	1.00
Hours Worked	6.35	4.67	0.00	8.00	12.00
Overtime Pay Hours	1.08	2.74	0.00	0.00	5.00
On Leave	0.02	0.13	0.00	0.00	0.00
Hours on Leave	0.08	0.78	0.00	0.00	0.00
Injured	0.00	0.03	0.00	0.00	0.00
Coworkers on Leave	8.36	8.36	0.00	5.00	21.00
Wage	30.10	2.30	26.64	30.54	32.22
Seniority Rank	30.44	25.36	3.00	22.00	71.00
Observations	259,861				

[1] Age as of January 1, 2015.
[2] Wage is maximum base rate observed on the date.
[3] Worked, On Leave and Injured are indicator variables.

Table 3.2 describes variation in time worked across officers. Panel A presents statistics for the distribution of the hours worked in a day. Panel B presents statistics for days worked in four-week periods. Because an injury causes officers to subsequently miss work, Panel B excludes data after the first observed injury.

From these tabulations of work patterns two things are apparent. First, there is much more variation in the days per shift than the hours per day. The inter-quartile range of shift length is 0, while the inter-quartile range of days worked in four weeks is 5. For this reason I focus on the variability in the number of days rather than the number of hours. Second, employees who experience injury tend to work fewer days per month than those who do not. This fact suggests more injury-prone officers work less.

Table 3.2 displays the distribution of work at the hourly and daily margins. It should

Table 3.2: Distribution of Time Worked

(a) Daily Hours Worked

	Mean	Std. Dev.	p10	p50	p90
Not Injured	9.00	2.70	8.00	8.00	13.00
Injured	8.94	2.62	8.00	8.00	13.00
Total	8.98	2.67	8.00	8.00	13.00
N	183659				

(b) Days Worked in Four Week Period

	Mean	Std. Dev.	p10	p50	p90
Not Injured	18.15	4.44	13.00	19.00	23.00
Injured	17.54	4.24	12.00	18.00	22.00
Total	18.03	4.41	13.00	19.00	23.00
N	8378				

be noted that in Panel A, the sample is restricted to days with positive hours worked. In Panel B, the sample is restricted to 4 week periods with at least one day with positive hours worked.

Table 3.3 describes officer compensation. Most individuals earn a wage that is a little less or a little more than $30 per hour. This is consistent with a common wage schedule which is set during negotiations between the union and the city. Overtime on average represents 12 percent of pay, but this masks a highly skewed distribution. At least 50 percent of officer-weeks do not have overtime pay, while 10 percent are comprised of more than 33 percent overtime pay. Again these statistics indicate that schedules vary most in terms of number of days worked rather than number of hours worked per day.

Table 3.3: Pay Composition Statistics

	Mean	Std. Dev.	p10	p50	p90
Hourly Wage	30.10	2.33	26.56	30.54	32.22
Regular Pay	1236.11	716.25	244.00	1220.00	2135.00
Overtime Pay	287.60	488.18	0.00	0.00	967.00
Proportion OT	0.11	0.14	0.00	0.00	0.33
Observations	43004				

Note: Overtime and straight time are classified based on Variation Description. Wage is the maximum observed base wage during that day. During non-work days it is interpolated.

3.3.3 Descriptive Evidence of Self Selection

The purpose of this section is to show selection against injury is a robust pattern observed in the data, and not an artifact of the framework introduced earlier. If individuals incorporate their own private information about injury risk into their labor supply decisions, and we assume individuals dislike being injured, we should observe positive selection in the data. That is, the probability of injury among officers who work on any given day should be lower than the probability that would result from randomly forcing an officer to work. Mathematically, it should be true that:

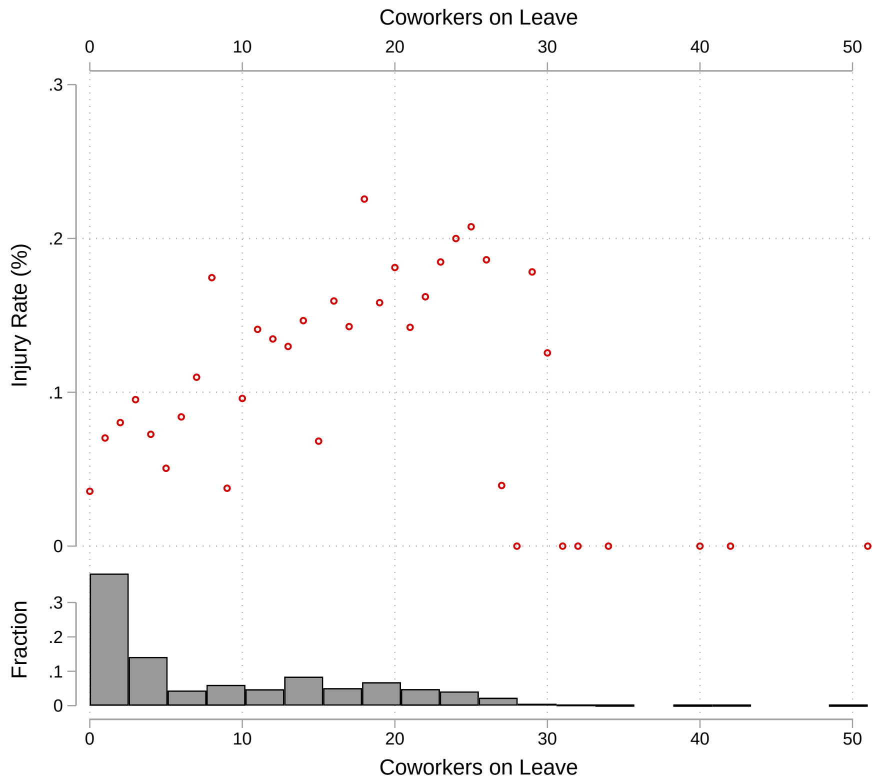
$$\mathbb{E}[Y^*] > \mathbb{E}[Y|W = 1]$$

To trace out how unobserved selection influences injury, I use variation in the leave of coworkers as an instrument. When more coworkers go on leave in a division, management is left with more open shifts and fewer officers to fill those shifts. This should increase the probability that any given officer who is not on leave works without impacting that officer's injury risk directly. We can visually check for selection by graphing a binned scatter plot of injury against the number of hours of leave taken by coworkers. I do this in Figure 3.2.

Figure 3.2 demonstrates that as leave of coworkers increase, the injury probability also rises. This is consistent with self-selection against injury. When officers are given the choice to work, they prefer not to work when their injury risk is elevated. When more coworkers go on leave management must force other officers to work, making the pool less selected thus bringing the injury rate closer to the underlying injury rate.

Comparing regression coefficients makes the same point. In Appendix Table 3.10 I regress injury on work. The coefficient on work is exactly the average underlying injury rate if selection is random. In Table 3.17, I perform several fixed-effects instrumental variable regressions. Comparing any two columns in the tables, we see that the naive estimate is much lower than the instrumental variables estimate. Just as in Figure 3.2, this

Figure 3.2: Evidence of Selection Against Injury



Note: This figure plots the average probability of injury conditional on different values of coworker leave. The probability rises with coworker leave, evidence that the officers select against injury.

result suggests that voluntary labor supply decisions result in selection against injury.

Although these arguments are helpful to establish the presence of selection, quantifying the magnitude and computing the average underlying injury rate requires the framework introduced earlier. I now turn to identification and estimation of the key parameters in the framework.

3.4 Empirical Strategy

The main threat to identification of the average underlying injury rate is that injury risk is likely correlated with the decision to work. Indeed, the entire premise of this paper is to understand the nature of this dependence. To identify and adjust for unobserved selection, I exploit variation in the number of coworkers who are out on leave. When more coworkers go on leave in a division, management is left with more open shifts and fewer officers to fill those shifts. This should increase the probability that any given officer who is not on leave works without impacting that officer's injury risk directly.

My identification strategy is best described as an instrumental variables approach in a binary panel data setting using leave of coworkers as the instrument. Estimation is performed using partial maximum likelihood, with expressions for the likelihood given in Appendix Section 3.9.3. As noted in Semykina and Wooldridge 2018, models of this type camn be estimated as pooled Heckman selection probit models. As a result, I estimate the parameters using Stata's built-in 'heckprobit' command with the addition of person-specific means ($\bar{Z}_i$) in the selection and outcome equations. Standard errors are clustered at the officer level to account for within officer serial-correlation.

The main assumptions required for identification in my model involve the excluded instrument. Leave of coworkers must be properly excluded from the injury equation, it must be sufficiently relevant to the work decision, and it must generate sufficient variation in the support of the propensity to work. I provide evidence that these assumptions are satisfied in Section 3.4.2. The panel structure of the data allows me to relax the exclusion restriction to allow for individual-specific mean dependence between the instrument and the unobserved components.

The reader may wonder why I impose parametric structure on the estimated model, given that the underlying strategy is essentially an instrumental variable approach. This is because injury is a rare outcome and the data is not large enough to accommodate more flexible approaches. This appears to be without loss: my main qualitative findings exist without the structure.[6] Other readers may wonder why I do not use a linear probability model. I do not do this because injury is a rare event and I am concerned with computing the injury rate across the distribution of willingness to work. A linear probability model will result in estimates that are outside the unit interval.

6. See the descriptive evidence section.

3.4.1 Variable Construction

On date t for officer i "leave of coworkers" is constructed as the number of officers in the same division as i on date t that are out on leave for eight or more hours. The majority of variation is therefore at the division-date level. I exclude leave taken by officers who are observed to have worked that day, since such leave might be correlated with injury risk on that specific date and thus violate the exclusion restriction.

In the estimation of the probability of injury, the X_{it}'s includes age, the officer's wage, the officer's seniority rank within the division on date t, division indicators (with small divisions grouped together), and a full set of date fixed effects. Because injury is rare and binary, including date fixed effects comes at the cost of lower statistical power. It weakens the assumptions for identification but it means dates which do not have injuries will not be used during estimation because their log-likelihood estimate is undefined. The effective estimation sample therefore reduces from 259,861 to 80,898 officer-days. All 553 officers remain in the effective sample.

The Z_{it} variables in the probability of working equation include everything contained in X_{it}'s as well as leave of coworkers to satisfy the exclusion restriction. $\bar{Z}_i$ includes officer-specific time averages of leave of coworkers and wage. Seniority rank, division, date indicators and age are excluded from $\bar{Z}_i$ because they are highly co-linear with time. For example, age can be perfectly predicted from average age and date indicators.[7]

3.4.2 Instrument Validity

Identification of the model requires leave of coworkers to be a valid instrument: it must be properly *excluded* from the injury equation and *relevant* to the work decision. I provide statistical and then theoretical arguments that these assumptions are satisfied. Fortunately, the requirements for identification are weaker than the typical exclusion restriction. The panel nature of the data allows the mean of the leave of coworkers to directly impact the

7. As a result of the lack of variation, trying to include these variables causes convergence problems.

injury outcome. Formally, this is reflected by the fact that $\bar{Z}_i$ enters the injury equation.

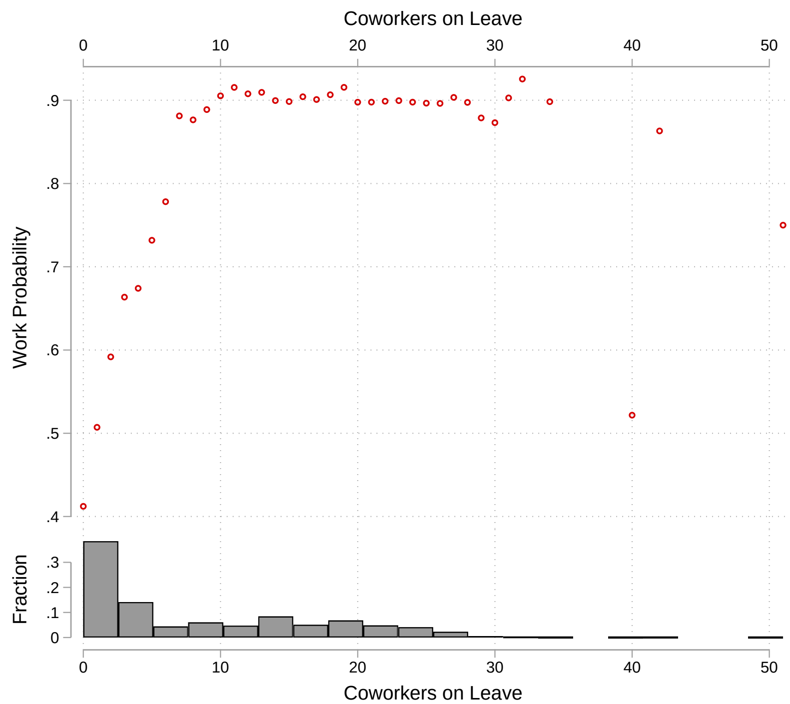
To statistically test the exclusion restriction I conduct a balance test using medical expenses of an injury as a proxy for injury severity. I first restrict the data to the officer-dates with an injury outcome. I then sum the medical expenses paid in the associated workers' compensation claim. Intuitively the exclusion restriction requires that leave of coworkers only impact the binary injury outcome by inducing officers to work. If leave of coworkers is correlated with the severity of an injury this suggests some sort of direct effect. I find no evidence that the exclusion restriction is violated. The results are presented in Appendix Table 3.18. There seems to be no correlation between the severity of an injury and the leave taken by coworkers.

Figure 3.3 presents graphical evidence that leave of coworkers is relevant to the work decision. The figure displays a binned scatter plot of work probability and coworkers on leave, with equally spaced bins by number of coworkers on leave. The relationship is generally upward sloping, indicating a positive link between the instrument and work probability.[8] As a statistical test, I present F-statistics of an analogous linear probability model of work on the leave of coworkers in Table 3.11. All F-statistics are greater than 99. The coefficient on coworker leave is also highly significant in all specifications. Overall the table suggests instrument relevance is satisfied. Additional formal tests of weak instruments are documented in Appendix Section 3.9.6. All tests support the assumption that leave of coworkers is relevant to the work decision and not a weak instrument.

These results appear consistent with our theoretical intuition. Conditional on X_{it} and $\bar{Z}_i$, "coworker leave" must only impact injury through the decision to work. For many forms of leave, like bereavement and jury duty, this seems likely to be satisfied. The death of a coworker's relative is unlikely to affect own work conditions or own health status. For other forms of leave, such as vacations or floating holidays, this requirement is conditionally satisfied. That is, people may tend to take vacations during times of the year when weather conditions contributing to injury risk prevail. For example, more

8. In the Appendix I provide a binned scatter plot which does not interpolate between bins. This plot also supports relevance.

