Essays in Labor and Finance

by

Bryan Seegmiller

B.S., Brigham Young University (2016)
S.M., Massachusetts Institute of Technology (2020)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author ....
Department of Management
April 28, 2022

Certified by ....
Leonid Kogan
Nippon Telegraph & Telephone Professor of Management
Professor of Finance
Thesis Supervisor

Accepted by ....
Catherine Tucker
Sloan Distinguished Professor of Management
Professor of Marketing
Chair, MIT Sloan PhD Program

Essays in Labor and Finance

by

Bryan Seegmiller

Submitted to the Department of Management
on April 28, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In chapter 1, I quantify the economic value that firms of different productivity levels derive from their labor market power by estimating the effect of unanticipated firm-level labor demand shocks on wages and employment at publicly listed U.S. firms. Productive firms face lower labor supply elasticities on average, and still lower elasticities for skilled workers, who are disproportionately employed at more productive firms. Using a dynamic wage posting model in which firms face upward-sloping labor supply and adjustment costs in hiring, I estimate that firms in the top and bottom quartiles of labor productivity pay 62% and 94% of marginal product, despite the fact that adjustment costs temper the exercise of labor market power. Markdown differentials can explain three-fifths of the average spread in log labor shares between high- and low-labor productivity firms, and the evolution of these differentials can explain most of the change in the aggregate labor share in the 1991–2014 period. Holding constant equilibrium labor demand, I estimate that about a third of capital income for the typical firm stems from wage markdowns. Aggregate wage markdowns are worth two-fifths of total capital income.

In chapter 2, joint work with Leonid Kogan, Dimitris Papanikolaou, and Larry Schmidt, we construct new technology indicators using textual analysis of patent documents and occupation task descriptions that span almost two centuries (1850–2010). At the industry level, improvements in technology are associated with higher labor productivity but a decline in the labor share. Exploiting variation in the extent certain technologies are related to specific occupations, we show that technological innovation has been largely associated with worse labor market outcomes—wages and employment—for incumbent workers in related occupations using a combination of public-use and confidential administrative data. Panel data on individual worker earnings reveal that less educated, older, and more highly-paid workers experience significantly greater declines in average earnings and earnings risk following related technological advances. We reconcile these facts with the standard view of technology-skill complementarity using a model that allows for skill displacement.

In chapter 3, I show that stocks with similar characteristics but different levels of ownership by financial institutions have returns and risk premia that comove very differently with shocks to the risk-bearing capacity of financial intermediaries. After accounting for observable stock characteristics, excess returns on more intermediated stocks have higher betas on contemporaneous shocks to intermediary willingness to take risk and are more predictable by state variables that proxy for intermediary health. The empirical evidence supports the predictions of asset pricing models featuring financial intermediaries as marginal investors who face frictions that induce changes in their risk-bearing capacity. This suggests that such models are useful for explaining price movements not only in markets for complex financial assets, but also within asset classes where households face comparatively low barriers to direct participation.

Thesis Supervisor: Leonid Kogan
Title: Nippon Telegraph & Telephone Professor of Management
Professor of Finance

Acknowledgments

I first wish to express most heartfelt thanks to my wife Hayley, and daughters June and Emma for bringing joy to my life and making the grind of the PhD worthwhile. Next, I want to thank my Savior Jesus Christ for being my Source of redemption and healing. My faith in Him helps me see the beauty in the world and the light in the people around me, which stands as a stark contrast to a profession that conditions one to be endlessly cynical and skeptical. And of course I would be remiss if I didn't thank my parents and siblings for motivating me, teaching me proper values, and instilling in me a desire to be better.

I also thank my advisors Leonid Kogan, David Autor, Larry Schmidt, and Dimitris Papanikoloau for their unfailing guidance and continued support. Without their many hours spent supporting me I would have never completed a PhD. Thanks are also due to Daron Acemoglu, Hui Chen, Taha Choukhmane, Maryam Farboodi, Dan Greenwald, Simon Jager, Eben Lazarus, Debbie Lucas, Chris Palmer, Jonathan Parker, Antoinette Schoar, Anna Stansbury, David Thesmar, Adrien Verdelhan, Emil Verner, Haoxiang Zhu, and Luigi Zingales for helpful comments on various parts of this thesis.

Disclaimer: The U.S. Census Bureau reviewed this data product for unauthorized disclosure of confidential information and approved the disclosure avoidance practices applied to this release (approval numbers CBDRB-FY21-163, CBDRB-FY21-POP001-0178, and CBDRB-FY21-POP001-0232).

Contents

Chapter 1

Valuing Labor Market Power: The Role of Productivity Advantages

Market power affects the distribution of firm cash flows between the various claimants on its output; in particular, increased market power is typically associated with a shift in the distribution of productive output away from the firm's workforce and towards owners of the firm's capital. A number of empirical patterns that have emerged over the past several decades suggest that US firms have been earning higher rents due to increased market power. This period of time has been characterized by a large rise in the value of the aggregate stock market and high firm profitability;[1] high firm valuations but with low rates of investment;[2] and, a decline in the share of output accruing to labor.[3] Much of the literature documenting these trends has focused on market power in the product market. In this paper I argue that a "superstar firms" view of labor market power—where firms with high labor productivity also face less labor market competition—can explain a significant share of the cross-sectional differences in firm labor shares and profitability, and can also speak to time series changes in the aggregate labor share. I test this view of labor market competition using a sample of firms containing wide variation in productivity and which features the largest and most

[1] See Greenwald, Lettau, and Ludvigson (2021) and Barkai (2020).

[2] Gutierrez and Philippon (2017)

[3] Karabarbounis and Neiman (2013)

highly productive firms in the world: publicly traded US corporations.

In support of the superstar firms view, I show empirically that labor productivity advantages confer a substantial competitive advantage in the labor market. In contrast with local labor market concentration, which has trended downward,[4] I find empirical patterns that parallel the recent shifts in the competitive structure of the US economy. An expanding productivity gap between the most- and least-productive firms has led to both an absolute and relative increase in productive firms' labor market power, resulting in a larger cross-sectional spread in labor shares as well as a lower aggregate labor share. Thus labor market power emerges as another factor in explaining recent macroeconomic trends in the competitive landscape. Furthermore, because labor compensation is the single largest cost of production for most firms, wage markdowns represent a substantial fraction of the capital income generated by US corporations.

I show all this in three steps. First, I motivate my choice to focus on productivity differentials by presenting evidence that firms with high labor productivity earn high economic rents relative to less productive firms. Next, I show empirically that labor market power is likely to play an important part in generating the rents earned by productive firms. Finally, I use a model combined with my empirical findings to value the cashflows that firms derive from their ability to set wages.

To show evidence suggesting that productive firms earn economic rents, I examine their labor shares, operating performance, financial valuation ratios, and investment rates, finding that productive firms have much lower labor shares; higher return on assets/equity, Tobin's Q, market-to-book, and market-to-sales ratios; but do not have significantly larger investment rates. Taken together, these features are consistent with productive firms earning higher economic rents than unproductive firms.

Because the supply elasticity governs wage markdowns, I next introduce a method for estimating supply elasticities that is applicable to a wide range of firms spanning multiple decades. The method provides enough statistical power to allow for heterogeneous cross-

[4] Berger, Herkenhoff, and Mongey (2021), Rinz (2018), and Lipsius (2018) all find declining local labor market concentration over the past several decades using Census administrative data.

sectional and time series estimates, and accounts for the key sources of endogeneity that would cause me to make erroneously large inferences regarding the magnitude of labor market power. Specifically, I use the ratio of the firm-level employment and wage responses to stock returns as a measure of the supply elasticity, controlling for industry-by-year fixed effects and other firm level covariates. This quantity directly estimates the supply elasticity under certain assumptions, and provides an upper bound on the elasticity under a broader set of conditions, which causes my estimates to be a lower bound on the magnitude of labor market power. The most important potential source of confounding variation comes from common market level productivity shocks, which could bias my supply elasticity estimates downward. Accordingly, I show that my baseline estimates are insensitive to a wide variety of additional controls for common shocks, implying that my estimates are driven by the firm-specific component of stock returns. I find that my baseline average supply elasticity estimate is well within the range found in the literature and is highly robust to different sets of controls. I also find that alternate labor demand shock proxies—such as using the stock returns of firms' customers, patent grants, or only including stock returns near quarterly earnings announcements—yield supply elasticity estimates that are nearly identical to my baseline specification, though with less statistical precision.

Applying my method to firms sorted on labor productivity, I show that highly productive firms face much lower labor supply elasticities than less productive firms. In models of monopsony, wages move further below marginal product as the firm-specific supply elasticity decreases, so this suggests that productive firms enjoy more monopsony power, a key source of their economic rents. Decomposing wages into worker- and firm-specific heterogeneity to back out worker skill, I find that productive firms also face lower supply elasticities for workers of all skill levels; the most skilled workers have the lowest supply elasticities of all, implying that labor market power is increasing in worker skill. In the time series I find that labor supply elasticities have decreased overall and for each worker skill level, and especially so for skilled workers. The spread in supply elasticities between productive and unproductive firms has been widening over time, leading to an increased gap in firm-level labor shares between top and bottom firms.

The decline in my estimated supply elasticities matches the secular shift towards increased rents and falling labor shares. The key role that skilled workers play in these phenomena can possibly help explain why supply elasticity estimates trend downward, even as local labor market concentration has decreased. For skilled workers, labor markets are much less local (see Malamud and Wozniak, 2012; Amior, 2020, for example), diminishing the importance of concentration defined over narrow geographic boundaries. Instead, I suggest that the increased importance of human capital specificity may be a key contributor. Consistent with this notion, the wage premium for incumbent workers has increased and worker separations rates have decreased, suggesting that it has both become more costly to replace an incumbent worker and more difficult for an incumbent worker to leave the firm.

The empirical incumbent wage premium suggests that it may be costly to replace workers within the firm (Kline, Petkova, Williams, and Zidar, 2019). Prior research also uncovers direct evidence that firms find it costly to adjust their labor force.[5] I further find that firm hiring decisions respond to a labor force analogue of Tobin's Q, also suggesting labor is costly to adjust by analogy to a long literature in finance on investment with costly capital adjustment. In the presence of labor adjustment costs the supply elasticity may not be a sufficient statistic for wage markdowns from marginal product, and so quantifying the extent of labor market power also requires accounting for the impact of a labor force that is costly to adjust. Accordingly, I quantify my empirical findings in the context of a dynamic wage posting monopsony framework, which is based on the static models of Kline et al. (2019) and Card, Cardoso, Heining, and Kline (2018) and features costly hiring of workers from outside the firm. I use the model to estimate an adjustment cost factor that allows me to convert empirically estimated supply elasticities into markdowns from marginal product.

In the model firms face upward sloping labor supply curves for both incumbent workers and potential recruits. Firms may offer different wages to these two groups because of

[5] In the labor literatureJager and Heining (2019) show that incumbent workers are costly to replace; in asset pricing, Belo, Lin, and Bazdresch (2014) argue that labor adjustment costs can explain the empirical relationship between firm hiring and expected stock returns.

adjustment costs in hiring workers from outside the firm. These adjustment costs could stem from the need to train new hires to utilize the firm's production technology and/or the costliness of recruiting outside workers. I calibrate the model to match empirical labor supply elasticity estimates for incumbents and recruits; average worker separations rates; and, the average incumbent wage premium. The model is also able to generate the empirical pattern of monotonically decreasing supply elasticities and labor shares as firm productivity increases, without explicitly targeting these moments. My main calibration matches data moments for the full sample period from 1991-2014; I also separately calibrate the model for the 1991-2002 and 2003-2014 subperiods. I then use the model along with my empirical estimates to quantify the value of cashflows that firms derive from paying workers less than their marginal product.

Combining my empirical results and the model calibration, my main quantitative findings are the following. The full-sample calibration of the dynamic model with adjustment costs shows that wage markdowns are about 16% smaller relative to a static model without costly adjustment. However, the dollar value of wage markdowns still represent a substantial portion of firm cash flows. I find that the average firm pays workers about 83% of their marginal product, with firms in the top productivity quartile paying about 62%, and the least productive firms paying 94%. These wage markdowns in turn represent a meaningful portion of firm cash flows. I perform a counterfactual exercise where I reverse wage markdowns to remunerate back to workers their marginal products, holding constant firms' production decisions. Overall, wage markdowns are worth about a third of the average firm's operating income. This figure is 17% for firms in the bottom quartile of the labor productivity distribution and 43% for firms in the top quartile. Aggregating the dollar value of wage markdowns across firms, wage markdowns are worth on average about two-fifths of aggregate operating income.

I further show that wage markdowns can explain about three-fifths of the gap in average labor shares between the firms in the top and bottom quartiles of the labor productivity distribution; meanwhile, my supply elasticity estimates and model calibrations for the 1991-2002 and 2003-2014 subperiods imply that changes in markdowns can explain the majority of the observed decline in the labor share over that time period.

The implication that productive firms exert more labor market power despite paying high wages to their employees appears counterintuitive. But productive firms' wages are low relative to a very high marginal revenue product of labor, not low in the absolute sense, which is entirely consistent with the empirical fact that productive firms also have low labor shares. It also suggests the need for some nuance when considering the welfare implications of this type of labor market power. For example, is labor is inelastic because of valuable firm-specific human capital, then breaking up productive firms to reduce labor market power could have negative consequences for both employees and employers if it devalues the workers' human capital in some way. On the other hand, the welfare implications are more clearly negative if firms generate some of their labor market power by using their resources to artificially curb worker mobility. In section 1.6 I show evidence that both forces are likely to be important.

I organize the main body of the paper as follows. In Section, 3.2 I discuss the data sources and basic summary statistics; in Section 1.2 I examine stylized facts that are consistent with productive firms earning economic rents; I explain my method for estimating supply elasticities and apply it to firms sorted on labor productivity in Section 1.3; in Section 3.1 I then introduce and calibrate a dynamic wage posting model with labor adjustment costs; in Section 1.5 I combine my model and empirical estimates to quantify the contribution of labor market power to firm capital income and its impact on aggregate and firm-level labor shares; in Section 1.6, I discuss economic and policy implications and alternative explanations for my findings. Finally, Section 3.5 concludes.

Related Literature

This paper contributes simultaneously to several different literatures. First of all, this paper adds to a growing literature in macro-finance which examines recent macroeconomic trends in competition, firm valuations/operating performance, and labor shares.[6] Most related to my paper is Hartman-Glaser, Lustig, and Xioalan (2019), who document that large firms have

[6] These include Barkai (2020), Grullon, Larkin, and Michaely (2019), Covarrubias, Gutiérrez, and Philippon (2020), Greenwald et al. (2021), Corhay, Kung, and Schmid (2020), and Farhi and Gourio (2018).

obtained an increasing share of their sales as capital income over the past several decades and propose an explanation based on firms providing insurance to their workers against increased risk. I similarly find that labor shares have decreased by more among highly productive firms. I argue that this is in large part driven by productive firms marking down wages further from marginal product. I also provide the novel finding that firms generate a substantial portion of their operating income from wage markdowns, which is especially the case for firms with high labor productivity. This competitive advantage generates rents which are reflected in higher financial valuation ratios and relatively low investment rates among firms with high labor productivity.

A parallel area of research in macroeconomics connects the secular decline in labor shares with changes in market competition. I differ from most of this literature in primarily focusing on the role of labor market power instead of product market power.[7] While some papers in this literature have focused on trends in productivity dispersion as a driving force behind changes in product market competition, my findings imply that increases in the labor market power of the most productive firms have also played a crucial role in my sample of publicly traded companies.

This paper also builds on recent research in labor economics. A number of prior papers estimate the elasticity of supply to the firm to infer the extent of monopsony power;[8]Another related empirical literature has examined how labor market concentration affects wages.[9] My paper differs from others in this literature by quantifying the impact that labor adjustment costs have on wage markdowns. I also estimate heterogeneity in labor supply elasticities based on firm characteristics, focusing on labor productivity in particular. Another large literature estimates the passthrough of firm-specific shocks to worker outcomes.[10] These

[7] See Autor, Dorn, Katz, Patterson, and Van Reenen (2020), De Loecker, Eeckhout, and Unger (2020), Covarrubias et al. (2020) and Kehrig and Vincent (2021) for examples. One exception is Stansbury and Summers (2020), who argue that worker bargaining power has played an important role.

[8] See Lamadon, Mogstad, and Setzler (2019), Berger et al. (2021), Kroft, Luo, Mogstad, and Setzler (2020), Bassier, Dube, and Naidu (2020), and Ransom and Sims (2010) for a few examples.

[9] See Rinz (2018), Benmelech, Bergman, and Kim (2018), Schubert, Stansbury, and Taska (2020), and Jarosch, Nimczik, and Sorkin (2019), for example.

[10] See Abowd and Lemieux (1993), Van Reenen (1996), Guiso, Pistaferri, and Schivardi (2005), Kline et al. (2019), Kogan, Papanikolaou, Schmidt, and Song (2020), Chan, Salgado, and Xu (2021), Garin and Silverio (2020), Friedrich, Laun, Meghir, and Pistaferri (2019), Balke and Lamadon (2020).

papers typically analyze the response of worker pay to firm-specific shocks, whereas my focus is on estimating the ratio of the employment and wage responses to stock return shocks in order to estimate the supply elasticity.

In a related prior paper, Gouin-Bonenfant (2020) argues theoretically for labor productivity dispersion as a determinant of labor market power in a search and matching model, and he shows empirically that productivity dispersion depresses labor shares at the aggregated industry level in Canada. I find that productivity advantages also reduce labor shares at the firm level in the United States, and this is at least partly direct result of differences in monopsony power. Thus my findings complement those of Gouin-Bonenfant (2020). My paper is also the first, to my knowledge, to explicitly estimate heterogeneous supply elasticities for firms of different productivity levels both cross-sectionally and across time.

Finally, this paper adds to a growing body of work in financial economics that examines the interplay between labor markets and firm financial outcomes and decision making.[11] My findings suggest that accounting for imperfect competition in labor markets may be of first-order importance for future research in this area.

1.1 Data and Summary Statistics

My analysis relies primarily on two data sources. The first is employer–employee matched wage data from the US Census Bureau's Longitudinal Employer Household Dynamics Database (LEHD). I link the firm identifiers in the LEHD to financial information in the CRSP/Compustat merged database obtained from Wharton Research Data Services. I describe the LEHD and CRSP/Compustat–LEHD merged data and samples below. Besides stock return and market cap data, which are from CRSP, all financial variables are from Compustat. Henceforth I refer to the CRSP/Compustat merged sample as the Compustat sample.

[11] References in asset pricing include Eisfeldt and Papanikoloau (2013), Belo et al. (2014), Donangelo, Gourio, Kehrig, and Palacios (2019), Kuehn, Simutin, and Wang (2017), Liu (2019); see Matsa (2010), Jeffers (2019), Mueller, Ouimet, and Simintzi (2017), Kim (2020), and Shen (2021) for examples in corporate finance.