Figure 3.3: Instrumental Relevance



Note: The figure visualizes the relationship between coworkers on leave and probability of working overtime, with relative frequencies underneath. Each circular dot represents each of the values taken on by coworker leave on the x-axis and the corresponding average probability of work for those observations. As the number of coworkers on leave increases, the probability an officer works rises.

vacation may be taken during the summer when heat exhaustion is a factor. But I control for date fixed-effects, and conditional on these, there is likely no dependence. Use of sick leave might violate the exclusion restriction if coworkers are likely to infect each other. To address this concern I estimate the main parameters using a leave instrument that does not include sick time. These estimates are in Appendix Table 3.12 and are discussed in more detail in the robustness section.

Leave of coworkers must also be relevant. Recall that the LADOT uses a "spin the wheel" system to assign overtime. As discussed at length earlier, if more individuals go on leave, the supervisor will need to select a larger number of volunteers and the pool of people available to work will shrink. Thus, conditional on volunteering, the probability of working an extra shift rises. Even if an officer does not volunteer, there is nothing in the memorandum of understanding preventing supervisors from forcing officers to work if the volunteer pool is exhausted. In fact, the MOU describes some overtime as "required," implying that management can force officers to work in certain situations. The MOU also states that many rules are suspended during emergencies. Thus, it is reasonable to assume that the city can force officers to work during incidents such as water main breaks, earthquakes, etc. The probability of any individual officer working in such circumstances will, once again, depend on the size of the pool of available workers. As a result, the probability of working for non-volunteers should also be increasing in the number of other officers on leave.

3.4.3 Identifying the Average Underlying Injury Rate

As mentioned earlier my empirical strategy is equivalent to a marginal treatment effects approach. The average underlying injury rate is then the analogue of the average treatment effect (ATE). Like the ATE, the average injury rate is counterfactual and requires strong conditions on the instrument to be identified and estimated. Fortunately, because workplace injury never occurs when someone does not work, I do not need to satisfy all of the typical marginal treatment effect propensity support conditions. However, identification still requires that the instrument bring the propensity score arbitrarily close to 1 for those who work (Heckman and Vytlacil 1999). If the instrument does not do this, estimating the average population injury rate will rely entirely on the parametric assumption to extrapolate beyond the support.[9]

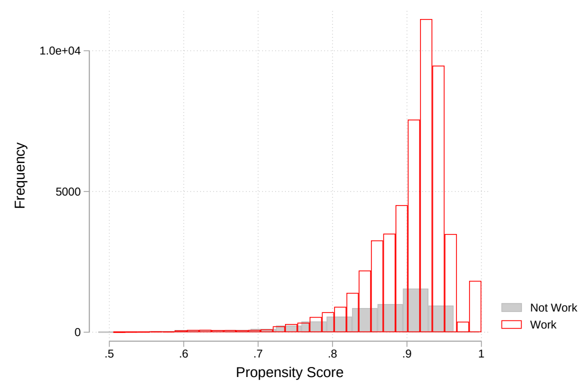
Figure 3.4 plots propensity scores for the weekends (Saturday and Sunday) and Monday through Thursday.[10] We see from this diagram that while the propensity score comes close to 1 for Monday through Thursday, it never rises above 0.6 for Saturday and Sunday. Therefore I do not claim to identify the unconditional average underlying injury rate. I only claim identification of the injury rate conditional on work being performed on a

9. This is often referred to identification at infinity. I thank an anonymous referee for pointing this out.

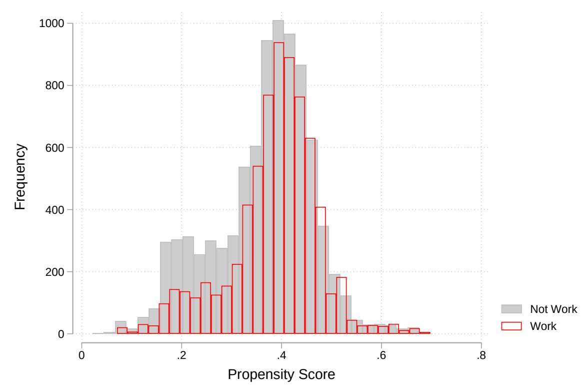
10. Friday is excluded because it is a hybrid between a weekday and a weekend. It has a higher average work probability then the weekdays but a lower probability then the other weekdays.

weekday (Monday through Thursday). Unless otherwise noted, all estimates and plots in the results section average only over the weekday observations. As a result they should be interpreted as objects that are conditional on being a weekday (Monday through Thursday).

Figure 3.4: Support of the Propensity Score



(a) Monday through Thursday



(b) Saturday and Sunday

Note: The figures display the support of the propensity score for weekdays (Mon through Thursday) and weekends (Saturday and Sunday).

Even for weekdays, some readers may still be concerned the support is sparse near 1. To alleviate this concern, I implement a bounding method proposed by Heckman and Vytlacil 1999 in the robustness section.

3.5 Results

This section presents the results. I start with the parameter estimates, then the impact of injury risk on labor supply, followed by the impact of labor supply on injury, and end with a decomposition of selection into predictable and private components. All results support the conclusion that officers use their labor supply decisions to avoid injury. The majority of this selection comes through private rather then predictable factors. Although the model is estimated using all days of the week, most of the results are conditional on the date being a weekday (Monday through Thursday) because of the identification issue discussed earlier. For more discussion, see Section 3.4.3.

3.5.1 Parameter Estimates

Estimates of the most important coefficients in Equations 3.3 and 3.1, as well as ρ (the unobserved correlation between injury propensity and work utility), are presented in Table 3.4. Recall that selection effects as well as decompositions hinge crucially on the estimate of ρ. When ρ is negative there is evidence that officers are utilizing private information about injury risk to avoid working risky shifts.

Due to the non-linear nature of the model, I also report average elasticities of the work probability with respect to several variables in Appendix Table 3.14. I find large wage elasticities: a 1 percent increase in the wage increases the probability a worker takes a shift by 2.27 percent. Leave of coworkers has the expected positive effect. Appendix Table 3.15 reports average elasticities of injury conditional on work. That is, how the observed injury rate responds to changes in the main covariates conditional on the officer having worked. A one percent increase in the number of coworkers on leave results in a 0.23 percent increase in the probability of injury given work.

Table 3.4: Workplace Injury and Labor Supply Model: Select Parameter Estimates

	Injury	Work
Avg. Coworkers on Leave	-0.0638***	0.0143**
	(0.0120)	(0.00611)
Avg. Wage	-0.0590	-0.101***
	(0.0584)	(0.0183)
Wage	0.0756	0.107***
	(0.0610)	(0.0157)
Seniority Rank	0.00165	0.000900
	(0.00142)	(0.000787)
Coworkers on Leave		0.0150***
		(0.00267)
Observations	80898	
Rho	-0.658	
Rho 95% CI	(-0.19, -0.882)	

Standard errors in parentheses

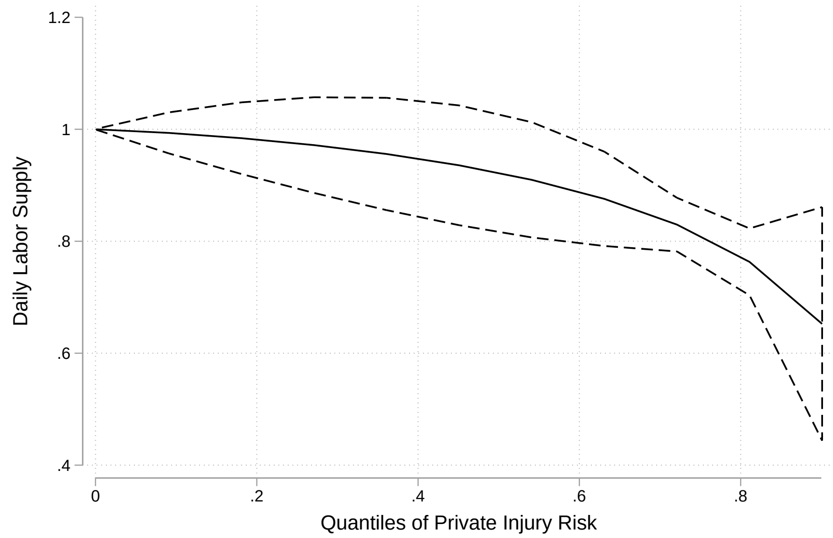
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table displays the main coefficient estimates of the injury and work equations. "Avg." variables are time averages for each person across time periods.

3.5.2 Impact of Injury Risk on Labor Supply

The most important takeaway from Table 3.4 is the estimate of ρ. Since $\hat{\rho} = -0.66$ and the estimate is significant, I can reject the null hypothesis that $\rho = 0$ at the 0.05 level. Recall by Lemma 11, a negative value for ρ leads to a labor supply ($L(v)$) which is downward sloping in private injury risk. Thus, the estimated value for ρ indicates that officers are less likely to work when they are more likely to be injured. Intuitively, this suggests officers are acting on private injury risk information in order to avoid injury. Figure 3.5 plots labor supply as a function of private injury risk. We can compare different points along the graph to understand how officers with different private injury risk make labor supply decisions. An officer at the 80th percentile of unobserved injury risk is 21 percentage points less likely to work on a particular date than an officer at the 20th percentile.

Figure 3.5: Average Daily Labor Supply and Private Injury Risk



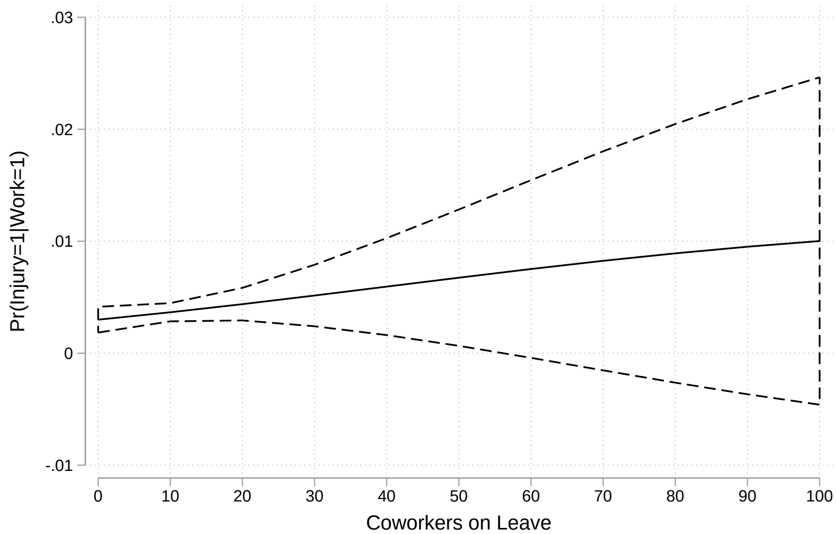
Note: This figure plots daily labor supply (in terms of probability of working) at different levels of private injury risk. Injury risk is in terms of v a quantile-based measure of private injury risk. Higher v indicates higher private injury risk. Dotted lines represent 95 percent confidence intervals with a Bonferroni correction for multiple hypothesis testing. The x-axis is truncated on the right for better visualization.

3.5.3 Impact of Labor Supply on Injury Risk

The last section illustrated how injury impacts labor supply. We can also ask how labor supply impacts injury rates. When more coworkers are on leave, an officer is less able to self-select out of work. This should increase the probability of injury. This is exactly what we observe in Figure 3.6. As the department has to dig deeper into the pool to fill open slots, it has to rely on officers who are less willing to work and thus more likely to be injured.

As I show in section 3.2.2, the effect of voluntary labor supply decisions on workplace injury within an organization is fully captured by the well-known marginal treatment effect function. My framework yields an explicit expression for the marginal treatment effect given by Equation 3.5. I use this expression averaged over the observed covariates values X_{it} to plot the marginal treatment effect in Figure 3.7. The upward slope of the MTE indicates that officers which are more resistant to work are more likely to be injured. If we think of workplace injury as a negative outcome, this represents positive selection.

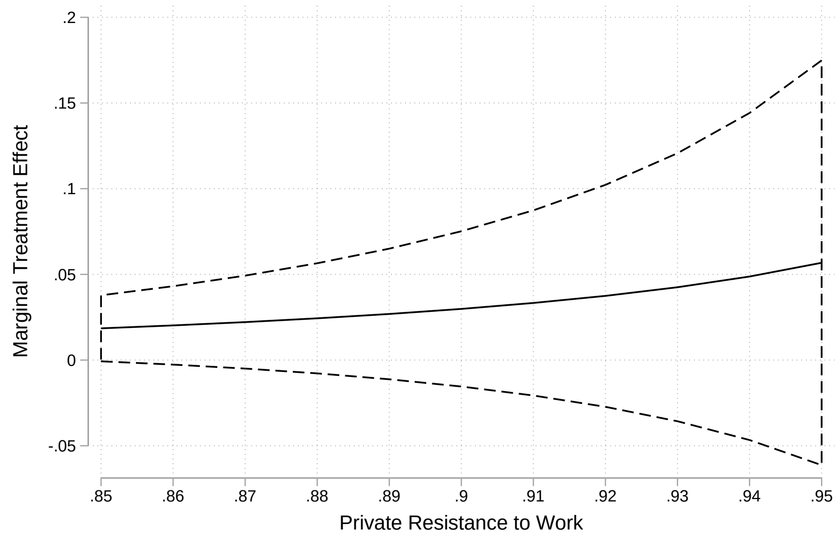
Figure 3.6: Expected Injury Rate Conditional on Different Instrument Values.



Note: Point estimates are averages of both unobserved heterogeneity and covariates. Dotted lines represent 95 percent confidence intervals with a Bonferroni correction for multiple hypothesis testing. The x-axis is truncated on the right for better visualization.

The most risky officers are the least likely to undergo the treatment which is working a shift. To measure the magnitude of selection, we can compare the observed injury rate (injuries divided by days worked) with the average underlying injury rate. Recall that this is the expected injury rate of a random officer forced to work on a random date. It is the average treatment effect of work on workplace injury, and a counterfactual quantity. Comparing the observed and underlying rates demonstrates that selection effects are large: the average underlying injury rate is 1.2 percent while the observed injury rate is 0.38 percent. This means that selection via voluntary labor supply decisions greatly mitigates injury risk among Los Angeles traffic officers. It also means that an officer who is forced to work will generally have a much higher injury risk than an officer who volunteers. In the robustness section I provide a lower bound on the average underlying injury rate which accounts for potential violations of the propensity score requirements. This lower bound is 0.71 percent.