1.1.1 Employee-Employer Matched Wage Data from the LEHD

The Longitudinal Employer Household Dynamics Database (LEHD) contains restricted-use microdata with wage and employer information for individuals in the United States. Wage data in the LEHD are collected from firms' unemployment insurance filings, and they contain all forms of compensation that are immediately taxable, including stock options. Individuals in the LEHD are linked to their employers through their State Employer Identification Number (SEIN). The LEHD provides crosswalks between the SEIN and the federal Employer Identification Number (EIN), which is also available for firm-level data sources such as Compustat. The LEHD data begin in 1990, although most states join later as the LEHD coverage becomes more comprehensive; the LEHD covers the majority of jobs in the United States by the mid- to late-1990s, and coverage ends in 2015.[12] The wage data are reported on a quarterly basis, and cover nearly 100 percent of private employees in state-quarters where the data are available. See Abowd, Stephens, Vilhuber, Andersson, McKinney, Roemer, and Woodcock (2009) for a more detailed overview of the construction of the LEHD.

1.1.2 Matched Compustat-LEHD Sample

I link firm identifiers in each year of the the LEHD to Compustat records using a crosswalk created by Larry Schmidt.[13] I then assign individuals to their corresponding Compustat gvkey and retain worker-years in which at least one of the worker's employers was linked to a Compustat firm. Following Sorkin (2018), I convert quarterly LEHD wages to their full-year equivalents. Adjusting wages for a given year following the Sorkin (2018) procedure requires information on wages in the year before and the year after the current year. See appendix section 1.7.1 for more details. Given the availability of the LEHD wage data from 1990-2015, this means my effective sample period spans 1991-2014. In order to be in the sample, I require a Compustat firm to have at least 15 workers for whom that firm is their primary employer (defined to be the firm where the worker earned the most income that

[12] An overview of LEHD coverage is available at https://www2.vrdc.cornell.edu/news/data/qwi-public-use-data/.

[13] Thanks to Larry for sharing his crosswalk code.

year). I additionally exclude financial firms and regulated utilities from my analysis following common practice. Because I use NAICS industry codes throughout, this excludes the 2-digit NAICS codes 22, 52, and 53. In appendix Figure 1-8 I show the shares of employment, market cap, and sales represented in my LEHD-Compustat matched sample. On average, my matched sample covers about 62% of employment, 63% of market cap, and 50% of sales represented in Compustat in a given year. I compare the distribution of firm characteristics for my matched sample and the overall Compustat database in appendix Table 1.10. Not surprisingly, firms in my matched sample skew a little bit larger relative to the mean or median firm in Compustat based on various measures of firm size such as assets, employment, or market cap. The distributions of annual excess stock returns are about the same across the two samples. I obtain the risk-free rate from Ken French's data library in order to calculate excess returns.

1.1.3 Variable Construction

My proxy for labor productivity is given by log value-added per worker, following a large literature in labor economics (see Card et al., 2018, for a review). Firms can have higher labor productivity by enjoying higher total-factor productivity, being more capital intensive, or by hiring more skilled workers. I define value-added for Compustat firms following Donangelo et al. (2019), as the sum of operating income before depreciation, changes in inventories, and labor expenses. Firm level labor shares are the ratio of labor expenses to value added. In order to estimate the skill level of individual workers (or the average skill of a given firm), I follow a long literature starting with Abowd, Kramarz, and Margolis (1999)—from now on AKM—in decomposing observed wages into worker- and firm-specific heterogeneity. I start with a modification of the AKM decomposition proposed by Lachowska, Mas, Saggio, and Woodbury (2020) and Engbom and Moser (2020) that allows for the firm-specific component of wages to vary by time. I use LEHD data to create measures of firm wage offers and total labor expenses. Details on my construction of all these variables can be found in appendix section 1.7.1.

1.2 Firm Labor Productivity and Labor Shares, Valuations, and Investment

In this section I document the following facts: productive firms 1) have lower labor shares; 2) have higher operating performance and valuation ratios; 3) but do not have high rates of investment. These facts in tandem are consistent with productive firms earning higher economic rents, and motivates my choice later on to examine supply elasticities for firms sorted on labor productivity. I emphasize that the evidence in this section is suggestive and correlational rather than causal, and do not by themselves constitute direct evidence for labor market power. Direct evidence on the magnitude of rents earned from labor market power instead come from when I estimate the labor supply elasticity to the firm in section 1.3. I focus on valuation ratios and labor shares because rents earned from market power should be associated with both higher valuations and lower shares of output accruing to labor, all else equal.[14] As shown by Crouzet and Eberly (2021) and Abel and Eberly (2011), the Tobin's Q valuation ratio reflects both rents and marginal Q, the marginal increase in the value of the firm with respect to an additional incremental unit of capital. In the presence of economic rents Tobin's Q overstates investment opportunities, and firms earning more rents have higher Tobin's Q without an attendant increase in investment.

In Table 1.1 I examine the relationship between labor productivity, labor shares, valuations, and operating profitability. In panel A I study the relationship between log value-added per worker and firm-level labor shares and log valuation ratios, respectively. Without controlling for industry-by-year fixed effects in the first column, the elasticity of the labor share with respect to labor productivity is a highly significant -.34. Size effects increase this slightly in the second column. Adding industry-by-year (where industry is defined at the 3-digit NAICS level) fixed effects increases the magnitude of this elasticity considerably, to about -.51 in

[14] For labor shares, with a Cobb-Douglass production function the labor share is $(1-\alpha)$, where $(1-\alpha)$ is the labor returns to scale. When there is a price markup over marginal cost due to product market power, the labor share is (price markup)$^{-1} \times (1-\alpha)$, where the price markup is the ratio of the price to marginal cost. With labor market power the labor share is wage markdown $\times (1-\alpha)$, where the wage markdown is the ratio of the wage to marginal revenue product of labor.

column 3. Note that even without market power, labor shares can also vary negatively with productivity as a result of production function variation.[15] Production function variability is likely to be larger between rather than within industries, but the negative productivity-labor share relationship is stronger within industries. This is consistent with the market power of more productive firms playing a role in driving down firm level labor shares.

In panel B I look at the relationship between log value-added per worker and the log of four different valuation ratios, focusing on the specification with size controls and industry-by-year fixed effects. Log value-added per worker is strongly associated with increased log total Tobin's Q (which is taken from Peters and Taylor (2017) and includes both intangible and physical capital) and in my employment-based Tobin's Q measure. Both have elasticities above 0.5 and highly significant t-stats. Productive firms have high enterprise values relative to their installed capital stock or skill-weighted labor force, as would be expected in the presence. The last two columns show that labor productivity is also associated with high market capitalization relative to book equity and sales.

Market power should also yield increases in profitability. I find that productive firms exhibit better operating performance in panel C of Table 1.1. This is true both contemporaneously and in the future, so the relationship is not merely a mechanical result arising from transitory increases in productivity that correlate with temporary contemporaneous increases in profitability. Instead, productivity and profitability show a persistent correlation intertemporally.

If the high valuations of productive firms are due to rents and not because of growth opportunities, we should not see a consummate rise in investment (or hiring) in response to valuation differences. In Panel A of Table 1.2 I sort firms into four quartiles of labor productivity and show the average Tobin's Q and average investment rates across these quartiles. In the first two rows I show the Peters and Taylor (2017) total (physical plus intangible) investment rates and average total Q. There is a strongly increasing average Q,

[15] If the production function is CES then labor shares are naturally decreasing in the productivity level when capital and labor are complements (Donangelo, 2020). Returns to scale could also differ so as to be correlated with log value-added per worker.

with an average of about 1.70 for the top quartile of productivity and 0.73 for the bottom quartile. The average investment rates are only slightly higher for the top quartile relative to the bottom quartile, this difference is not statistically significant, and the middle two quartiles of productivity have lower average investment rates than the bottom quartile. The results are similar, when I examine the hiring rates relative to my measure of employment Q in the bottom two rows of Table 1.2. The employment Q is several times larger for highly productive firms, and yet their hiring rates are actually significantly lower.

These findings are consistent with the result in Crouzet and Eberly (2021) showing that when firms earn rents their Tobin's Q overstates marginal Q, and hence overpredicts their investment decisions relative to actual investment. I examine this relationship further in panel B of Table 1.2, by running a regression of investment rates on Q ratios with productivity quartile-specific slopes and for investment types $k = Total, Hire$. As in Peters and Taylor (2017) I include firm and year effects. I also add labor productivity quartile fixed effects and further control for firm cashflows.

$$\text{Inv Rate}^k_{t+1} = \alpha^k + \alpha^k_{q(j,t)} + \alpha^k_j + \alpha^k_t + \sum_{q=1}^{4} \mathbf{1}(q(j,t) = q) \times \left(\beta^k_q Q^k_{j,t} + \delta^k_q CF_{j,t}\right) + \epsilon^k_{j,t} \quad (1.1)$$

The $q(j,t)$ denotes firm j's time-t productivity quartile, and $\alpha^k_{q(j,t)}$ are labor productivity quartile specific intercepts. Finally, $Q^k_{j,t}$ denotes the firm's total or employment Tobin's Q ratio, and $CF_{j,t}$ are the firm's cash flows, defined as in Peters and Taylor (2017). Panel B of Table 1.2 shows that β^k_q coefficients are lowest for the most productive firms and larger for the least productive firms, for both investment in capital and for hiring new workers. The fifth column gives the p-value on a test that the top and bottom labor productivity quartile slope coefficients are equal, which is rejected at high levels of statistical significance in both regressions. These findings could also be explained by productive firms facing higher adjustment costs, so in appendix Tables 1.12 and 1.13 I control for adjustment cost proxies, still finding the same basic result. I describe this test in appendix section 1.7.2. In sum, Table 1.2 provides suggestive evidence that average Q ratios overstate marginal Q for productive firms because they reflect in part the value of economic rents earned from current assets in place and not future investment opportunities.