Figure 3.7: Marginal Treatment Effect of Work on Workplace Injury



Note: This figure visualizes selection against injury: officers who are more resistant to work are more likely to be injured. Dotted lines represent 95 percent confidence intervals with a Bonferroni correction for multiple hypothesis testing.

3.5.4 Decomposing Selection

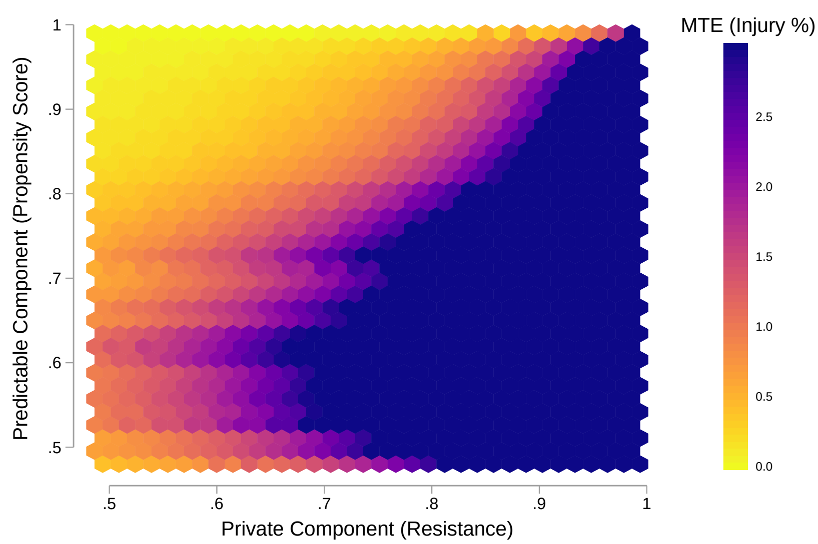
Recall that the work decision is based on two pieces: a predictable component, $Z_{it}'\alpha$, and a private component $C_{i1} + U_{it1}$. When the sum of the two components is positive, the officer works. In this section we establish how much of selection comes through each component. If we assume $Z_{it}'\alpha, X_{it}'\beta$ are jointly normal, the fraction of selection due to the private component is given by:

$$\lambda := \frac{\rho^2}{\rho^2 + Corr(Z_{it}'\alpha, X_{it}'\beta)}$$

I can estimate λ by replacing all the variables on the right hand side with their empirical counterparts from the estimated model. The only additional step is to compute the correlation between the linear predictions. I find that $\hat{\lambda} = 0.96$, meaning 96 percent of selection is due to private factors while the remaining 4 percent is due to predictable factors. We can also follow Zhou and Xie 2019 and express the marginal treatment effect as a function of just the propensity score (another measure of the predictable component) and resistance to work (another measure of the private component). Because we have an exact expression for the MTE, I can create a grid of the average MTE for various values and see how much variation is explained by each component. Using this method, I find 82 percent of variation in the marginal treatment effect is attributable to private factors while 18 percent is attributable to predictable factors. Both methods confirm that the majority of selection is due to private factors: things like private health information and demographics that a manager either could not predict or could not legally use to assign work. I explore in the discussion section how this finding implies that carefully designed overtime assignment mechanisms can help reduce injury rates.

Motivated by a similar diagram in Zhou and Xie 2019, I visualize the patterns of selection across the private and predictable components in Figure 3.8. In the diagram we can

Figure 3.8: Decomposition of Selection



Note: This figure plots the marginal treatment effect for different values of the predictable component (expected labor supply given observables) and the private component (unobserved resistance or willingness to work).

think of each point as representing a different type of officer. As we move along the x-axis from left to right, officers become less willing to work along the private component. The unobserved parts of their net utility from work become lower. As we move from bottom to top along the y-axis, officers become more willing to work along the predictable component. The shading represents the expected injury rate of each type of officer. Darker shading indicates a higher injury rate while light shading indicates a lower injury rate.

Figure 3.8 paints a fuller picture of the two dimensions of selection. The fact that the darkest portions of the diagram are in the right bottom corner indicates that private and predictable selection move in the same direction. Officers who we expect to work (higher propensity score) generally have a lower marginal treatment effect (average injury rate) for all values of resistance. Officers who have a higher unobserved resistance tend to have a lower marginal treatment effect for all values of the propensity score.

Figure 3.8 can be used to think about the problem of a central planner trying to minimize the injury rate of an organization subject to some labor supply constraint. Suppose the planner can choose who to assign a shift, but can only base their decision on observable factors. Such a planner will only be able to exploit selection along the y-axis (predictable component). This is quite limiting: the planner could greatly reduce injury rates if it is able to use the x-axis as well (private factors).

The mechanism design literature tells us that there exist mechanisms, like shift auctions, which will induce officers to reveal their private willingness to work. Because the private components are highly correlated, the social planner can design mechanisms which also extract officer's private injury risk. Section 3.7.1 provides a concrete example of these ideas.

3.6 Robustness

I estimate several variations of the model to test the sensitivity of the main results and detect any potential threats to identification. A summary of parameter estimates under each specification is provided in Appendix Table 3.13. For each specification, I report the coefficient on leave of coworkers as well as $\hat{\rho}$ and the average underlying injury rate.

First, I construct a more conservative version of the leave instrument, which excludes sick time. I do this out of concern that sick leave violates the exclusion restriction. For example, diseases may be contagious and thus there could be a direct effect of the sick days taken by others on own injury risk. Alternatively, increased sick leave might make the remaining pool of officers on average more healthy. This conservative instrument has considerably less variation, because sick time represents a fourth to a third of leave.[11] For this analysis only I additionally provide all main coefficient estimates in Appendix Table 3.12. All estimates remain relatively stable,

Second, I test the sensitivity of my results to changes in the definition of injury. Because I measure injuries as workers' compensation claims, there is a concern that false reporting of injuries might be biasing my results. Claims are verified by medical professionals, but for hard-to-verify injuries, like strains and mental stress, over-reporting might still be a concern. If this is true, the selection I observe could be generated by correlation between an officer's propensity to file false claims and their unwillingness to work. To address this, I estimate my model again with claims described as "Strain" not considered injuries. Out of 243 injuries, 118 are classified as a "Strain." This removes almost 50 percent of the injuries, so it is not surprising that my estimates fall in magnitude and statistical significance. However, it is reassuring that all estimates remain qualitatively similar: $\hat{\rho}$ remains negative and the average underlying injury rate remains of a similar magnitude.

Third, I run the analysis reclassifying injuries based on a thresholds for medical expenses. We can assume that more expensive claims are more serious injuries, and more serious injuries are less likely to be falsely reported. I re-estimate the model with low value claim–those more likely to be fraudulent–reclassified as non-injuries. First I estimate the model reclassifying claims incurring \$0, then reclassifying those incurring less than \$200, and finally, those less than \$400. Surprisingly, $\hat{\rho}$ actually rises as I raise the minimum expense threshold. Similarly, the average underlying injury rate also rises as more claims fall below the minim threshold. This result suggests that if there is fraudulent reporting of workplace injury, it is likely causing me to underestimate selection against

11. See Appendix Table 3.9.

risk.

Lastly, I address concerns about the support of the propensity score. As mentioned previously, I only claim to identify the weekday average underlying injury rate. This is mainly because my instrument does not generate sufficient support of the propensity score for weekend dates. Still, even among weekdays there are very few observations where an individual works and the propensity score is above 0.98. This might leave some concerned that the support condition is still not fully satisfied, and that the resulting estimates rely on extrapolation and identification at infinity.

To account for this I estimate bounds on the average injury probability. Appealing to the fact that the average underlying injury rate is the average treatment effect of work on injury, I use the results derived in Heckman and Vytlacil 1999. Define $\bar{p}(x)$ as the maximum observed propensity score for covariate pattern x among individual-days with $W = 1$. Then we have:

$$\mathbb{E}[Y^*|X = x] = \int_0^{\bar{p}(x)} \mathbb{E}[Y^*|X = x, U = \tilde{u}]d\tilde{u} + \int_{\bar{p}(x)}^1 \mathbb{E}[Y^*|X = x, U = \tilde{u}]d\tilde{u}$$

The first integral is always observed. Because injury is a binary event, the second integral is bounded between $[0, 1 - \bar{p}(x)]$ which implies that the average underlying injury rate for covariate pattern x is bounded in the following way:

$$\int_0^{\bar{p}(x)} \mathbb{E}[Y^*|X = x, U = \tilde{u}]d\tilde{u} \leq \mathbb{E}[Y^*|X = x] \leq \int_0^{\bar{p}(x)} \mathbb{E}[Y^*|X = x, U = \tilde{u}]d\tilde{u} + 1 - \bar{p}(x)$$

Note the interval collapses to a point when the maximum observed propensity score is 1. Because workplace injury is a rare event, the upper bound is not informative. However because I am generally concerned about whether the average underlying injury rate is higher than the observed injury rate, the lower bound is my focus. I set $\bar{p} = 0.98$ based on the plots of the propensity score. I then approximate the first integral using the midpoint method. The procedure generates a lower bound for the weekday average underlying injury rate of 0.71 percent. This is lower than the main estimate of 1.2 percent I report but still nearly double the observed weekday injury rate of 0.37 percent. This is evidence that the main qualitative result does not rest on identification at infinity or functional form extrapolation.

3.7 Discussion

The traffic officers I analyze are assigned overtime through a relatively simple system: extra shifts are given to volunteers based on seniority and whether or not the person has already worked overtime during the relevant period. This system is not designed to reduce injury rates. It is designed to maximize ex-ante fairness. Strikingly, however, it still generates a large amount of positive selection which drives the observed injury rate to be much lower than the underlying injury rate. This is because it gives officers opportunities to self-select out either by not volunteering or declining a shift.

This result lends some nuance to news stories about overtime among public safety professionals. Many articles are alarmed by the massive amount of overtime worked by certain fire fighters and police officers (Ashton and Reese, Steinbach 2019). My analysis suggests such massive overtime is not necessarily a problem for workplace injury. I analyze 553 officers over 609 days. The median number of days worked is 379, but the top 10 percent of officers work more than 447 days. One officer worked 601 of the 609 days. The data cannot speak to the quality of the work performed by an officer who works almost everyday. However, my results indicate this overtime inequality reflects a process which is helping to reduce injury.

To see this, notice what I call the average underlying injury rate is also the counterfactual injury rate we would observe if work assignments were determined mechanically by a random number generator, and all officer choice was removed. Such a system assures equality in overtime outcome: all officers can expect to work the same number of shifts. The observed injury rate is the rate which arises under the spin the wheel mechanism, which assures equal overtime opportunity but not equal overtime outcomes. The fact that the observed rate is so much lower than the underlying rate implies that achieving overtime equality will come at the cost of more injuries.

Thus, inequality in the distribution of overtime ex-post is not necessarily bad in terms of injury rates, as long as the inequality is generated by a voluntary process. Indeed, my results highlight that the distinction between mandatory and voluntary work is of first-order concern when it comes to injury. Many descriptive analyses have shown a positive relationship between excessive work and workplace injury. These include studies using the NLSY (Dembe et al. 2005), a survey of fire fighters in Korea (Kim et al. 2016) and an analysis using the PSID (Conway et al. 2017). Importantly, these studies do not distinguish between mandatory and voluntary overtime. Under my framework, we can think of mandatory overtime as shifts worked when willingness to work is low. I have shown both private and predictable willingness to work is negatively correlated with injury risk. This implies mandatory overtime is more dangerous than voluntary overtime. Because of this, analyses which lump mandatory and voluntary overtime together will always be estimating a weighted average of the mandatory and voluntary effect. Additionally, two identical companies employing identical populations of employees could still have completely different observed injury rates if they allocate work differently. Organizations which rely on voluntary mechanisms will tend to have lower injury rates, while organizations which force employees will tend to have higher rates.

Because I have variation in wages, I am able to estimate the value of a statistical injury for different traffic officers. Because this is not the main focus of the paper, the calculations and estimates are provided in Appendix Section 3.9.8. One observation worth noting is that even within a single occupation and a single organization, the value of a statistical injury can be heterogeneous. This is illustrated in Appendix Figure 3.12.

3.7.1 Shift Auctions

The decomposition of selection into a private and predictable component revealed that the majority of selection is private. I have shown in this paper that the current mechanism used to assign traffic officers to shifts leverages *some* of this private selection. In this section, I demonstrate both theoretically and via simulation that organizations can reduce the injury rate further by using a better mechanism, specifically an auction.

First, recall a few results from auction theory. In a second-price auction with private values, it is weakly dominant to bid one's true value. As a result, the auction will generally assign the object to whoever values it most. When officers are the bidders and shifts are the objects, net utility from work is the value. Suppose officers bid by posting a wage, and the winning officer is the one who posts the lowest wage. Suppose it is a second-price auction, so the winning officer works the shift at the second-lowest wage bid. From auction theory we know that in equilibrium, the winning officer will be the one with the highest net utility.

We can use the earlier estimated results to analyze the efficacy of such a mechanism. I have shown net utility is negatively correlated with injury risk (through both the private and predictable component). The coefficient on wages in the work decision is also positive, indicating officers value wages. This means officers will trade-off wages and injury risk, and the winning officer will have one of the lowest expected injury risks. In theory this should increase selection against injury, because we no longer assign shifts sequentially but rather let officers compete for the shift. Intuitively, shift auctions induce officers to truthfully reveal their injury risk.[12] In this way a shift auction should theoretically improve upon the status quo. Revisiting Figure 3.8, a shift auction would free a manager to use variation along both dimensions to minimize injury rates.

I now show via simulation that auctions reduce injury rates. I compare a shift auction like the one described previously to a random list mechanism. I employ a simple random list mechanism similar in spirit to the status quo spin the wheel system. A full description of the simulation of the list and shift auction mechanisms is given in Appendix Section 3.9.7. I perform 1,000 simulations. On average, the shift auction mechanism generates an average daily injury rate of 0.49 percent, while the random list mechanism generates an average daily injury rate of 0.55 percent. This means the shift auction results in a 12

12. This intuition comes via the revelation principle.

percent decrease in the injury rate.