Finally, I look at trends in the spread in productivity and Tobin's Q ratios between the most and least productive firms. I again sort firms into labor productivity quartiles and take the average log value-added per worker and log total/employment Q ratios for each year. I then take the difference between firms in the top- and bottom-quartiles of labor productivity over a three-year moving average centered at the current year. I plot the resulting series in Figure 1-1. The spread in value-added per worker has increased, echoing the findings of De Loecker et al. (2020), Autor et al. (2020), and Gouin-Bonenfant (2020), among others. Lastly, the spread in both of the average log Q ratios has also increased, suggesting a widening gap in economic rents between the most- and least-productive firms, likely driven in part by the increasing gap in productivity between the two groups.

Overall, the empirical patterns documented in this section are consistent with high and potentially increasing economic rents being earned by firms with high labor productivity. Productive firms have high market valuations without high investment rates, and also have lower labor shares. These stylized facts are correlational rather than causal, and could be consistent with the presence product market power, as in Barkai (2020), De Loecker et al. (2020), and Autor et al. (2020), as well as labor market power. In the following sections I examine to what extent, if any, labor market power could play a role in driving some of these empirical patterns. Direct evidence on the extent of labor market power comes from estimating the elasticity of labor supply to the individual firm, which I do in the next section.

1.3 Firm Productivity and Labor Market Power

In this section I first discuss my method for estimating supply elasticities. I then apply my method to firms sorted on labor productivity.

1.3.1 Relationship Between Wage Markdowns and the Supply Elasticity

The firm-specific elasticity of supply is the key quantity that determines wage markdowns in standard models of monopsony. To see this, consider the basic static monopsony framework (Robinson, 1969). The firm has a revenue function $F(L)$ and faces a labor supply function $L(w)$, where w is the wage offer. The firm solves:

$$\max_{w} \qquad F(L(w)) - wL(w) \tag{1.2}$$

First order conditions give

$$w = \frac{\epsilon}{1+\epsilon} F_L \tag{1.3}$$

where

$$\epsilon = \frac{dL}{dW}\frac{w}{L} \text{ is the supply elasticity.} \tag{1.4}$$

Define

$$\frac{\epsilon}{1+\epsilon} \equiv \text{ wage markdown} \tag{1.5}$$

Thus the wage markdown represents the fraction of labor's marginal revenue productivity that workers obtain in wages. In perfect competition $\epsilon \to \infty$, and the wage equals F_L, the marginal revenue product of labor.

1.3.2 Estimating Supply Elasticities

As emphasized by Manning (2003), supply elasticities can be identified by shocks to the marginal revenue product of labor, holding the firm-specific labor supply curve constant. In order to estimate supply elasticities, labor demand shocks to should be: 1) a persistent, unanticipated shock to firm productivity; 2) firm-specific; 3) not correlated with shocks to firm-specific labor supply. An objective of this paper is to estimate heterogeneous supply elasticities for a panel of Compustat firms sorted on labor productivity, as well as to allow for time-variation in these estimates, so my labor demand shock also needs to be available for a wide variety of firms and for a long period of time.

Stock Returns As a Labor Demand Shock

Firm-specific stock returns satisfy the above criteria for a labor demand shock quite well. In particular, they reflect a revision in expected discounted future cashflows, and hence inherently correlate with changes in current and expected future revenue productivity.[16] Consider the standard asset pricing equation that determines the price of a stock:

$$P_t = E_t \left[\sum_{k=0}^{\infty} M_{t,t+k} D_{t+k} \right]$$

Here D_{t+k} is a cashflow occuring at time $t+k$ and $M_{t,t+k}$ is the time t discount factor for time $t+k$ cashflows. The time t stock return is

$$R_t = \frac{D_t + P_t}{P_{t-1}}$$

Most of the variation in stock returns is firm-specific, and the common components can be removed rather easily. Because stock returns react to new information, they are inherently unexpected, and because they are forward looking they represent persistent changes in the outlook of the firm going forward. All this suggests that stock returns are good candidates for unexpected shocks to firm revenue productivity, which in turn moves firm labor demand.

However, firms may also gain value from the willingness of workers to be employed at the firm for any given wage, which affects the level of the labor supply curve. Unexpected changes in the level of labor supply could bias supply elasticity estimates when using stock returns as an instrument for labor demand shocks. To build intuition about the form this

[16] Previous research provides evidence that while discount rate shocks drive much of the variation in the aggregate stock market, news containing information about future cashflows drives the vast majority of the variation in firm-level stock returns. For example, Vuolteenaho (2002) finds that the idiosyncratic component of individual stock returns is almost entirely driven by firm-specific cash flow news, while movements in discount rates are primarily common across firms. Recent research (Neuhierl, Scherbina, and Schiusene, 2013; Boudoukh, Feldman, Kogan, and Richardson, 2018) also shows that fundamental information about firms, such as unexpectedly high earnings, acquisitions, or new product announcements, are an important driving force behind stock price movements. All these shocks can be expected to relate to firms' marginal revenue productivity.

bias may take, I derive closed form expressions for the bias in Appendix 1.7.3 with a simple model of the labor market that takes inspiration from Card et al. (2018) and Lamadon et al. (2019). There are two key insights from this exercise. The first is that failing to account for market-wide shocks is likely to bias stock-return based elasticity estimates towards zero, causing me to make erroneously large inferences about the importance of labor market power. Because of this I demonstrate the insensitivity of my baseline elasticity estimates to a host of specifications using different proxies for market-level shocks.

The second insight is that once market shocks are accounted for, any remaining bias is likely to cause my estimates to be conservative. This follows from the fairly innocuous assumption that firms do not cut amenities (or workers don't reduce their common perception of firms' amenities) by too much on average when idiosyncratic productivity improves. Hence workers shouldn't view the firm as a worse place to work at when firm-specific productivity goes up. This assumption implies my elasticity estimates are an upper bound on the true elasticity. My estimation strategy correctly identifies the supply elasticity when I further invoke the identifying assumption that only market-specific shocks move the level of the labor supply curve. It should be noted that this assumption is invoked in most papers that estimate passthrough parameters of firm-specific shocks to workers.[17] This is because unobservable firm-specific amenities shocks bias passthrough coefficients. This is an issue even when using plausibly exogenous variation in firm revenue productivity, because one still cannot rule out in general that amenities could also move as a response to the shock.

Because of this potential bias, in robustness checks detailed in Appendix Section 1.7.5 I examine elasticity estimates implied by different shocks to revenue productivity, and discuss the different identifying assumptions each implicitly invokes in order to estimate the supply elasticity. Finally, I identify sets of firms that were likely to have experienced observable firm-specific labor supply shocks and check how sensitive elasticity estimates are to the inclusion of these observable controls.

[17] For example: Lamadon et al. (2019), Kline et al. (2019), Garin and Silverio (2020). Papers in the asset pricing literature which incorporate labor market frictions (such as Kuehn et al. (2017), Belo et al. (2014), or Donangelo (2014)) also assume that labor supply shocks are determined at the market level.

My specification to obtain a baseline average supply elasticity estimate is as follows:

$$\log(Y_{j,t+1}) - \log(Y_{j,t}) = \alpha + \alpha_{I(j),t} + \beta \text{Stock Ret}_{j,t\to t+1} + \Gamma X_{j,t} + \epsilon_{j,t} \tag{1.6}$$

Here Y = firm Compustat employment or firm average full-time full-year equivalent adjusted wage. The ratio of the estimates $\widehat{\beta}^{Emp}/\widehat{\beta}^{Wage}$ gives the supply elasticity. The $\alpha_{I(i),t}$ denote 3-digit NAICS industry × year fixed effects, which are intended to control for common market shocks. The controls $X_{j,t}$ include the contemporaneous change in log firm average AKM worker effects $\alpha_{j,t}$ (defined in equation (1.41) in appendix section 1.7.1); and, lagged growth rates in employment, wages, and firm assets. I include the contemporaneous changes in worker effects in case stock returns lead to changes in the skill composition of the workforce. If this is the case, then failing to account for skill changes may cause employment and wage growth to be misstated in terms of constant efficiency units of labor. I include lagged growth rates in firm wages, employment, and assets to control for any pre-trends in labor demand or firm expansions that may be slightly correlated with future stock returns. The identifying assumption of the estimation strategy is that labor supply shocks operate at the market level, defined by industry-year. The annual stock returns in (1.8) are given by the sum of the firm's monthly returns in excess of the risk-free rate from July of year t through June of year $t+1$. Due to some extreme outliers in the tails of monthly returns, I cross-sectionally winsorize the outer 0.25% from each tail before summing excess returns to the annual level, although my findings are not meaningfully affected in any way by this decision.

Before estimating elasticities via (1.6), I first look at wages and employment responses over different horizons to see if there are pre-trends. Similar to Kogan et al. (2020), I examine the growth in average wages or employment over different forward- and backward-looking horizons of h years:

$$\log\left(\frac{1}{|h|}\sum_{k=1}^{|h|} Y_{j,t+k\times\text{sign}(h)}\right) - \log(Y_{j,t}) = \alpha + \alpha_{I(j),t} + \beta_h \text{Stock Ret}_{j,t\to t+1} + \Gamma X_{j,t} + \epsilon_{j,t} \tag{1.7}$$

Here Y = firm employment or average wage. This specification gives the growth of average employment or wages over the horizon and highlights persistent changes to wages and employment induced by stock return shocks. I estimate (1.7) for horizons $h = -5$ to $h = 5$ years. For $h = 1$ year the specification (1.7) is the same as (1.6). I plot the results in Figure 1-2. Because of the unforeseen nature of a shock to firm-specific stock returns, the pre-trends are non existent for both wages and employment. Meanwhile, both employment and wages display large, statistically significant and persistently positive responses at the time of and following the stock return shock. The joint positive employment and wage responses to movements in stock prices are consistent with my argument above that stock returns constitute a powerful and essentially unpredictable labor demand shock.