I also compare shift auctions to what I call the *full information benchmark*. The full information benchmark is the injury rate that would be observed if a manager could assign additional shifts directly to the employees with the lowest injury risk. To simulate it, I randomly assign regular shifts among officers who are willing to work, and then I assign the additional shifts to the officers with the lowest injury risk. The full information benchmark results in an average daily injury rate of 0.40 percent.

The simulation results are summarized in Figure 3.9. The figure displays the simulated injury rate under all three regimes plotted for 1,000 simulations (assuming the number of shifts worked is constant). The injury rate distribution when officers bid for shifts approaches the full information benchmark, and yields much lower injury rates than the random list. This exercise highlights the practical implication of my results: because so much of selection is driven by private, unobserved factors, carefully designed mechanisms which induce officers to act on their private information can reduce an organization's injury rate. The shift auction is one such mechanism: because officers value wages but dislike injury, the winning bidder will have low injury risk.

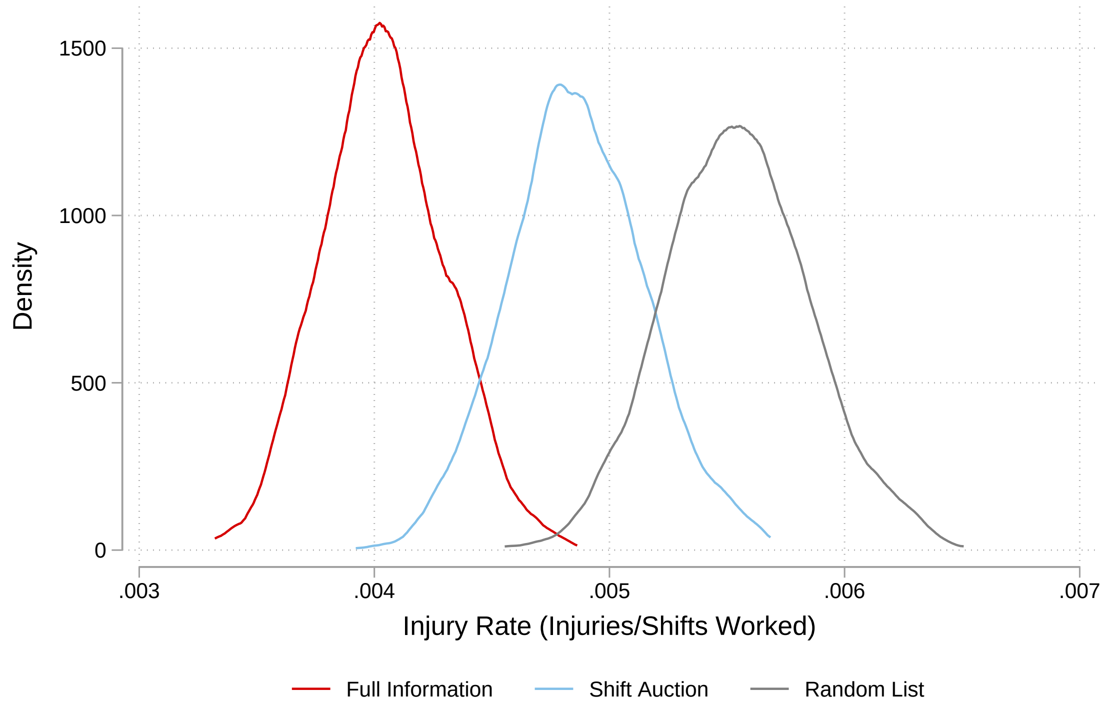
In the simulated shift auction, a manager posts the available shifts, and officers may place a wage bid for each. The shift is then assigned to the officer who bids the lowest wage. Although shift auctions may seem unorthodox, many scheduling software companies already include such a system as a built-in option.[13]

3.7.2 Labor Supply Elasticity

So far I have established that, all else being equal, officers will work less when they have elevated injury risk. That is, the labor supply curve slopes downward in injury risk. In this section, I quantify how injury risk impacts labor supply elasticities with respect to the

13. Some examples: Stay Staffed, which produces a nurse scheduling software; Celayix Software, a multi-industry workforce management software company; EPay Software, a human capital management provider.

Figure 3.9: Simulated Injury Rates Under Three Mechanisms



	mean	p5	p95
Random List	0.0055	0.0051	0.0061
Shift Auction	0.0049	0.0044	0.0054
Full Information	0.0040	0.0036	0.0045
Simulations	1000		

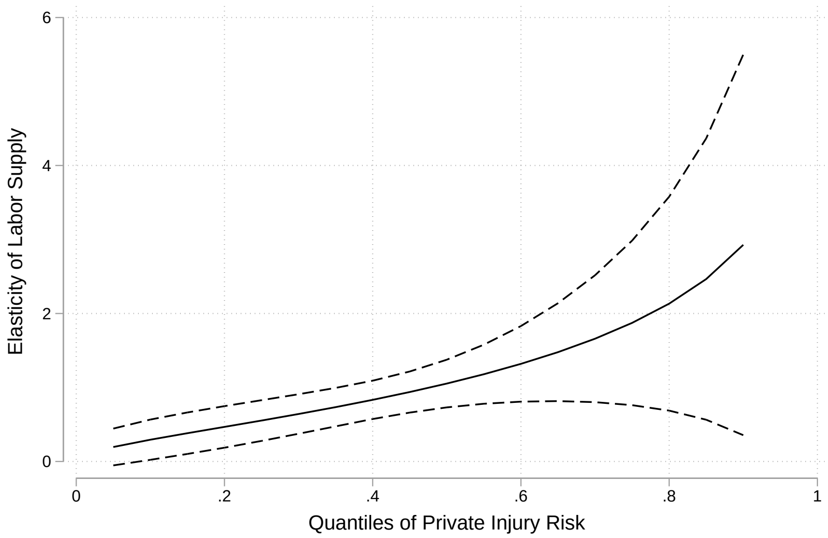
Note: The figure plots the simulated distribution of the injury rate under three different overtime assignment mechanisms. The full information mechanism is the ideal case, when a planner assigns shifts to the officers with the lowest risk. The random list mechanism is similar to the mechanism currently used by the City of Los Angeles, where shifts are given randomly to everyone who volunteers. The shift auction assigns extra shifts to the officers who bid the lowest wage.

wage. My model allows me to estimate the elasticity of the probability of working a shift with respect to the wage conditional on different unobserved propensities to be injured. This allows me to see how elasticities vary at different levels of risk. Formally, I calculate the quantity:

$$e_{wage}(Z_{it}, v) = \frac{wage_{it}}{Pr(W_{it}|Z_{it}, v_{2it} = v)} \frac{\partial}{\partial wage_{it}} Pr(W_{it}|Z_{it}, v_{2it} = v)$$

and average over observed Z_{it}. This yields an average labor supply elasticity for each value of v. I plot this relationship in Figure 3.10 and see that the elasticity is increasing in private injury risk.

Figure 3.10: Average Labor Supply Elasticity by Injury Risk Propensities



Note: The figure displays the average work probability (labor supply) elasticity conditional on different values of unobserved injury propensity. The dotted lines represent a 95 confidence interval with a Bonferroni correction for multiple hypothesis testing.

Appendix Table 3.19 contains the point estimates from Figure 3.10. Individuals with a private injury risk at the 30th percentile have an expected elasticity of 0.642, while those with an injury propensity around the 60th percentile will have an elasticity of around 1.32. As injury risk rises, labor supply becomes more sensitive to wage changes. This illustrates that heterogeneity in injury risk can be an important confounder when estimating intensive margin labor supply elasticities.

3.8 Conclusion

This paper provides evidence traffic officers consider their individual injury risk when deciding whether to work. I identify and estimate a labor supply model utilizing the unique structure of overtime assignment employed by the Los Angeles Department of

Transportation. I establish that daily labor supply is downward sloping in unobserved injury risk, implying officers work less when they are more likely to be injured. This behavior implies the population of officers working on any given day is positively selected (less injury prone) compared to the underlying workforce. I then show this plays a significant role in mitigating observed injury rates.

I also illustrate the practical implications of the main result. I propose shift auctions with workers bidding the wage at which they are willing to work as a mechanism which can leverage self-selection to reduce injury even more than traditional overtime assignment schemes. I show by simulation that a shift auction reduces the observed injury rate compared to a typical assignment mechanism.

To my knowledge, this paper is the first to explore how workers within a single organization working the same job incorporate private injury risk into high-frequency labor supply decisions. The fact that idiosyncratic injury risk plays such a large role in labor supply decisions raises a number of questions across both economics and public health. Across both disciplines, it suggests current estimates of injury rates are biased downwards. This is because the estimates use observational data, and the observed injury rate will tend to overweight low-risk workers who choose to take on additional shifts. Within economics, it implies injury risk within some jobs is a choice variable, which workers can control through their labor supply. Within public health, the fact that some public safety professionals work massive amounts of overtime may not be bad for injury rates. If it is the result of voluntary labor supply decisions, ex-post inequality in days worked can be evidence of self-selection acting to mitigate injury.

It has long been established that workers sort across occupations based on injury risk concerns. Future work should explore how such extensive margin sorting interacts with the intensive margin sorting within an organization established here. It is not clear how such sorting shapes and is shaped by labor market equilibrium.

3.9 Appendix

3.9.1 Labor supply as a function of unobserved injury propensity

The following lemma establishes that whether $L(v)$ is increasing or decreasing depends only on ρ.

Lemma 12 *$L(v)$ is strictly decreasing if and only if $\rho < 0$.*

Proof. Note that:

$$\frac{\partial}{\partial v}\Phi\left(\frac{\zeta_1 + Z_{i,t}'\alpha + \bar{Z}_i'\gamma_1 + \rho\Phi^{-1}(v)}{(1-\rho^2)^{1/2}}\right) < 0 \qquad \forall v$$

for any value of $\zeta_1 + Z_{i,t}'\alpha + \bar{Z}_i'\gamma_1$ if and only if $\rho < 0$. Then the expectation is just an integral over values of $\zeta_1 + Z_{i,t}'\alpha + \bar{Z}_i'\gamma_1$, and I can invoke dominated convergence to say that:

$$\frac{\partial L(v)}{\partial v} = \mathbb{E}_{Z_{i,t},\bar{Z}_i}\left[\frac{\partial}{\partial v}\Phi\left(\frac{\zeta_1 + Z_{i,t}'\alpha + \bar{Z}_i'\gamma_1 + \rho\Phi^{-1}(v)}{(1-\rho^2)^{1/2}}\right)\right] < 0 \qquad \forall v \qquad Q.E.D.$$

3.9.2 Additional Traffic Officer Details from the Memorandum of Understanding

The Memorandum also outlines payment guidelines surrounding minimum payments and "early report" pay. The city is required to pay a minimum of my hours of premium pay if an employee is required to return to work "following the termination of their shift and their departure from the work location" (MOU, 30). If an officer is required to come into work earlier than their regularly scheduled time, they must be paid one and a half times their hourly rate for the amount of time worked prior to the regularly scheduled time (MOU, 32). Workers compensation rules are briefly described. For any injuries on duty, salary continuation payments "shall be in an amount equal to the employee's bi-weekly, take-home pay at the time of incurring the disability condition" (MOU, 59).

In regards to the assignment of overtime, the Memorandum has this to say: "Management will attempt to assign overtime work as equitably as possible among all qualified employees in the same classification, in the same organizational unit and work location" (MOU, 27). Employees must also be notified 48 hours in advance for non-emergency overtime and unofficial overtime that is not sanctioned by a supervisor is "absolutely prohibited" (MOU, 28). Workers cannot add additional hours to their shift unless authorized. For this reason my paper focuses on the decision to work additional shifts rather than the decision to work additional hours.

3.9.3 The Partial Likelihoods

$$Pr(y_{it} = 1 | w_{it} = 1, Z_i) = \frac{1}{\Phi(Z_{it}'\alpha + \zeta_1 + \bar{Z}_i'\gamma_1)} \int_{-\infty}^{Z_{it}'\alpha + \zeta_1 + \bar{Z}_i'\gamma_1} \Phi\left(\frac{\zeta_2 + X_{it}'\beta + \bar{Z}_i'\gamma_2 + \rho v}{(1-\rho^2)^{-1/2}}\right)\phi(v)dv$$

$$Pr(y_{it} = 0 | w_{it} = 1, Z_i) = \frac{1}{\Phi(Z_{it}'\alpha + \zeta_1 + \bar{Z}_i'\gamma_1)} \int_{-\infty}^{Z_{it}'\alpha + \zeta_1 + \bar{Z}_i'\gamma_1} \left[1 - \Phi\left(\frac{\zeta_2 + X_{it}'\beta + \bar{Z}_i'\gamma_2 + \rho v}{(1-\rho^2)^{1/2}}\right)\right]\phi(v)dv$$

$$Pr(w_{it} = 1 | Z_i) = \Phi(Z_{it}'\alpha + \zeta_1 + \bar{Z}_i'\gamma_1)$$

$$Pr(w_{it} = 0 | Z_i) = 1 - \Phi(Z_{it}'\alpha + \zeta_1 + \bar{Z}_i'\gamma_1)$$

3.9.4 Data Cleaning and Population Definition

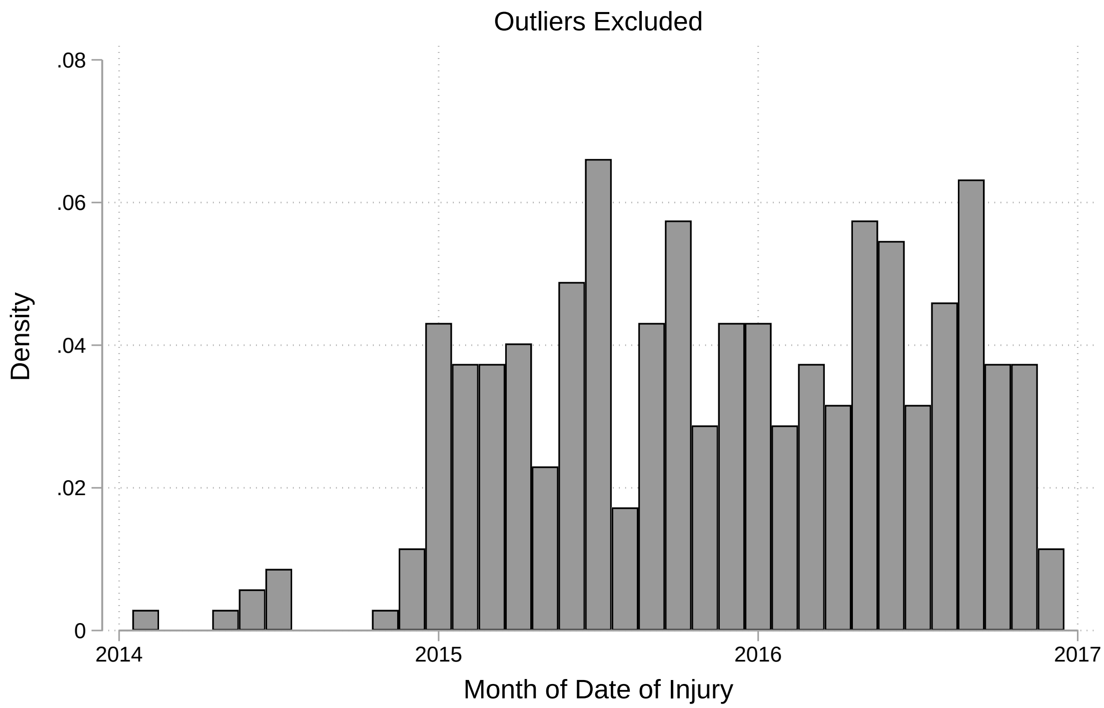
The worker's compensation and payroll data was provided by the City of Los Angeles. The data was de-identified, and spans from 2014 to 2016. It was first provided to a city employee, who performed the de-identification and merged together the two sources. Originally, only the worker's compensation files contained information on employee age and hire date. To the extent an employee was never injured, there would be no age information. A third file was acquired and merged on to fill in gaps of information for employees that were not injured.

The workers' compensation data includes the date of the injury,[14] the date on which

14. It also includes time of injury, but this field says 12:00 AM the majority of the time, suggesting it is not reliable.

the employee gained knowledge of the injury, the nature of the injury, and the cause of the injury. After removing duplicate records, there are 351 distinct worker compensation claims across 246 traffic officers in the time period. Of these, 295 have a non-zero value for "Med Pd" suggesting some sort of expense was paid out to the employee. Figure 3.11 displays the distribution of claims across the period. The claim counts appear abnormally low prior to January 2015 and after September 2016.

Figure 3.11: Workers' Compensation Claims by Month



Note: The figure plots the number of workers' compensation by month. Unlike subsequent months, there are almost no claims prior to January 2015. To avoid confounding the results with observations from a different data generating or reporting system, I limit the analysis window to January 1, 2015 to September 1, 2016.

The pay data includes records for each type of pay received on each day. It also includes the number of hours, amount of pay, rate of pay, division worked, and *Variation Description*. Variation Description is a pay code which describes the reason for a payment. I use Variation Description to classify records as work-related, leave-related, or neither. Table 3.16 displays the classification process.

I aggregate the pay and workers' compensation records into an officer-day panel data set with measures of daily hours worked and hours taken as leave. This process is non-trivial, and requires some assumptions which are outlined in the data-building section of the Appendix. I then perform several important exclusions to create the working sample. First, I limit the study period to workdays and injuries between January 1, 2015 for the reason described above. Second, I exclude all part-time employees (defined as officers having more than three four week periods with less than 60 hours of leave and work) due to their highly irregular schedules.

I include only officer-days where the officer works or does not work, and exclude days where they are on leave. I exclude non-work officer-days that occur after an injury but before the first day worked after injury. I also exclude the first day worked after injury. The decision to return to work after an injury follows a different process than the normal decision to work a shift. The days off work may be medically required. The first day returned also is part of the workers' compensation process and not subject to the normal labor supply decision. Omitting these days allows me to focus on the decision to work a shift, rather than the decision to use a sick or vacation day. Finally, ten injuries occurred on dates without positive work hours. Four of these injuries are associated with the day prior (it appears that the work may have crossed over midnight). Six injuries are assumed to have happened immediately, and the date is considered worked.

3.9.5 Justifying Identification

If one is willing to ignore Equation 3.1 and instead assume a linear probability model for the injury outcome, my model would be a special case of the switching model described in Chen, Zhou, and Ji 2018. Then I could achieve non-parametric identification with a single exclusion restriction and a symmetry condition on the unobservables. But I am not willing to make this simplification, because unlike in other applications, injury for a particular officer on a particular day is quite unlikely, so that $Pr(y_{it}) \approx 0$. Because there are continuous covariates in X_{it}, $X_{it}'\beta$ is unlikely to be bounded between $[0, 1]$ almost surely. According to Horrace and Oaxaca 2006, this makes the linear probability model implausible.

3.9.6 Statistical Tests of the Instrument Validity

Another test of instrument independence examines the balance of other officer-day characteristics across values of the instruments. One such variable is *medical expenses paid*, which is included in the workers' compensation data for each documented injury. Medical expenses are a proxy for the seriousness of injury. For example, injuries with *Claim Cause* "Repetitive Motion - Other" had an average expense of \$2,726, while those with "Collision or Sideswipe" had an average expense of \$3,385. In theory, leave of others and cumulative potential contacts should only impact injury by inducing more people to go into work. Both instruments should not impact the severity of the injury. If they do, then there is reason to suspect that the exclusion restriction does not hold. In Table 3.18, I regress medical expenses paid on the leave instrument with different sets of controls.

For linear models, there are many formal under-identification, over-identification, and weak instrument tests. Unfortunately, my model is nonlinear. In Appendix Table 3.17, I report results from what I call a "proxy" model. It is a fixed effects 2SLS specification (the model I would fit if y_{it} were continuous). Across all specifications, using the Kleinbergen-Paap rK LM test, I reject the null hypothesis of under-identification. Using the Kleinberg-Paap rk Wald F test, I reject the null hypothesis that the instruments are weak. Overall I find no evidence the identifying assumptions are violated in the proxy model.

I can use the proxy model to see how instrument strength impacts the coefficients. Using the tables presented in Stock and Yogo 2002, for my preferred specification (the third model in Table 3.17) the maximum relative bias of the IV estimator is less than 10% (relative to OLS). The Cragg-Donald F-Statistic of my preferred specification is 230. According to Lee et al. 2020, this means I can safely use the 1.96 critical-value for testing hypotheses while maintaining a Type 1 error of 5 percent. This means I have sufficient instrument strength to reject the null hypothesis of random selection into work at the 0.05 level.

3.9.7 Description of Shift Auction Simulation

I first describe the equilibria of the random list and shift auction mechanisms. For shift auctions, I restrict attention to $k+1$-price auctions, where the k overtime shifts in a division are assigned to the lowest k bidders and they are paid the bid of the $k+1$ lowest bidder. Assuming independent values, the unique Bayesian Nash Equilibrium is clearly for each officer to bid their value. The winner in equilibrium will be the officers with the k lowest values. Further, since injury risk is negatively correlated with value, the k winners will have the lowest injury risks among all bidders. In the list mechanism, officers will accept the shift if they are offered it and their value exceeds their outside option. If their value does not exceed their outside option, the shift passes to the next person. Whenever there are more officers willing to work at their normal wage then there are shifts to fill, the officers selected from an auction will have a lower expected injury rate than from the random list. If there are more shifts than officers, and it is assumed that in both mechanisms the shortage is filled by forcing employees to work, then the mechanisms deliver ex-ante the same injury rates. As a result, injury rates will be weakly lower with shift auctions.

To formalize this, consider a fixed day t (from here on I suppress the t subscript). Denote the monetary value of a shift to officer i as θ_i. I can derive the monetary value by setting utility equal to 0 and solving for the wage variable. This yields: $\theta_i := (z_i'\alpha + \zeta_1 + \bar{Z}_i'\gamma_1 - v_{i1})/\alpha_{wage}$ where z_i does not include the wage variable and α_{wage} is the coefficient on the wage variable. The utility from working at bid wage b_i is given by $U_i = \theta_i + b_i$. Recall that the injury outcome is denoted y_i. θ_i and y_i are correlated both through the shared elements of Z_i that enter both the work and injury outcomes and through unobserved correlation.

There are a number of complexities related to how overtime shifts can be assigned. I abstract from these complexities, and consider a simple situation where each division on each date requires $s_{d,t}$ officers, where $s_{d,t}$ is determined as the number of people observed working. Denote total shifts in the the entire analysis period in division d as S_d. I assume that some number of the positions, denoted $r_{d,t}$ are filled by regular officers. The remainder, denoted $k_{d,t}$, are filled with additional officers. Because I do not observe how many shifts are regularly scheduled, I assume that, within each division, it can be approximated as the number of hours coded as "CURRENT ACTUAL HOURS WORKED ONLY" divided by 8.[15] Call this numbers R_d. I also assume the fraction of shifts which are regular is time invariant. This allows us to approximate $r_{d,t}$ as $R_d/S_d \times s_{d,t}$ rounded to the nearest whole number. $k_{d,t}$ is then $s_{d,t} - r_{d,t}$. With these in hand, the simulation procedure I use to obtain injury rates under the random list and shift auctions is as follows:

1. For all officer-days, randomly draw i.i.d. pairs of (v_{it1}, v_{it2}). Then, within each division-date, do steps 2-4.

2. To simulate the list mechanism, randomly select $s_{d,t}$ officers from among those with $z_i'\alpha + \zeta_1 + \bar{Z}_i'\gamma_1 - v_{i1} > 0$ with wage included in z_{it}. If there are not enough officers that satisfy the criteria, fill the remaining slot with randomly chosen officers. Calculate the list-mechanism injuries using the v_{it2} draws of the selected officers.

3. To simulate a shift auction, order the officers according to $z_i'\alpha + \zeta_1 + \bar{Z}_i'\gamma_1 - v_{i1}$. Assign the $r_{d,t}$ shifts to the "winners", the lowest $r_{d,t}$ officers. Calculate the shift auction injuries using the v_{it2} draws of the auction winners.

4. Compute the injury rate change as the difference in the number of injuries under the two systems divided by the total number of officer-work days.

3.9.8 The Value of a Statistical Injury

I use an approach similar to that observed in the literature (Kniesner and Viscusi 2019) and define the *value of a statistical injury* (VSI) as the amount of money an officer would be willing to pay to decrease the probability of injury on a work day by $1/n$ multiplied by n. I set n to be 259,861. This is the number of officer-days in my analysis population. Thus the VSI I present has the usual interpretation: it is the amount of money a large number

15. This code appears to correspond to regular hours, or non-overtime, hours.

of officers are willing to collectively pay to avoid one additional injury in the 609-day period.

In my setting, variation in wages allows us to back out the value of a statistical injury using a willingness to pay approach. Since unobserved injury risk is negatively correlated with utility and the coefficient on wages in utility is positive, the typical officer will require a positive payment to take on injury risk. The methodology I use to calculate the value of statistical injury is listed in Appendix Section 3.9.9. I estimate that on average, the implied value of a statistical injury for Los Angeles traffic officers is between \$125,445 and \$250,891.[16] These aggregate figures mask significant individual and temporal heterogeneity. Figure 3.12 displays a density plot of willingness to pay estimates across officer-days. The distribution is bimodal, with a peak near \$0.1 and another near \$0.5. This is a cautionary tale: even though the analysis is restricted to a single occupation in a specific city, willingness to pay for injury risk reduction varies greatly from person to person. My results also suggest that as working arrangements become more flexible and under the worker's control (through gig-economy growth and the transition to contractor employment), workplace injury should fall.

Table 3.5: Value of a Statistical Injury

Lower Bound (M = 1)		Upper Bound (M = 2)	
Willingness to Pay	VSI	Willingness to Pay	VSI
1.151	67195.5	2.301	134391.0
(1.971)	(115136.0)	(3.943)	(230272.0)

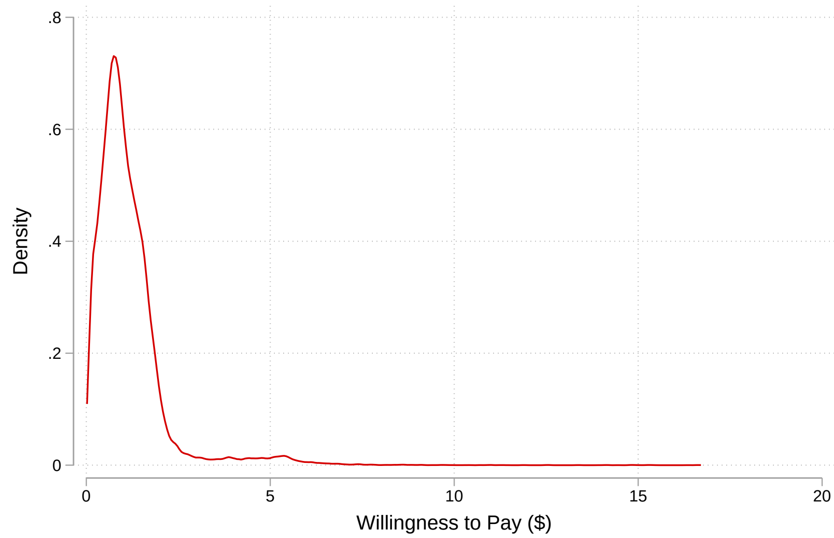
Note: This table displays the willingness to pay for an injury risk reduction, which is the average amount an officer who is indifferent between working and not would pay to reduce injury risk by $1/259,861$. The value of a statistical injury (VSI) is the willingness to pay multiplied by $259,861$.

Viscusi and Aldy 2003, which surveyed VSI estimates as of 2003, report developed country VSI estimates ranging from \$8,148 to \$242,671 (using year 2000 US dollars). Most of the estimates they report are between \$20,000 and \$50,000. My estimates adjusted to 2000 dollars[17] yield an estimated VSI range of \$90,606 to \$181,212. It is hard to compare

16. Dollars are as of 2015 and unadjusted for inflation.

17. using the U.S. Bureau of Labor Statistics' CPI Inflation Calculator.

Figure 3.12: Distribution of Willingness to Pay Across Officer-Days



Note: The figure plots the distribution of willingness to pay for a $1/259,861$ reduction in risk. The unit of observation is officer-day. The Epanechnikov kernel is used to estimate the density. Values above $2 (less than 3% of the data) are removed for better visualization.

VSI estimates, because they depend heavily on the severity of injuries faced as well as the risk tolerance of the population analyzed. Individuals sort into occupations partly based on risk tolerance. Therefore, because I analyze a specific occupation, my estimates are not representative of the average working population's value of a statistical injury.