To examine pre-trends and shock responses at different horizons by worker skill, in Figure 1-3 I again estimate the employment and wage responses over forward- and backward-looking horizons as in (1.7), except broken down by each worker skill level. I define workers in the bottom two quintiles of estimated worker wage effects in (1.38) to be low skill, the third and fourth quintiles to be middle skill, and the top quintile to be high skill. Similar to Figure 1-2, I find no pre-trends and a strong positive wage and employment responses to a stock return shock. The skilled worker wage response is the largest in magnitude, generating a lower supply elasticity for skilled workers as found in Table 1.4.

Baseline Elasticity Estimates

In Panel A of Table 1.3 I estimate supply elasticities for workers of different skill levels. My baseline pooled average supply elasticity estimate is about 2.5, found in the first column of Panel A in 1.3. This is well in line with the range of estimates found in previous research, although on the slightly smaller end.[18] This is likely because of my focus on publicly traded firms, which are much larger than the typical firm in the US economy. The remaining columns show that supply elasticities decline in worker skill. This may cause one to worry that the larger wage passthroughs for skilled workers come because skilled workers who are

[18] Sokolova and Sorensen (2021) find the median supply elasticity estimate is around 1.7 in a meta-analysis. Bassier et al. (2020) point out that recent quasi-experimental evidence—such as Kroft et al. (2020), Cho (2018), and Dube, Manning, and Naidu (2018), and Caldwell and Oehlsen (2018)—has found supply elasticities between 2 and 5.

insiders are simply able to appropriate more rents generated from positive firm specific shocks. In Panels B and C of Table 1.3 I split workers into incumbent and recruit status, where incumbents have been employed at the firm at least since the previous year. I find lower elasticities for incumbent workers. However, consistent with a monopsony explanation for the low elasticities skilled workers, I also find that skilled workers face the lowest supply elasticities among recruits. My wage posting model in section 3.1 matches the pattern of high high wage passthroughs and low incumbent supply elasticities, which is caused by the costliness of hiring workers from outside the firm and the presence of an incumbent wage premium.

Robustness Checks

I run a host of robustness checks for my estimates of (1.6), which I cover in more detail in in Section 1.7.5 of the appendix. Since a major source of confounding variation comes from aggregate shocks, I carefully check to see how my estimates vary with different controls for common market level shocks in 1.7.5. One may also worry about the quantitative impact of other sources of endogeneity. In 1.7.5 I instead estimate supply elasticities with several alternative labor demand shifter—including customers' stock returns, R&D tax credits, patenting, and earnings announcement returns—finding very similar estimates. In 1.7.5 I also show that controlling for salient observable firm-specific labor supply shocks, such as unionization events or changes in non-compete enforceability, have no effect on the average supply elasticity estimate. The common finding from these exercises is that my baseline average supply elasticity estimate is very close economically and statistically to the various alternatives, suggesting that bias in my estimates is likely to be quantitatively small.

1.3.3 Supply Elasticities for Firms Sorted on Labor Productivity and Over Time

The stylized facts in Section 1.2 imply that such a sorting on labor productivity should induce wide variation in labor shares, firm valuations, and profitability, suggesting these firms earn rents but does not necessarily imply anything about their labor market. To provide direct evidence on differential labor market power across firms of varying labor productivity levels, I now estimate heterogeneous elasticities for firms sorted by labor productivity as proxied by log value-added per worker. In Table 1.11 I find that productive firms hire more skilled workers on average, and so I also break down my estimates by worker skill level. I sort firms into quartiles on log value-added per worker and estimate heterogeneous coefficients for each quartile of the productivity distribution. My main specification is the following:

$$\log(Y_{j,t+1})-\log(Y_{j,t}) = \alpha+\alpha_{q(j,t)}+\alpha_{I(j),t}+\sum_{q=1}^{4}\mathbf{1}(q(i,t)=q)\times\beta_q\text{Stock Ret}_{j,t\rightarrow t+1}+\Gamma X_{j,t}+\epsilon_{j,t} \tag{1.8}$$

where again Y = firm Compustat employment or firm average full-time full-year equivalent adjusted wage. Here $q(j,t)$ denotes the productivity quartile of firm j at time t. The controls $X_{j,t}$ are the same as introduced in (1.6), including lagged employment, wage, and asset growth, and the contemporaneous change in firm average worker skill. Finally, I add quartile-specific fixed effects $\alpha_{q(j,t)}$ to allow each productivity quartile to have a different intercept and slope. I also estimate (1.8) separately for workers of different skill levels. I define workers in the bottom two quintiles of the distribution of worker effects in the AKM decomposition (1.38) to be low-skilled; the third and fourth quintiles to be middle-skilled; and the top quintile to be high skilled. I separately compute the average wages, average AKM worker effects, and employment by worker skill level.[19]

In Table 1.4 I estimate (1.8) for the overall firm and for each skill type. Table 1.4 shows that more productive firms face lower supply elasticities at the firm level and at for each skill level. The overall supply elasticity for a highly productive firm is about 1.17, while it is 4.32 for a low productivity firm. These estimates mask heterogeneity by worker skill level; the

[19] Since I can't observe the share of Compustat employment by worker skill, I assume the number of workers of a given skill level are proportional to their observed shares in the LEHD. So if the firm has $N_{j,t}$ workers in the sample of the AKM decomposition (1.38) in year t, $N^s_{j,t}$ of which are from skill group s, then I assume Compustat employment in group s is equal to $EMP^s_{j,t} = EMP_{j,t}\times N^s_{j,t}/N_{j,t}$. I continue to use Compustat employment because it is reported at year end, which allows more time for employment to respond to a stock return shock than LEHD employment. Compustat employment also reflects workers at all the firm's establishments instead of just those at the establishments that are covered by the LEHD in that year. I make the same assumption when breaking down workers into their status as either incumbents or recruits.

highest skill workers have much lower supply elasticities, implying greater wage markdowns for the most skilled workers. However, even for low- and middle-skill workers the elasticities are decreasing in firm productivity. The test on coefficient differences between high- and low-productivity firms in fifth column also shows that differences in estimates elasticities are driven primarily by higher wage responses in the denominator—highly productive firms move wages by more to induce a given change in employment relative to unproductive firms.

The lower supply elasticities that productive firms face provide a competitive advantage, allowing them to markdown wages by more relative to marginal product, generating substantial economic rents to the firm. This finding squares with the results in Section 1.2 where I documented evidence that these same firms exhibit behaviors consistent with them earning rents derived from market power, such as having low investment and hiring rates given valuations and much lower labor shares.

Table 1.4 also shows that more skilled workers face lower supply elasticities at any given productivity level relative to less skilled workers. This implies that highly skilled employees have a greater difficulty of finding suitable substitutes among different potential employers. Skilled workers also exhibit much larger attachment to their employers. I find that the average separations rate for a firm's low-skilled workers is about 35.5%, whereas it is 24.6% for skilled workers. This suggests that firm-specific human capital could be a key driver of monopsony power among this group. Highly skilled workers are more educated and have generally made more specific human capital investments. This has the benefit of making skilled workers productive to their employers, but also leads to more difficulty in substituting their labor across firms. However, there is a flip side to this coin, which is that employees with specific skills are also hard for employers to replace. This can mitigate some monopsony power. In section 3.1 I present a dynamic wage posting model where hiring outside workers is costly, causing incumbent workers to enjoy a wage premium relative to the standard static marginal product markdown implied by the elasticity alone. I calibrate the model to quantify the attenuating influence such adjustment costs may have on firms' ability to fully exercise their monopsony power.

I now estimate time-varying heterogeneous supply elasticities. I estimate the baseline specification (1.6) for overlapping moving 3-year windows, both for the whole firm and each individual skill level. In Figure 1-4 I find that estimated supply elasticities have been decreasing for each skill group, and especially for skilled workers. Movements in labor market power may therefore have explanatory power for observed declines in the aggregate labor share documented by Karabarbounis and Neiman (2013), Autor et al. (2020), and numerous others. I examine how changes in elasticities impact the aggregate labor share in section 1.5.

My supply elasticity estimates can also explain cross-sectional dispersion in firm-level labor shares over time. To demonstrate this, I estimate (1.8) for overlapping moving 3-year windows and take the elasticity estimates for the top- and bottom-productivity quartiles. I use the standard markdown formula to estimate the elasticity-implied spread in log markdowns between the two productivity types and compare this to the spread in log labor shares. The log markdown spread is

$$\text{Log Markdown Spread}_t = \log\left(\frac{\widehat{\epsilon_{4,t}}}{1+\widehat{\epsilon_{4,t}}}\right) - \log\left(\frac{\widehat{\epsilon_{1,t}}}{1+\widehat{\epsilon_{1,t}}}\right) \tag{1.9}$$

where $\widehat{\epsilon_{q,t}}$ is the estimated supply elasticity for productivity quartile q in the 3-year moving window centered at time t. At the same time, I compute the average log labor share of the high- and low-productivity firms over the same 3-year window:

$$\text{Log Lshare Spread}_t = (1/3)\sum_{\tau=t-1}^{t+1}\left(E[\log(lshare_{4,\tau})] - E[\log(lshare_{1,\tau})]\right) \tag{1.10}$$

Figure 1-5 plots the two standardized series. Due to the availability of LEHD wage data and the requirements of my regression specification, I compute supply elasticity estimates for the years 1992-2013. There is a clear tight link between the cross-sectional dispersion in labor shares and the wage markdowns for firms sorted on labor productivity, with the two series tracking each other closely over time. The correlation is about 0.73, with a Newey-West t-stat of 4.10. Since there are only 22 observations I compute the small-sample adjusted t-stat, and I choose a lag length of 5 years due to the persistence of the two series. The widening gap in labor shares and elasticity-implied markdowns in Figure 1-5 echoes the changing relative valuations and productivity levels of high- and low-productivity firms in Figure 1-1, suggesting the phenomena may be jointly linked. In accordance with this, the model I introduce and calibrate in Section 3.1 forges a direct link between productivity advantages, labor market power, and firm labor shares.

My results in this section imply that labor market power is likely to be a quantitatively important determinant of firm rents and cash flows, especially for the most productive firms. While my estimates are informative about the magnitudes of labor market power, they may not be sufficient by themselves to quantify the cashflows firms derive from wage markdowns. As discussed before, in the presence of labor adjustment costs the supply elasticity is not the only determinant of markdowns from marginal product. I formalize this concept in section 3.1, where I propose a model to map my empirical supply elasticity estimates into quantitative markdowns in the presence of labor market adjustment costs.

Robustness Checks for Labor Productivity Sorted Supply Elasticities

I now briefly discuss robustness checks for labor productivity-sorted supply elasticity estimates; for further details on all these estimates see 1.7.5 in the appendix. In appendix Table 1.20 I sort firms into labor productivity quartiles within 2-digit NAICS industry, and find that every key finding from my baseline sorts in Table 1.4 is unchanged. In appendix Table 1.21 I perform several more robustness checks on elasticities sorted on productivity. I first estimate elasticities at the 3-year instead of 1-year horizon to test whether short term frictions drive the lower elasticity estimates for unproductive firms. Next I replace replacing average log wage changes with time-varying AKM firm wage effects to verify that sorting patterns are not likely to be driven by differences in incentive pay or match components of wages. Then I sort firms on TFP estimates from İmrohoroğlu and Tüzel (2014) instead of value added per worker. Finally, I replace Compustat employment changes with LBD employment changes, finding that Compustat-based elasticity estimates are more conservative. In all cases I find the same strongly monotonic decreasing pattern in supply elasticities as productivity increases.

1.4 Dynamic Wage Posting Model

In this section I introduce a dynamic wage posting model where a single firm hires from a fixed pool of potential employees. My primary focus is on quantifying how much adjustment costs affect markdowns from marginal product relative to the standard static elasticity-based markdown formula. The model is a dynamic extension on the static framework in Kline et al. (2019), and is also closely related to the setup of Card et al. (2018).

1.4.1 Setup

Consider the dynamic wage posting problem of a labor market with a single firm. There are a fixed total of L workers in the market, with a fraction λ_t currently employed at the firm at the start of period t. The firm can make different wage offers to outside hires ("O") and incumbents ("I"). The firm faces labor supply functions for these two classes of worker. For incumbents:

$$L_t^I = \lambda_t L \frac{(w_t^I - b)^\beta}{\kappa} \tag{1.11}$$

For outside hires:

$$L_t^O = (1 - \lambda_t) L \frac{(w_t^O - b)^\beta}{\kappa} \tag{1.12}$$

This labor supply functional form is borrowed from Card et al. (2018) and Kline et al. (2019). The $\frac{(w_t^k - b)^\beta}{\kappa}$ represents the probability that a worker of type $k = I, O$ accepts the offer to work at the firm. Card et al. (2018) derive the above labor supply function in a model of worker discrete choice with unobserved taste shocks between firms; like Kline et al. (2019) I take the functional form in equations (1.11) and (1.12) as given. Unlike Kline et al. (2019), I assume that firms choose the wage offer for both their incumbent workers and outside hires. Card et al. (2018) show that this supply functional form has some nice properties. In particular, it has a representation of the optimal wage that is isomorphic to a bargaining model, where $\beta/(1 + \beta)$ acts like the workers' bargaining weight and b functions like the workers' outside option. In Card et al. (2018) the parameter β is inversely related with the variance in workers' unobservable preferences among firms; higher values of β lead to more elastic demand. The parameter κ is simply a scaling factor to ensure that the probability of a given worker choosing to be employed at that firm is bounded between 0 and 1. In order to focus on the firm's decision, I deliberately make the worker employment decision simple. An incumbent or potential recruit takes the wage offer at time t and decides whether or not to be employed at the firm for that period according to the choice probabilities implied in (1.11) and (1.12). Though β and b could in theory be worker-type specific, I restrict the them to be the same for incumbent workers and outside hires in order to limit the number of free parameters when I calibrate the model.

The supply elasticity facing the firm for workers of type $k = I, O$ is

$$\frac{\beta w_t^k}{w_t^k - b} \tag{1.13}$$

This is decreasing in the wage and is increasing in β and b.

Firms may pay outside hires differently because they are imperfect substitutes for incumbents. Specifically, hiring outside workers incurs quadratic adjustment costs:

$$c_t(L_t^O) = \exp(\sigma_c z_t)\bar{C}\lambda_t L \left(\frac{L_t^O}{\lambda_t L}\right)^2 \tag{1.14}$$

The adjustment costs may represent the cost of training workers in firm-specific production technologies or recruitment/hiring costs. I include the term $\exp(\sigma_c z_t)$ to allow adjustment costs to be increasing in firm productivity. Given the Belo, Li, Lin, and Zhao (2017) and Jager and Heining (2019) findings that adjustment costs are likely higher for skilled labor, this feature is a reduced form way of getting adjustment costs to increase in firm "skill" in a setting where there is only one type of labor. The $\lambda_t L$ term gives the number of incumbent workers employed at the firm at the start of time t, so adjustment costs are quadratic in the ratio of outside hires to start of period incumbents. This is similar to the formulation in Kline et al. (2019).

The firm produces output according to a Cobb-Douglas production function in the sum of inside and outside hires. Firm log productivity z_t follows an AR(1) process. There is no aggregate risk and firm cash flows are priced at the constant risk-free discount rate r. The firm solves

$$V_t = \max_{w_t^I, w_t^O} F_t(L_t^I, L_t^O) - \left(w_t^I L_t^I + w_t^O L_t^O + c_t(L_t^O) \right) + \exp(-r) E_t[V_{t+1}] \tag{1.15}$$

Subject to $L_t^O = (1 - \lambda_t) L \frac{(w_t^O - b)^\beta}{\kappa}$ and $L_t^I = \lambda_t L \frac{(w_t^I - b)^\beta}{\kappa}$

$$\lambda_{t+1} = (L_t^O + L_t^I)/L \tag{1.16}$$

Where

$$F_t(L_t^I, L_t^O) = \exp(z_t) \left(\bar{z}(L_t^I + L_t^O) \right)^{(1-\alpha)} \tag{1.17}$$

$$z_{t+1} = \rho_z z_t + \sigma_z \epsilon_{z,t+1} \tag{1.18}$$

and $c_t(L_t^O)$ is the cost of hiring L_t^O workers. The law of motion for λ_t means that a time t retained incumbent or outside hire becomes an incumbent at the start of $t + 1$.

1.4.2 Wage Markdowns with Adjustment Costs

The presence of adjustment costs changes the markdown formula in the standard monopsony framework. To see this, we take the first-order conditions of (1.15) for worker types $k = I, O$:

$$w_t^k = \frac{\varepsilon_k(w_t^k)}{\varepsilon_k(w_t^k) + 1} \left(MRP_t^k - \frac{\partial c_t}{\partial L_t^k} + \exp(-r) E_t \left[\frac{\partial V_{t+1}}{\partial L_{t+1}^I} \right] \right) \tag{1.19}$$

Here MRP_t^k is $\frac{\partial F_t}{\partial L_t^k}$, which is the same for incumbents and recruits given the production function $F_t(L_t^I, L_t^O) = \exp(z_t) \left(\bar{z}(L_t^I + L_t^O) \right)^{(1-\alpha)}$. There are two additional terms in (1.19) relative to the static markdown formula. The first is $\frac{\partial c_t}{\partial L_t^k}$, the derivative of the adjustment cost function with respect to an additional unit of labor of type k. This term is zero for incumbents and positive for outside hires. The presence of adjustment costs leads the firm to prefer retaining incumbent workers over replacing them with outside hires, generating a wage premium for incumbent workers over recruits.

The second additional term relative to a standard static wage markdown in the dynamic wage offer equation is $\exp(-r)E_t\left[\frac{\partial V_{t+1}}{\partial L^I_{t+1}}\right]$, which is the time t expected marginal benefit of having an additional incumbent worker at time $t+1$. Because both a retained incumbent or recruit hired today becomes an additional worker in the available pool of incumbents tomorrow, this term is the same for both types of workers. This term is also always positive because entering the next period with more incumbents is strictly better than having fewer incumbents. This implies that wage markdowns from marginal product will be strictly smaller for incumbent workers than would be inferred by supply elasticities alone, and may be larger or smaller for recruits depending on the relative size of $\frac{\partial c_t}{\partial L^O_t}$ and $\exp(-r)E_t\left[\frac{\partial V_{t+1}}{\partial L^I_{t+1}}\right]$. Note also that since incumbents and outside hires face different wage offers, the supply elasticities themselves also differ because the elasticity (1.13) is a decreasing function of the wage offer for each worker type.