There are several potential reasons why my estimates are higher than past estimates. First, the VSI estimates in the Viscusi and Aldy 2003 survey use the coefficients from hedonic wage regressions. This approach implicitly assumes that risk within occupations is exogenous. In the case of traffic officers at least, individuals can control their own risk through daily labor supply decisions. The fact that our VSI estimates are high relative to others suggests this endogeneity causes a downward bias. Second, a good portion of the injuries I analyze are severe and related to vehicle accidents. Such injuries have the potential to be fatal, and are much more likely to have long term consequences for quality of life.

3.9.9 Description of Value of Statistical Injury Calculations

For the purposes of these calculations, I assume that all officers are indifferent between working and not working *prior to the probability change.* Mathematically, this means that $\zeta_2 + x'\beta + \bar{Z}_i'\gamma_2 = v_{it1}$. Such officers are willing to accept an increase of $\alpha_w q$ in v_{it1} in exchange for a \$$q$ increase in the wage. This increase in v_{it1} translates into injury probability because it is correlated with v_{it2}. Thus an increase in v_{it1} (unobserved willingness to work) shifts the conditional distribution of v_{it2} (unobserved injury resistance). Specifically, it decreases the mean of injury resistance by $\rho\alpha_w q$. The proportional change in the probability of injury for an officer with covariates x_{it} and initial value of unobserved work utility v_{it1} is:

$$\Delta(x_{it}, q, v) := \Phi\left(\frac{\zeta_2 + x'\beta + \bar{Z}_i'\gamma_2 - \rho v + q(\beta_w - \rho\alpha_w)}{(1-\rho^2)^{1/2}}\right) - \Phi\left(\frac{\zeta_2 + x'\beta + \bar{Z}_i'\gamma_2 - \rho v}{(1-\rho^2)^{1/2}}\right)$$

The willingness to pay for a $1/n$ increase in injury probability for an officer with covariates x_{it} and unobserved resistance to work v is then given by $q(x_{it}, v)$ which solves:

$$\Delta(x_{it}, q(x_{it}, v), v) = \frac{1}{n}$$

This is uniquely defined because the CDF is strictly increasing. Solving for q (willingness to pay) yields:

$$q(x_{it}, v) = -\frac{1}{\beta_w - \rho\alpha_w}\left((\zeta_2 + x_{it}'\beta + \bar{Z}_i'\gamma_2 - \rho v) - (1-\rho^2)^{1/2}\Phi^{-1}\left\{\Phi\left(\frac{\zeta_2 + x_{it}'\beta + \bar{Z}_i'\gamma_2 - \rho v}{(1-\rho^2)^{1/2}}\right) + \frac{1}{n}\right\}\right)$$

To calculate VSI, I assume that officers expect to work 8 hours ex-ante. Finally, the value of a statistical injury is given by:

$$VSI = M \cdot n \cdot 8 \cdot \mathbb{E}_{x,v}[q(x, v)]$$

where note that I have integrated out v, the unobserved utility from work.[18] M represents a multiplier on the wage. For some shifts, officers will expect to be paid their typical wage rate, so $M = 1$. For others, officers may expect to be paid an overtime or special events premium, so $M = 1.5$ or $M = 2$. Because the coefficient on wage is positive, I can bound the VSI from above by setting $M = 2$ and below by setting $M = 1$. The upper and lower bounds of the average VSI (and the associated willingness to pay) for Los Angeles traffic officers are presented in Table 3.5.

3.9.10 Additional Tables

Table 3.6: Number of Unique Injuries

	Officer Count	Percent
0	366	66.18
1	134	24.23
2	39	7.05
3	12	2.17
4	1	0.18
5	1	0.18
Total	553	100.00

Note: Distribution of injuries across officers. Most officers experience no injuries or only one injury.

18. For my estimates, I integrate out v using Gauss-Hermite quadrature with 5 nodes.

Table 3.7: Types of Injuries

	Count	Percent
Strain or Injury By, NOC	53	20.95
Collision or Sideswipe w	40	15.81
Repetitive Motion - Other	24	9.49
Fall, Slip, Trip, NOC	18	7.11
Motor Vehicle, NOC	16	6.32
Other-Miscellaneous, NOC	12	4.74
Animal or Insect	10	3.95
Object Being Lifted or	8	3.16
Other Than Physical Cause	8	3.16
Fellow Worker, Patient, or	7	2.77
Person in Act of a Crime	7	2.77
Cumulative, NOC	5	1.98
Dust, Gases, Fumes or	5	1.98
Exposure, Absorption,	4	1.58
Twisting	4	1.58
Foreign Matter in Eye(s)	3	1.19
Struck or Injured, NOC	3	1.19
Using Tool or Machinery	3	1.19
Bicycling	2	0.79
Broken Glass	2	0.79
Lifting	2	0.79
Pushing or Pulling	2	0.79
Repetitive Motion	2	0.79
Temperature Extremes	2	0.79
Other (Catch-all)	11	4.40
Total	253	100.00

(a) Injuries by "Claim Cause"

	Count	Percent
Strain	119	47.04
Contusion	32	12.65
Sprain	30	11.86
Mental Stress	14	5.53
No Physical Injury	11	4.35
Inflammation	7	2.77
All Other Specific Inj.	5	1.98
Bee Sting	4	1.58
Dermatitis	4	1.58
Foreign Body	4	1.58
Heat Prostration	4	1.58
Multiple Physical Inj.	4	1.58
Carpal Tunnel	3	1.19
All Other Cumulative	2	0.79
Infection	2	0.79
Respiratory Disorders	2	0.79
Asbestosis	1	0.40
Bloodborne Pathogens	1	0.40
Hypertension	1	0.40
Laceration	1	0.40
Mult Injuries	1	0.40
Stroke	1	0.40
Total	253	100.00

(b) Injuries by "Nature of Injury"

Note: The table displays the distribution of injuries across two injury classification variables.

Table 3.8: Days Worked by Day of the Week

	Count	Percent	Cum. Pct.
Tuesday	32364	17.62	17.62
Wednesday	31548	17.18	34.80
Thursday	31329	17.06	51.86
Monday	30933	16.84	68.70
Friday	29757	16.20	84.90
Saturday	16478	8.97	93.87
Sunday	11250	6.13	100.00
Total	183659	100.00	

Note: This table describes the distribution of officer-days by day of the week.

Table 3.9: Number of Officers on Leave By Division

	mean	sd	p10	p50	p90
811					
Officers with Positive Leave	4.54	3.67	1.00	4.00	8.00
Officers with Positive Sick	1.57	1.45	0.00	1.00	4.00
Total Leave Hours	52.35	34.17	2.00	52.00	94.00
812					
Officers with Positive Leave	11.25	7.55	1.00	12.00	20.00
Officers with Positive Sick	3.54	2.79	0.00	3.00	7.00
Total Leave Hours	112.26	76.29	6.00	123.00	203.00
814					
Officers with Positive Leave	16.76	10.15	1.00	21.00	28.00
Officers with Positive Sick	5.59	3.61	0.00	6.00	10.00
Total Leave Hours	169.11	101.10	16.00	203.50	281.00
816					
Officers with Positive Leave	9.37	5.93	0.00	11.00	16.00
Officers with Positive Sick	2.40	2.04	0.00	2.00	5.00
Total Leave Hours	90.70	58.55	0.00	104.00	155.00
818					
Officers with Positive Leave	4.75	3.35	0.00	5.00	9.00
Officers with Positive Sick	1.49	1.39	0.00	1.00	3.00
Total Leave Hours	47.69	33.65	0.00	49.00	88.00
819					
Officers with Positive Leave	17.01	10.49	1.00	21.00	28.00
Officers with Positive Sick	5.79	3.79	1.00	6.00	10.00
Total Leave Hours	173.82	106.87	16.00	206.00	293.00
800 - 810, 824, 828,					
Officers with Positive Leave	1.48	1.42	0.00	1.00	3.00
Officers with Positive Sick	0.63	0.81	0.00	0.00	2.00
Total Leave Hours	16.14	15.82	0.00	16.00	40.00
Other					
Officers with Positive Leave	2.42	1.77	0.00	2.00	5.00
Officers with Positive Sick	0.68	0.84	0.00	0.00	2.00
Total Leave Hours	24.28	18.55	0.00	24.00	48.00
Total					
Officers with Positive Leave	8.45	8.66	0.00	5.00	23.00
Officers with Positive Sick	2.71	3.05	0.00	2.00	7.00
Total Leave Hours	85.79	86.87	0.00	52.00	227.00
Observations	4864				

Note: This table describes the distribution of the number of officers on leave by division. It gives a sense of how leave varies spatially (differences in the distribution across divisions) and temporally (variation within division across time). The category "Other" contains several small division codes.

Table 3.10: Regressions of Injury on Work

	(1)	(2)	(3)	(4)
Work	0.00140***	0.00141***	0.00113***	0.00116***
	(0.000100)	(0.000103)	(0.000106)	(0.000109)
Age	0.0000108	0.0000116	0.0000119	0.0000117
	(0.00000769)	(0.00000982)	(0.00000981)	(0.00000987)
Observations	259861	259861	259861	259861
F-Stat.	97.78	9.668	5.509	.
Division FE	No	Yes	Yes	Yes
Day of Week/Month FE	No	No	Yes	No
Date FE	No	No	No	Yes

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents results of ordinary least squares regressions of injury on work. The coefficient on work provides a naive estimate of the observed injury rate. Standard errors are clustered at the officer level.

Table 3.11: Linear Probability Models of Work Decision

	(1)	(2)	(3)	(4)	(5)
Coworkers on Leave	0.0191***	0.0192***	0.0218***	0.00148***	0.00380***
	(0.000474)	(0.000437)	(0.000461)	(0.000523)	(0.000669)
Age		0.000322	0.000312	0.000385	0.000366
		(0.000424)	(0.000290)	(0.000285)	(0.000283)
Wage		0.0577***	0.0448***	0.0281***	0.0267***
		(0.00487)	(0.00464)	(0.00357)	(0.00363)
Seniority Rank		-0.0000898	0.000160	0.000223	0.000221
		(0.000187)	(0.000164)	(0.000164)	(0.000163)
Observations	80898	80898	80898	80898	80898
First-Stage F.	583.7	390.8	114.0	112.1	99.45
Division FE	No	No	Yes	Yes	Yes
Month/Day of Week FE	No	No	No	Yes	No
Date FE	No	No	No	No	Yes

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents estimates of a linear probability model of the work decision. Time averages of age, leave of coworkers, seniority rank and wage are included in all specifications. The table suggests that the instruments are relevant to the work decision. The sample is limited to dates where an injury is observed. Standard errors are clustered at the officer level.

Table 3.12: Model Parameters with Sick Time Excluded

	Injury	Work
Avg. Leave of Coworkers	-0.0360***	0.00520
	(0.00972)	(0.00493)
Avg. Wage	-0.0694	-0.104***
	(0.0627)	(0.0182)
Age	0.00308	0.00173
	(0.00285)	(0.00137)
Wage	0.0870	0.111***
	(0.0652)	(0.0155)
Seniority Rank	0.00138	0.000949
	(0.00143)	(0.000794)
Leave of Coworkers		0.0150***
		(0.00264)
Observations	80898	
Rho	-0.653	
Rho 95% CI	(-0.18, -0.880)	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Parameter estimates when sick time is excluded from the leave instrument. $\hat{\rho}$ remains negative and significantly different from 0.

Table 3.13: Robustness Analyses

	Leave Coef.	Coef SE	Rho	Rho SE	Avg. Pop. Inj. Rate	Avg. Pop. Inj. Rate SE
Base Model	0.0150	0.0027	-0.6583	0.1722	0.0117	0.0069
Sick Time Excluded from Leave	0.0150	0.0026	-0.6534	0.1736	0.0115	0.0068
Broader Date FE	0.0116	0.0025	-0.5800	0.3364	0.0093	0.0090
Strains Not Considered Injuries	0.0152	0.0026	-0.5697	0.4996	0.0026	0.0046
Med Exp $\leq$0 Not Injury	0.0116	0.0025	-0.8013	0.1710	0.0194	0.0164
Med Exp $\leq$200 Not Injury	0.0116	0.0025	-0.7915	0.1752	0.0177	0.0152
Med Exp $\leq$400 Not Injury	0.0116	0.0025	-0.8264	0.1404	0.0191	0.0152

The table displays results of a number of robustness analyses. The first row provides the reference values from the primary specification. The second row removes sick time from the leave instrument. The third row (and all following rows) utilizes week and day of the week fixed effects rather than date fixed effects. The fourth row excludes strains as injuries. The fifth through sixth rows recode injuries with medical expenditures less than different amounts as non-injuries.

Table 3.14: Average Labor Supply Elasticities

Effect	Analytical Representation	Model Estimate
Leave of Coworkers	$\mathbb{E}_{z_{it}}\left[\frac{leave_{it}}{Pr(w_{it}=1\|z_{it})}\frac{\partial Pr(w_{it}=1\|z_{it})}{\partial leave_{it}}\right]$	.0327 (.00583)
Wage	$\mathbb{E}_{z_{it}}\left[\frac{wage_{it}}{Pr(w_{it}=1\|z_{it})}\frac{\partial Pr(w_{it}=1\|z_{it})}{\partial wage_{it}}\right]$	.6253 (.09263)
Seniority	$\mathbb{E}_{z_{it}}\left[\frac{senior_{it}}{Pr(w_{it}=1\|z_{it})}\frac{\partial Pr(w_{it}=1\|z_{it})}{\partial senior_{it}}\right]$	.0049 (.00425)

This table reports averages elasticities of the work outcome. Estimates are averages over all covariates and officer-days, with standard errors accounting for sampling of covariates. The values can be interpreted as a 1% increase in the variable changes the probability of working by x%.

Table 3.15: Average Elasticities: Injury Conditional on Working

Effect	Analytical Representation	Model Estimate
Wage	$\mathbb{E}_{z_{it}}\left[\frac{wage_{it}}{Pr(y_{it}=1\|w_{it}=1,z_{it})}\frac{\partial Pr(y_{it}=1\|w_{it}=1,z_{it})}{\partial wage_{it}}\right]$	12.37 (5.3440)
Leave of Coworkers	$\mathbb{E}_{z_{it}}\left[\frac{leave_{it}}{Pr(y_{it}=1\|w_{it}=1,z_{it})}\frac{\partial Pr(y_{it}=1\|w_{it}=1,z_{it})}{\partial leave_{it}}\right]$	.2347 (.12577)
Seniority	$\mathbb{E}_{z_{it}}\left[\frac{senior_{it}}{Pr(y_{it}=1\|w_{it}=1,z_{it})}\frac{\partial Pr(y_{it}=1\|w_{it}=1,z_{it})}{\partial senior_{it}}\right]$	.2165 (.13800)

This table reports averages elasticities of the injury outcome conditional on working. The elasticities are averages over all covariates and officer-days, with standard errors accounting for sampling of covariates. The values can be interpreted as a 1% increase in the variable changes the conditional probability of injury by x%.

Table 3.16: Variation Descriptions

Work	Leave	Other
ADJUSTMENT PERMANENT VARIATION IN RATE	100% SICK TIME (CREDIT OR CHARGE)	100% SICK TIME BALANCE PAID AT RETIREMENT
CURRENT ACTUAL HOURS WORKED ONLY	75% SICK TIME (CREDIT OR CHARGE)	50% SICK TIME BALANCE PAID AT RETIREMENT
DAY SHIFT HOURS WORKED	ABSENT WITHOUT PAY (POS OR NEG)	ADJUST VACATION EARNED BALANCE (+) OR (-)
HOLIDAY HOURS (CREDIT OR CHARGE)	ABSENT WITHOUT PAY - BANKED EXCESS SICK TIME	ADJUST VC MAX BALANCE (-) WAIVED
LIGHT DUTY RETURN TO WORK PROGRAM	ABSENT WITHOUT PAY - CPTO	BANKED EXCESS SICK TIME - PAID AT TERMINATION/RETIREMENT
NIGHT OR GRAVE PAY 5.5% NOT FOR SWORN	ABSENT WITHOUT PAY - FAMILY ILLNESS ¿ 40.0 HOURS	BANKED EXCESS SICK TIME - TIME OFF
OVERTIME (1.0) WORKED AND PAID	ABSENT WITHOUT PAY - FAMILY LEAVE-C CLASS	BIKE/WORK NON-TAX REIMBURSEMENT
OVERTIME (1.5) WORKED AND PAID	ABSENT WITHOUT PAY - FLOATING HOLIDAY	BIKE/WORK TAXABLE REIMBURSEMENT
OVERTIME WORKED (1.5)	ABSENT WITHOUT PAY - OVERTIME OFF 1.5	BONUS OR MARKSMANSHIP
OVERTIME WORKED (STRAIGHT)	ABSENT WITHOUT PAY - PREVENTIVE MEDICINE ¿ LIMIT	CALIFORNIA STATE TAX ADJUSTMENT (POS OR NEG)
PAID OVERTIME (HOLIDAY 1.5)	ABSENT WITHOUT PAY - SICK LEAVE	CASH-IN-LIEU PAYMENT
SEDENTARY DUTY	ABSENT WITHOUT PAY - VACATION	CATASTROPHIC TIME TRANSFERRED FROM BANK TO RECEIVING EMPLO
TEMPORARY VARIATION IN RATE - UP	ADDITIONAL BEREAVEMENT LEAVE OUT OF SICK TIME	CATASTROPHIC TIME USED BY CIVILIAN FROM CATASTROPHIC
	ADMINISTRATIVE LEAVE WITH PAY (POS OR NEG)	CPTO - CHANGE PERMANENT BALANCE + OR -
	BEREAVEMENT LEAVE (POS OR NEG)	CURR YR IOD CONVERSION ADJUSTMENT
	CPTO - COMPENSATED PERSONAL TIME OFF	ELECTRONIC PARKING SENSORS
	DECEASED EMPLOYEE / HOURS DID NOT WORK	FEDERAL TAX ADJUSTMENT (POS OR NEG)
	FAMILY ILLNESS (POS OR NEG)	FICA/MEDICARE YTD WAGE ADJUSTMENT (POS OR NEG)
	FML USING 1.0 BANKED OT	FLOATING HOLIDAY ACCRUED HOURS BALANCE (REPLACE)
	FML USING 1.5 BANKED OT	FLOATING HOLIDAY HOURS TAKEN THIS PAY PERIOD
	FML USING 100% SICK	Floating Holiday Lost
	FML USING 75% SICK	GROSS WAGE ADJUSTMENT
	FML USING FAMILY ILLNESS	NEW HIRE CODE / HOURS NO PAY IN INITIAL PAY PERIOD
	FML USING FLOATING HOLIDAY	OVERTIME (1.5) BALANCE PAID AT TERMINATION/RETIREMENT
	FML USING HOLIDAY	OVERTIME (STRAIGHT) BALANCE PAID AT TERMINATION/RETIREMENT
	FML USING VACATION	OVERTIME PAYMENT CONVERTED FROM OT (1.5)
	FML WITHOUT PAY	PMT OF EXES SICKLEAVE OVER 800 HRS AT 100% PAID AT 50
	JURY DUTY	PRIOR YR IOD CONVERSION ADJUSTMENT
	LEAVE WITH PAY (POS OR NEG)	PROFESSIONAL DEVELOPMENT STIPEND
	LEAVE WITHOUT PAY (POS OR NEG)	REDUCTION FROM TERMINATION PAYOUTS BAL OWED- CURR YR IOD CONV ADJ
	MILITARY LEAVE WITH PAY (POS OR NEG)	REDUCTION FROM TERMINATION PAYOUTS BAL OWED- PRIOR YR IOD CONV ADJ
	MILITARY LEAVE WITHOUT PAY (POS OR NEG)	REFUND DEDUCTION
	NET IOD (POS OR NEG)	SETTLEMENT
	OVERTIME TAKEN OFF (1.5)	SICK 100% ACCUMULATED
	OVERTIME TAKEN OFF (STRAIGHT)	SICK 100% CURRENT
	PREVENTIVE MEDICINE (POS OR NEG)	SICK 75% ACCUMULATED
	SUSPENSION (POS OR NEG) / HOURS NO PAY	SICK 75% CURRENT
	UNION NEGOTIATION TIME	STRAIGHT MONEY ADJUSTMENT OR EMPLOYEE EARNINGS (PO
	UNION RELEASE TIME	TERMINATION CODE / HOURS NO PAY
	VACATION (POS AND NEG)	TRANSIT BENEFIT ADJUSTMENT DOLLAR AMOUNT (NET PAY BENEFIT)
	WORKERS' COMPENSATION (POS OR NEG)	TRANSIT SPENDING SUBSIDY POSTTAX
		TRAVEL ALLOWANCE
		UNIFORM ALLOWANCE
		VACATION BALANCE PAID AT TERMINATION/RETIREMENT
		W2 MEDICAL SUBSIDY ADJUSTMENT
		YTD IMPUTED GROUP TERM LIFE - W2

The table lists the way each Variation Description is categorized. Variation Descriptions are pay codes describing the reason for payment. "Work" codes are used to construct hours worked and determine which days were worked. "Leave" codes are used to construct the leave instrument.

Table 3.17: Fixed Effects IV: Testing Instrument Validity

	(1)	(2)	(3)	(4)
Work	0.00263***	0.00233***	0.00740*	0.00265
	(0.000336)	(0.000299)	(0.00379)	(0.00224)
N	259861	259861	259861	259861
Underid K-P LM-stat	336.7	342.5	28.20	57.41
C-G F-Stat	60651.3	67321.7	497.8	1377.4
Weak id. K-P F-stat	3394.3	3470.9	29.66	67.40
Division FE	No	Yes	Yes	Yes
Day of Week/Month FE	No	No	Yes	No
Date FE	No	No	No	Yes

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The table displays estimates from a fixed effects instrumental variables regression. Work is instrumented with leave of coworkers, seniority and cumulative potential contacts. Column 4 is called the proxy model in the paper, as it denotes the model which would have been estimated if the outcome was continuous. Several weak instrument and overidentification tests are displayed under the coefficient estimates. Each column adds additional controls. Standard errors are clustered at the officer level.

Table 3.18: Balance Test: Regression of Medical Expenses Paid on Instruments

	(1)	(2)	(3)	(4)	(5)
Coworkers on Leave	-34.17	-34.17	-15.19	3.865	-112.8
	(33.82)	(33.82)	(51.04)	(49.36)	(136.4)
Age	34.85	34.85	24.94	12.21	110.6
	(33.77)	(33.77)	(26.41)	(27.25)	(128.2)
Wage	69.89	69.89	25.04	8.804	-62.47
	(107.4)	(107.4)	(116.4)	(123.3)	(330.8)
Seniority Rank	4.654	4.654	8.538	5.421	23.74
	(11.28)	(11.28)	(12.80)	(12.87)	(32.36)
Observations	257	257	257	257	257
First-Stage F.	0.447	0.447	.	.	.
Division FE	No	No	Yes	Yes	Yes
Month FE	No	No	No	Yes	No
Date FE	No	No	No	No	Yes

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table presents regressions of medical expenses on the instruments. Time averages of age, leave of coworkers, cumulative officer potential contacts, seniority rank and wage are included in all specifications. This is a balance test of the instruments, and if the exclusion restriction holds we would see no relationship between each variable and the outcome. The lack of significant coefficients is evidence in favor of the exclusion restriction. Each column adds additional controls. Standard errors are clustered at the officer level.

Table 3.19: Labor Supply Elasticities

Private Injury Risk Quantile	Elasticity
0.15	0.382
	(0.143)
0.30	0.642
	(0.133)
0.45	0.939
	(0.164)
0.60	1.319
	(0.337)
0.75	1.873
	(0.738)
0.90	2.928
	(1.312)

The table displays the average work probability (labor supply) elasticity conditional on different values of unobserved injury propensity. Labor supply becomes less elastic as injury propensity rises.

Bibliography

Adenbaum, Jacob. 2021. "Endogenous Firm Structure and Worker Specialization."

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

Aravindakshan, Ashwin, and Brian Ratchford. 2011. "Solving share equations in logit models using the LambertW function." *Review of Marketing Science* 9.

Argyres, Nicholas S., Teppo Felin, Nicolai Foss, and Todd Zenger. 2012. "Organizational Economics of Capability and Heterogeneity." *https://doi.org/10.1287/orsc.1120.0746* 23, no. 5 (October): 1213–1226.

Armstrong, Mark, and John Vickers. 2010. "A model of delegated project choice." *Econometrica* 78 (1): 213–244.

Ashton, Adam, and Phillip Reese. *Soaring overtime fattens paychecks of California cops and firefighters. But at a cost.*

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *The Quarterly Journal of Economics* 118, no. 4 (November): 1279–1333.

Baker, George, Robert Gibbons, and Kevin J. Murphy. 2002. "Relational Contracts and the Theory of the Firm." *The Quarterly Journal of Economics* 117, no. 1 (February): 39–84.

Baker, George P., and Thomas N. Hubbard. 2003. "Make Versus Buy in Trucking: Asset Ownership, Job Design, and Information." *American Economic Review* 93, no. 3 (June): 551–572.

Bargain, Olivier, Kristian Orsini, and Andreas Peichl. 2014. "Comparing labor supply elasticities in europe and the united states new results." *Journal of Human Resources* 49 (3): 723–838.

Becker, Gary S., and Kevin M. Murphy. 1992. "The division of labor, coordination costs, and knowledge." *Quarterly Journal of Economics* 107 (4): 1137–1160.

Belzunce, Felix, Carolina Martinez-Riquelme, Jose M Ruiz, and Miguel A Sordo. 2016. "On sufficient conditions for the comparison in the excess wealth order and spacings." *Journal of Applied Probability* 53 (1): 33–46.

Bencharit, Lucy Zhang, Yuen Wan Ho, Helene H Fung, Dannii Y Yeung, Nicole M Stephens, Rainer Romero-Canyas, and Jeanne L Tsai. 2019. "Should job applicants be excited or calm? The role of culture and ideal affect in employment settings." *Emotion* 19 (3): 377.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841–890.

Black, Ines, Sharique Hasan, and Rembrand Koning. 2020. "Hunting for talent: Firm-driven labor market search in America." *Available at SSRN.*

Blackwell, David. 1953. "Equivalent comparisons of experiments." *The annals of mathematical statistics,* 265–272.

Blahut, Richard E. 1972. "Computation of Channel Capacity and Rate-Distortion Functions." *IEEE Transactions on Information Theory,* no. 4.

Bloedel, Alexander, and Weijie Zhong. 2021. "The Cost of Optimally Acquired Information*." June.

Blundell, Richard, Antoine Bozio, and Guy Laroque. 2011. "Labor supply and the extensive margin." *American Economic Review* 101 (3): 482–486.

Bureau, US Census. 2021. *2018 SUSB Annual Data Tables by Establishment Industry,* May.

Caliendo, Lorenzo, Giordano Mion, Luca David Opromolla, and Esteban Rossi-Hansberg. 2020. “Productivity and organization in portuguese firms.” *Journal of Political Economy* 128, no. 11 (November): 4211–4257.

Caliendo, Lorenzo, Ferdinando Monte, and Esteban Rossi-Hansberg. 2015. “The Anatomy of French Production Hierarchies.” *The Journal of Political Economy* 123, no. 4 (July).

Caliendo, Lorenzo, Esteban Rossi-Hansberg, Lorenzo Caliendo, and Esteban Rossi-Hansberg. 2012. “The Impact of Trade on Organization and Productivity.” *The Quarterly Journal of Economics* 127, no. 3 (August): 1393–1467.

Caplin, Andrew, and Barry Nalebuff. 1991. “Aggregation and Imperfect Competition: On the Existence of Equilibrium.” *Econometrica* 59, no. 1 (January): 25.

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. 2018. “Firms and labor market inequality: Evidence and some theory.” *Journal of Labor Economics* 36, no. S1 (January): S13–S70.

Carlsson, Magnus, Luca Fumarco, and Dan-Olof Rooth. 2014. “Does the design of correspondence studies influence the measurement of discrimination?” *IZA Journal of Migration* 3 (1): 1–17.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. “The Effect of Minimum Wages on Low-Wage Jobs.” *The Quarterly Journal of Economics* 134, no. 3 (August): 1405–1454.

Charles, Kerwin K, Matthew S Johnson, Melvin Stephens Jr, and Do Q Lee. 2019. “Demand conditions and worker safety: Evidence from price shocks in mining.”