Define for worker types $k = I, O$

$$\textit{dynamic markdown}^k_t \equiv \frac{w^k_t}{MRP^k_t - \frac{\partial c_t}{\partial L^k_t} + \exp(-r)E_t\left[\frac{\partial V_{t+1}}{\partial L^I_{t+1}}\right]} = \frac{\varepsilon_k(w^k_t)}{\varepsilon_k(w^k_t)+1} \tag{1.20}$$

$$\textit{MRP markdown}^k_t \equiv \frac{w^k_t}{MRP^k_t} \tag{1.21}$$

$$\textit{markdown wedge}^k_t \equiv \frac{\textit{MRP markdown}^k_t}{\textit{dynamic markdown}^k_t} \tag{1.22}$$

The markdown expressions (1.20), (1.21), and (1.22), are worker type specific, so I introduce their firm-level aggregates:

$$\textit{dynamic markdown}_t \equiv \frac{\sum_k w^k_t L^k_t}{\sum_k \left(MRP^k_t - \frac{\partial c_t}{\partial L^k_t} + \exp(-r)E_t\left[\frac{\partial V_{t+1}}{\partial L^I_{t+1}}\right]\right) L^k_t} \tag{1.23}$$

$$\textit{MRP markdown}_t \equiv \frac{\sum_k w^k_t L^k_t}{\sum_k MRP^k_t L^k_t} \tag{1.24}$$

$$\textit{markdown wedge}_t \equiv \nu_t \equiv \frac{\textit{MRP markdown}_t}{\textit{dynamic markdown}_t} \tag{1.25}$$

I examine the properties of these objects in the calibration. The markdown wedge is a particularly important quantity, as it allows me to map my empirically estimated supply elasticities to model implied markdowns.

1.4.3 Calibration and Model Fit

Closed form solutions for the firm's objective function (1.15) are not readily available, so I solve the model for a given calibration through value function iteration. I calibrate the model to match the average separations rate; the incumbent wage premium; overall supply elasticity, incumbent/recruit specific supply elasticities; the relative stock-return wage passthroughs of incumbents and recruits; and the average log labor share.[20] I estimate model-implied supply elasticities by running a regression of model-generated employment and wage growth on stock returns (growth in the value function). I use the Rouwenhourst (1995) method to create a discretized grid of 9 points to approximate the AR(1) productivity process. For pre-calibrated parameters, I normalize the labor force size L to equal 1; set the annual discount rate to .02; set the supply scale factor κ to .741[21]; I set $\rho_z = 0.9$ to match the persistence of annual log value-added per worker in the data; and finally, I set both the baseline labor productivity $\bar{z}$ and convex adjustment cost volatility σ_c equal to 1.

Table 1.5 shows the model calibration of the externally and internally calibrated parameters, and Panel A of Table 1.6 displays the model fit to targeted moments. There are 5 internally calibrated parameters for 7 targeted moments. The fit quantitative fit is very good, with almost exact matches on average incumbent wage premium, separations rates, relative wage passthroughs, and average log labor shares; the model calibration does slightly underestimate

[20] The incumbent wage premium is skill-adjusted and represents the intercept from a regression of the wage premium on the incumbent minus recruit difference in average AKM worker effects. I obtain incumbent and recruit elasticities by estimating (1.6) just for incumbents and recruits, so the growth rates in the left hand side are the growth in wages for incumbents/recruits or growth in the number of incumbents/recruits. The independent firm-specific control variables $X_{j,t}$ are also their analogues just for incumbents and recruits, with the exception of lagged asset growth which is not worker type specific.

[21] The parameter κ is technically a function of a pre-calibrated parameter and an internally calibrated parameter. I obtain κ by setting $\kappa = (w_{max} - b)^{\beta}$, where w_{max} is pre-set to 1. in equilibrium this ensures that the probability of being employed at the firm is always bounded between 0 and 1 for both worker types and all productivity levels.

average labor supply elasticities, but is still pretty close quantitatively. I calibrate the parameters $\beta = 0.38$, $b = 0.545$, and $\sigma_z = 0.145$ to help match the average supply elasticities and the incumbent/recruit stock return wage passthrough ratio; the parameter $\bar{C} = 0.55$ is identified primarily by the incumbent-recruit wage premium. The parameter $\alpha = 0.22$, which is one minus labor returns to scale, and productivity volatility $\sigma_z =$ jointly help me get close to the the average empirical supply elasticity and log labor share. Finally, the average wage levels implied by the combination of parameters help to jointly match the empirical average firm level separations rate.

The calibrated parameters are also in a reasonable range relative to analogous parameters in prior research. For comparison I look at the calibrations of the dynamic models in two other papers, Kuehn et al. (2017) and Belo et al. (2017), which both have model features that are related to mine. Kuehn et al. (2017) calibrate the benefit of unemployment to be 0.71; my analogous parameter is the outside option of b, calibrated at 0.545. Kuehn et al. (2017) also have a worker bargain weight of 0.11. The most comparable parameter in my calibration is $\beta/(1+\beta) = 0.27$, which is a little higher. However, the parameter β also affects the supply elasticity, and so $\beta/(1+\beta)$ doesn't have the exact same structural interpretation as it would in a pure bargaining model because it also directly affects the endogenous labor supply response to wage changes. Belo et al. (2014) calibrate the high-skill convex adjustment cost parameter to 1.8 and the low-skill convex adjustment cost parameter to 0.17; my calibration of $\bar{C} = 0.55$ lies in between the two. Finally, my calibration of returns to scale of $1 - \alpha = 0.78$ lies in between those of Kuehn et al. (2017) and Belo et al. (2017) (0.75 and 0.85, respectively). Again note that these parameters don't exactly map between one another in the papers, although they are related in important ways; I merely present these parameter estimates to compare to some reasonably close analogues in recent prior literature.

A supply elasticity that decreases in the wage is a key feature of the model that helps match the data moments. Consistent with the data, recruits face higher average supply elasticities. In the model this is because their wage offers are closer to the reservation wage level b, and elasticities can be arbitrarily high as wage offers approach b. This feature of the labor supply function also implies that firms face lower elasticities in more productive states. In Panel B Table 1.6 I show that the model is able to match the monotonic pattern in supply elasticities and log labor shares when sorting by labor productivity without explicitly targeting these moments. Interestingly, although the model matches the monotonic pattern in firm labor shares it misses on the magnitude of the spread. The log labor share spread is 0.37 in the model but 0.72 in the data. In the next section I show that this is also the case for my empirical supply elasticity estimates–although they imply a spread in labor shares, the spread is wider in the data than implied by my empirical estimates. This motivates a decomposition of the labor share spread in Section 1.5, where I examine how much of a role there is left over for differences in markdowns or labor returns to scale after accounting for the implied spread in markdowns.

I now analyze the behavior of markdowns in the model. In the top panel of Figure 1-6 I compare average markdowns implied by estimating supply elasticities in a regression of employment and wage growth on stock returns versus the actual average markdown from the true supply elasticities determined by (1.13). Because of adjustment costs, the regression implied elasticities slightly underestimate the wages paid to incumbent and overestimate the wages paid to recruits. However, when aggregating to the firm level the linear regression elasticity-implied markdown closely reflect the average dynamic markdowns for the whole firm. In the second panel of Figure 1-6 I compare the dynamic markdowns the marginal revenue product markdowns. This comparison shows the difference between the markdowns that would be implied by supply elasticities alone versus the actual markdown from marginal product. At the firm level, the average dynamic markdown is 0.68, which implies that firm pays about 68% of the denominator in (1.23) to its workers in wages on average. However, this again masks heterogeneity by worker type. Because incumbents are costly to replace, their wages are only marked down by 83% from marginal product; without accounting for dynamics due to adjustment costs, supply elasticities alone would imply a wage of 64% of marginal product. The reverse is true for recruits: supply elasticities would imply a wage that is about 80% of marginal product, but recruits' wages are actually about 71% of marginal product.

A key object for empirical quantification is the average firm-level markdown wedge implied by the model. This is about $0.79/0.68 \approx 1.16$, so wage markdowns at the firm level are about 16% smaller in magnitude after accounting for adjustment costs than would be inferred by the simple static markdown formula $\epsilon/(1+\epsilon)$. I label the markdown wedge ν. In Table 1.7 I show the markdowns from marginal revenue product that result from my supply elasticity estimates in section 1.3 combined with the markdown wedge ν.[22] Even after adjusting for dynamics due to adjustment costs, the markdowns still imply a considerable amount of monopsony power. However, adjustment costs do negate labor market power quite a bit. The baseline elasticity estimate of 2.52 yields a markdown that is 72% of marginal product; after adjusting for dynamics this becomes 83% of marginal product. A supply elasticity of about 4.9 would give this markdown in a static setting. Table 1.7 combines the model markdown wedge with my empirical elasticity estimates. The table also shows that while low productivity firms do mark wages down, they still pay 94% of marginal product. This is reasonably close to a competitive benchmark. On the other hand, the most productive firms pay 62% of marginal product even after adjusting using the markdown wedge ν.

Calibration for 1991-2002 and 2003-2014 Subperiods

Given the downward trends in supply elasticities and estimated markdown spreads that I find in Figures 1-4 and 1-5, I introduce calibrations to match moments for the first and second halves of the sample (1991-2002 and 2003-2014, respectively). The calibrated parameters by subperiod are in appendix Table 1.14. The average cross-sectional standard deviation of log value added per worker increases by about 5% over the two subperiods, so I decrease σ, the productivity dispersion parameter, by about 2.5% relative to the main calibration for the 1991-2002 period and increase it by 2.5% for the 2003-2014 calibration. I calibrate the remaining parameters to match the same moments as in my main calibration.

In 1.15 I give the corresponding model and data moments for the subperiods. Supply

[22] For simplicity I report results assuming the markdown wedge is constant across productivity levels. Results are very similar quantitatively if I allow the markdown wedge to vary by productivity.

elasticities and separations rates decline empiricially across the two periods; I decrease in the parameter β, which governs the average supply elasticity, and b, which governs the worker outside options to match these moments. My calibration of β decreases from 0.5 to 0.34 from the first half to the second half of the sample, and b decreases from 0.549 to 0.515. This agrees in principle with Stansbury and Summers (2020), who suggest that worker power in the workplace has declined over recent decades.