Chateauneuf, Alain, Michele Cohen, and Isaac Meilijson. 2004. “Four notions of mean-preserving increase in risk, risk attitudes and applications to the rank-dependent expected utility model.” *Journal of Mathematical Economics* 40 (5): 547–571.

Chausse, Pierre. 2021. *Package 'gmm'.*

Che, Yeon-Koo, Wouter Dessein, and Navin Kartik. 2013. “Pandering to persuade.” *American Economic Review* 103 (1): 47–79.

Chen, Po-Ning. n.d. “Chapter 6 Lossy Data Compression and Transmission.”

Chen, Songnian, Yahong Zhou, and Yuanyuan Ji. 2018. “Nonparametric identification and estimation of sample selection models under symmetry.” *Journal of Econometrics* 202 (2): 148–160.

Chetty, Raj. 2012. “Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply.” *Econometrica* 80 (3): 969–1018.

Conway, Sadie H, Lisa A Pompeii, David de Porras, Jack L Follis, and Robert E Roberts. 2017. “The identification of a threshold of long work hours for predicting elevated risks of adverse health outcomes.” *American Journal of Epidemiology* 186 (2): 173–183.

Cowgill, Bo, and Patryk Perkowski. 2020. “Delegation in Hiring: Evidence from a Two-Sided Audit.” *Columbia Business School Research Paper,* no. 898.

Crawford, Gregory S, Oleksandr Shcherbakov, and Matthew Shum. 1997. “Quality Overprovision in Cable Television Markets †.” *American Economic Review* 2019 (3).

Csaba, Dániel. 2021. “Attention Elasticities and Invariant Information Costs *.”

d’Aspremont, C., J. Jaskold Gabszewicz, and J.-F. Thisse. 1979. “On Hotelling’s ”Stability in Competition”.” *Econometrica* 47, no. 5 (September): 1145.

Dai, Tianjiao, and Juuso Toikka. 2022. “Robust Incentives for Teams.” *Econometrica* 90, no. 4 (July): 1583–1613.

Dembe, Allard E, J Bianca Erickson, Rachel G Delbos, and Steven M Banks. 2005. “The impact of overtime and long work hours on occupational injuries and illnesses: new evidence from the United States.” *Occupational and environmental medicine* 62 (9): 588–597.

Denti, Tommaso, Massimo Marinacci, and Luigi Montrucchio. 2020. "A note on rational inattention and rate distortion theory." *Decisions in Economics and Finance* 43, no. 1 (June): 75–89.

Dessein, Wouter, and Tano Santos. 2006. "Adaptive organizations." *Journal of Political Economy* 114, no. 5 (October): 956–995.

Du, Yu, and Ravi Varadhan. 2020. "SQUAREM: An R package for off-the-shelf acceleration of EM, MM and other EM-like monotone algorithms." *Journal of Statistical Software* 92:1–41.

Dufrane, Amy, Michael Bonarti, Jim DeLoach, and Rebecca DeCook. 2021. *The HR Function's Compliance Role,* June.

Ellickson, Paul B. 2007. "Does Sutton apply to supermarkets?" *The RAND Journal of Economics* 38, no. 1 (March): 43–59.

Frankel, Alex. 2021. "Selecting Applicants." *Econometrica* 89 (2): 615–645.

Frankel, Alexander. 2014. "Aligned delegation." *American Economic Review* 104 (1): 66–83.

Fredriksson, Peter, Lena Hensvik, and Oskar Nordström Skans. 2018. "Mismatch of talent: Evidence on match quality, entry wages, and job mobility." *American Economic Review* 108 (11): 3303–3338.

Freund, Lukas. 2022. "Superstar Teams: The Micro Origins and Macro Implications of Coworker Complementarities." *SSRN Electronic Journal* (December).

Galperin, Ron. 2015. *DOT Traffic Control for Special Events.*

Garicano, L. 2000. "Hierarchies and the organization of knowledge in production." *Journal of Political Economy* 108 (5): 874–904.

Garicano, Luis, and Thomas N. Hubbard. 2016. "The Returns to Knowledge Hierarchies." *The Journal of Law, Economics, and Organization* 32, no. 4 (November): 653–684.

Garicano, Luis, and Esteban Rossi-Hansberg. 2006. "Organization and Inequality in a Knowledge Economy."

Garicano, Luis, and Yanhui Wu. 2012. "Knowledge, communication, and organizational capabilities." *Organization Science* 23, no. 5 (February): 1382–1397.

Gentzkow, Matthew, and Emir Kamenica. 2016. "A Rothschild-Stiglitz approach to Bayesian persuasion." *American Economic Review* 106 (5): 597–601.

Gibbons, Robert. 2020. "March-ing toward organizational economics." *Industrial and Corporate Change* 29, no. 1 (February): 89–94.

Gregory, Terry, and Ulrich Zierahn. 2022. "When the minimum wage really bites hard: The negative spillover effect on high-skilled workers." *Journal of Public Economics* 206 (February): 104582.

Guardado, José R, and Nicolas R Ziebarth. 2019. "Worker investments in safety, workplace accidents, and compensating wage differentials." *International Economic Review* 60 (1): 133–155.

Haanwinckel, Daniel. 2020. "Supply, Demand, Institutions, and Firms: A Theory of Labor Market Sorting and the Wage Distribution."

Heart Disease Facts. 2020.

Heckman, James J. 1998. "Detecting discrimination." *Journal of economic perspectives* 12 (2): 101–116.

Heckman, James J, and Edward J Vytlacil. 1999. "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the national Academy of Sciences* 96 (8): 4730–4734.

Hirshman, Carolyn. 2018. *Incentives for Recruiters?,* April.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." *JL Econ. \& Org.* 7:24.

———. 1994. "The Firm as an Incentive System." *American Economic Review* 84 (4): 972–991.

Horrace, William C, and Ronald L Oaxaca. 2006. "Results on the bias and inconsistency of ordinary least squares for the linear probability model." *Economics Letters* 90 (3): 321–327.

How America Pays for College 2020.

Incidence rates for nonfatal occupational injuries and illnesses. 2020.

Johnson, Matthew S., and Michael Lipsitz. 2022. "Why Are Low-Wage Workers Signing Noncompete Agreements?" *Journal of Human Resources* 57, no. 3 (May): 689–724.

Jovanovic, Boyan. 1979. "Job matching and the theory of turnover." *Journal of political economy* 87 (5, Part 1): 972–990.

Jung, Junehyuk, Jeong Ho (John) Kim, Filip Matějka, and Christopher A. Sims. 2019. "Discrete Actions in Information-Constrained Decision Problems." *The Review of Economic Studies* 86, no. 6 (November): 2643–2667.

Kim, Woorim, Eun-Cheol Park, Tae-Hoon Lee, and Tae Hyun Kim. 2016. "Effect of working hours and precarious employment on depressive symptoms in South Korean employees: a longitudinal study." *Occupational and environmental medicine* 73 (12): 816–822.

Kniesner, Thomas J, and W Kip Viscusi. 2019. "The value of a statistical life." *Forthcoming, Oxford Research Encyclopedia of Economics and Finance,* 15–19.

Kremer, Michael. 1993. "The o-ring theory of economic development." *Quarterly Journal of Economics* 108 (3): 551–575.

Kugler, Maurice, and Eric Verhoogen. 2012. "Prices, Plant Size, and Product Quality." *The Review of Economic Studies* 79, no. 1 (January): 307–339.

Kuhn, Andreas, and Oliver Ruf. 2013. "The value of a statistical injury: New evidence from the Swiss labor market." *Swiss Journal of Economics and Statistics* 149 (1): 57–86.

Kuhn, Moritz, Jinfeng Luo, Iourii Manovskii, and Xincheng Qiu. 2022. "Coordinated Firm-Level Work Processes and Macroeconomic Resilience *."

Labor Statistics, U.S. Bureau of. 2019. *May 2018 National Occupational Employment and Wage Estimates,* April.

Lavetti, Kurt. 2020. "The estimation of compensating wage differentials: Lessons from the Deadliest Catch." *Journal of Business & Economic Statistics* 38 (1): 165–182.

Lazear, Edward P. 2009. "Firm-specific human capital: A skill-weights approach." *Journal of Political Economy* 117, no. 5 (October): 914–940.

Lee, David L, Justin McCrary, Marcelo J Moreira, and Jack Porter. 2020. "Valid t-ratio Inference for IV." *arXiv preprint arXiv:2010.05058.*

Lee, Jonathan M, and Laura O Taylor. 2019. "Randomized safety inspections and risk exposure on the job: Quasi-experimental estimates of the value of a statistical life." *American Economic Journal: Economic Policy* 11 (4): 350–74.

Li, Fei, and Can Tian. 2013. "Directed search and job rotation." *Journal of Economic Theory* 148, no. 3 (May): 1268–1281.

Liebman, Jeffrey B, Erzo F P Luttmer, and David G Seif. 2009. "Labor supply responses to marginal Social Security benefits: Evidence from discontinuities." *Journal of Public Economics* 93 (11-12): 1208–1223.

Lindenlaub, Ilse. 2017. "Sorting Multidimensional Types: Theory and Application." *The Review of Economic Studies* 84, no. 2 (April): 718–789.

Lipnowski, Elliot, and Doron Ravid. 2022. "Predicting Choice from Information Costs" (May).

Lucas, Robert E. 1978. "On the Size Distribution of Business Firms." *The Bell Journal of Economics* 9 (2): 508.

Martinez, Elizabeth A., Nancy Beaulieu, Robert Gibbons, Peter Pronovost, and Thomas Wang. 2015. "Organizational Culture and Performance." *American Economic Review* 105, no. 5 (May): 331–35.

Matêjka, Filip, and Alisdair McKay. 2015. "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." *American Economic Review* 105, no. 1 (January): 272–98.

McCall, John Joseph. 1970. "Economics of information and job search." *The Quarterly Journal of Economics,* 113–126.

McFadden, D. 1973. "Conditional logit analysis of qualitative choice behavior."

Meier, Stephan, Matthew Stephenson, and Patryk Perkowski. 2019. "Culture of trust and division of labor in nonhierarchical teams." *Strategic Management Journal* 40, no. 8 (August): 1171–1193.

Moran, Molly J, and Carlos Monje. 2016. "Guidance on treatment of the economic value of a statistical life (vsl) in us department of transportation analyses-2016 adjustment." *US Department of Transportation.*

Muro, Mark, Alan Berube, and Jacob Whiton. 2018. *Black and Hispanic underrepresentation in tech: It's time to change the equation,* April.

New York City Department of Finance. 2022. *New York State Sales and Use Tax,* May.

Nocke, Volker, and Nicolas Schutz. 2018. "Multiproduct-Firm Oligopoly: An Aggregative Games Approach." *Econometrica* 86, no. 2 (March): 523–557.

Ocampo, Sergio. 2022. “A Task-Based Theory of Occupations with Multidimensional Heterogeneity *.”

Parada-Contzen, Marcela, Andrés Riquelme-Won, and Felipe Vasquez-Lavin. 2013. “The value of a statistical life in Chile.” *Empirical Economics* 45 (3): 1073–1087.

Pomatto, Luciano, Philipp Strack, and Omer Tamuz. 2022. “The Cost of Information.”

Postel–Vinay, Fabien, and Jean–Marc Robin. 2002. “Equilibrium wage dispersion with worker and employer heterogeneity.” *Econometrica* 70 (6): 2295–2350.

Rosen, Sherwin. 1982. “Authority, Control, and the Distribution of Earnings.” *Bell Journal of Economics* 13 (2): 311–323.

Sattinger, Michael. 1975. “Comparative Advantage and the Distributions of Earnings and Abilities.” *Econometrica* 43, no. 3 (May): 455.

Semykina, Anastasia, and Jeffrey M Wooldridge. 2018. “Binary response panel data models with sample selection and self-selection.” *Journal of Applied Econometrics* 33 (2): 179–197.

Shaked, Avner, and John Sutton. 1987. “Product Differentiation and Industrial Structure.” *The Journal of Industrial Economics* 36, no. 2 (December): 131.

Shaked, Moshe, and J George Shanthikumar. 2007. *Stochastic orders.* Springer Science & Business Media.

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27 (3): 379–423.

Shimer, Robert, and Lones Smith. 2000. “Assortative matching and search.” *Econometrica* 68 (2): 343–369.

Shkel, Yanina, and Sergio Verdú. 2018. “A Coding Theorem for f-Separable Distortion Measures.” *Entropy* 20, no. 2 (February).

Song, Jae, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter. 2019. "Firming Up Inequality." *The Quarterly Journal of Economics* 134, no. 1 (February): 1–50.

Steinbach, Alison. 2019. *Police, firefighters earn most overtime among public employees, at times doubling salaries.*

Stock, James H, and Motohiro Yogo. 2002. *Testing for weak instruments in linear IV regression.* Technical report. National Bureau of Economic Research.

Teulings, Coen N. 2000. "Aggregation Bias in Elasticities of Substitution and the Minimum Wage Paradox." *International Economic Review* 41, no. 2 (May): 359–398.

Tian, Jianrong. 2019. "Attention Costs without Distraction and Entropy."

Tirole, Jean. 2009. "Cognition and Incomplete Contracts." *American Economic Review* 99, no. 1 (March): 265–94.

Tishby, Naftali, Fernando C Pereira, and William Bialek. 2000. "The information bottleneck method."

Ulbricht, Robert. 2016. "Optimal delegated search with adverse selection and moral hazard." *Theoretical Economics* 11 (1): 253–278.

Viscusi, W Kip, and Joseph E Aldy. 2003. "The value of a statistical life: a critical review of market estimates throughout the world." *Journal of risk and uncertainty* 27 (1): 5–76.

Viscusi, W Kip, and Joni Hersch. 2001. "Cigarette smokers as job risk takers." *Review of Economics and Statistics* 83 (2): 269–280.

Williams, Alex. 2020. *Our Ultimate Guide to the Best Data Science Bootcamps.*

Williamson, Ann, Rena Friswell, Jake Olivier, and Raphael Grzebieta. 2014. "Are drivers aware of sleepiness and increasing crash risk while driving?" *Accident Analysis & Prevention* 70:225–234.

Williamson, Oliver. 1984. "The Incentive Limits of Firms: A Comparative Institutional Assessment of Bureaucracy," 736–763.

Zhou, Jidong. 2020. "Improved Information in Search Markets."

Zhou, Xiang, and Yu Xie. 2019. "Marginal treatment effects from a propensity score perspective." *Journal of Political Economy* 127 (6): 3070–3084.