The incumbent log wage premium increased by about 30% (from 0.13 to 0.17) across the two periods. This suggests increased adjustment costs, and I increase $\bar{C}$, the convex adjustment cost parameter, from 0.38 to 0.745 to match these moments. The resulting model-implied markdown wedge ν increases from 1.11 to 1.24 between the first and second halves of the sample period, mitigating some, but not all, of the impact of decreasing supply elasticities. I use these subperiod-specific elasticity estimates and markdown wedges when I quantify the model in the next section.

1.5 Quantifying Empirical Estimates

In this section I use the model-implied markdown adjustment parameter ν in conjunction with my empirical elasticity estimates from Section 1.3 to quantify the value that firms derive from their wage markdowns. I also use the model and estimated elasticities to back out the share of variation in labor shares across time and between firms of different productivity types that can be ascribed to differences in markdowns.

1.5.1 Value of Cashflows Derived from Labor Market Power

With the markdown wedges ν from the model and the empirical elasticity estimates for firms sorted on labor productivity, I now examine the counterfactual cash flows firms would bring in if they were not able to mark down wages from marginal product, but still held their production decisions constant. This is different from the competitive counterfactual, because in the competitive equilibrium both quantities and prices of labor would adjust. Instead, this counterfactual transfers the capital income attributable to wage markdowns from the actual owners of the firm's capital to the firm's workers after production has occurred. I use two measures of cash flows. My primary measure is based off the firm's operating income. I use the Compustat variable OIBDP adjusted for changes inventories, which was also used as the basis for my computation of firm value-added in (1.34). Hartman-Glaser et al. (2019) use OIBDP as a proxy for capital income available to the owners of debt and equity issued by the firm. The ratio of the value of wage markdowns to operating income can then be thought of as representing the share of capital income coming from wage markdowns. Going forward I use operating income and capital income interchangeably.

Denote $\widehat{\epsilon_{q,t}}$ the elasticity estimate for productivity quartile q at time t, where I use the elasticities estimates for either the 1991-2002 or 2003-2014 subperiods depending on which period year t falls in. Let $q(j,t)$ give the productivity quartile of firm j at time t. Finally, let ν_t denote the markdown wedge estimated over the subperiod where year t falls. Counterfactual labor expenses without markdowns are

$$\widetilde{LABEX}_{j,t} = \nu_t^{-1} \frac{\widehat{\epsilon_{q(j,t),\, t}} + 1}{\widehat{\epsilon_{q(j,t),\, t}}} LABEX_{j,t} \tag{1.26}$$

And the resulting markdown share of operating income:

$$\text{Markdown Share}^{OI}_{j,t} = \frac{\widetilde{LABEX}_{j,t} - LABEX_{j,t}}{OI_{j,t}} \tag{1.27}$$

I then compute the average markdown share of operating income:

$$\text{Average Markdown Share of OI} = \frac{1}{N} \sum_t \sum_j \text{Markdown Share}^{OI}_{j,t} \tag{1.28}$$

Here N denotes the total number of firm-year observations in the sample. Some firms report negative operating income, and occasional extreme values of Markdown Share$^{OI}_{j,t}$ skew the firm-level averages, so in (1.28) I focus on the subset of firms for which $OI_{j,t} > \widetilde{LABEX}_{j,t} - LABEX_{j,t}$, which are likely to have lower measurement error skewing the average. Instead of relying on a subsample of firms, I also compute the median fraction of operating income derived from wage markdowns among a all firms.

Finally, I look at the aggregate share of operating income that comes from wage markdowns by summing true and counterfactual income across the whole population of firms:

$$\text{Aggregate Markdown Share of OI} = \frac{1}{T}\sum_t \left[\frac{\sum_j \widetilde{LABEX}_{j,t} - LABEX_{j,t}}{\sum_j OI_{j,t}}\right] \tag{1.29}$$

with T the number of years in the sample. I compute (1.28) and (1.29) for all firms and separately for firms in the top and bottom quartiles of labor productivity.

I report these averages in Table 1.8 overall and for firms in the top and bottom quartiles of labor productivity. In Panel A I focus on results for the full sample period. The first two rows of Panel A show the firm-level mean and median shares of operating income derived from paying wages that are different from marginal product. The average firm in the sample earns about 34% of its operating income by paying wages lower than marginal product; this figure is 30% for the median firm. There is large heterogeneity between the most- and least productive firms, with those in the bottom quartile of productivity earning only 17% of their income from wage markdowns, while firms in the top quartile earn about 43%. Since the operating income represents earnings available to the firm's capital owners, these estimates suggest that about a third of the capital income generated by the typical firm comes from wage markdowns.

In the third row of Panel A in Table 1.8 I compute the aggregate operating income share of wage markdowns, as in equation (1.29). Because markdowns are concentrated amongst larger firms, the aggregate average is higher than the firm-level average, at about 40%. Hence two-fifths of the capital income generated by publicly traded firms were attributable to wage markdowns over the 1991-2014 period. Assuming these cash flows would be discounted at the same rate as capital income overall, this has the interpretation that the dollar value of wage markdowns at the monopsony equilibrium is worth roughly 40% of the total enterprise value of publicly traded firms. However, note that this is a very different statement than saying that the aggregate value of publicly traded firms would be 40% lower without wage markdowns, because the exercise here holds equilibrium labor demand constant. The decline in total enterprise value in the counterfactual competitive equilibrium would be substantially smaller than 40%. I discuss this point further later on in this section.

Panel B of Table 1.8 breaks down these figures by the 1991-2002 and 2003-2014 subperiods. At the firm level, the average markdown share of operating income increases slightly from 0.34 to 0.35 between the 1991-2002 and 2003-2014. The median increases slightly more, from 0.29 to 0.32; finally, the aggregate operating income share also increases from 0.39 to 0.41. These slight overall increases mask heterogeneity by productivity type: the mean, median, and aggregate average share of operating income generated from wage markdowns actually dropped for the least productive firms, while all three rose for the most productive firms. The largest increase was for the median high productivity firm, which saw their wage markdowns as a share of operating income increase from 0.45 to 0.52 between the two periods. This divergence parallels the increase in productivity dispersion between top and bottom firms over the sample period, as shown for example in Figure 1-5.

Interpreting the Magnitude of Cash Flows Generated from Wage Markdowns

Given the size of wage markdowns as a fraction of operating income in Table 1.8, it is helpful to provide some context for interpretation and to compare the magnitude against other results.

First of all, note that these markdown shares of capital income do not represent the change in the value of the firm relative to the competitive equilibrium. Constructing a reasonable competitive counterfactual would require imposing far more structure on my model, including assumptions about labor supply across firms and product market competition. Rather, I am asking the question "What is the total value of the gap between wages and marginal products relative to total capital income, *holding equilibrium quantities fixed*?" In a competitive equilibrium quantities as well as prices would adjust. Hence my counterfactual does not imply, for example, that the aggregate enterprise value of publicly traded firms would be 40% lower if we imposed perfect competition. To illustrate, in appendix Section 1.7.4, I make a back of the envelope calculation in a simple static representative firm setting where wage markdowns are calibrated to be exactly 40% of operating income, and the production function is calibrated to roughly match the average aggregate labor share in my sample. In this setup, firm value declines by 16% in the counterfactual competitive equilibrium relative to the monopsony equilibrium.

How does the magnitude of my estimates compare with other estimates of market power? In the next subsection I introduce two measures of aggregate labor shares; depending on the measure the mean aggregate labor share is about 49-55% of value-added, which implies that wage markdowns average about 18-20% of total output. In comparison, Crouzet and Eberly (2021) point out that the De Loecker et al. (2020) price markup estimates imply rents from product market power were worth almost 40% of value added by the end of their sample period, so my estimates are about half as large as this figure. The De Loecker et al. (2020) estimates are also closely related to the ratio of sales to cost of goods sold, and they assume there is no monopsony power. Imperfect competition in labor markets can also affect the ratio of sales to costs of goods sold, so their estimates likely reflect both labor and product market power. Herschbein, Macaluso, and Yeh (2020) jointly estimate wage markdowns and price markups for manufacturing firms using a production function approach that is related to the De Loecker et al. (2020) method. They find that the average firm pays about 65% of marginal product. Since they estimate that markdowns increase in firm size, the value-weighted gap would be even wider. In comparison, I find that total aggregate wages are about 70% of aggregate marginal product, which is smaller but of a comparable magnitude.

1.5.2 Decomposition of Labor Share Differences in the Cross-Section and Time Series

Next, I look at what fraction of the gap in labor shares between high- and low-productivity firms can be explained by differences in wage markdowns; I also examine how changes in the aggregate labor share in the time series were impacted by changing markdowns. In this section I use two measures of labor shares. My baseline, defined previously in (1.37), comes from imputed labor expenses $LABEX_{j,t}$ derived from LEHD earnings and Compustat employment data. I find that $LABEX_{j,t}$ is on average a little lower than Compustat reported staff and labor expense (Compustat variable XLR), but the variable XLR is only available for a very small fraction of firms. To get around this I create an imputed version of XLR, which uses the fact that $LABEX_{j,t}$ and $XLR_{j,t}$ are very highly correlated when $XLR_{j,t}$ is observable. Details on the imputation are in appendix section 1.7.1. I label this imputed version $\widehat{XLR}_{j,t}$, and create my alternative labor share measure by simply replacing $LABEX_{j,t}$ with $\widehat{XLR}_{j,t}$ in equations (1.34) and (1.37). I call this measure LShare $(\widehat{XLR})_t$. The logs of the two firm-level labor share measures share a high correlation of 0.91, though the time series behavior of the aggregate labor share from summing the two versions of labor expenses is a bit different; I elaborate on this point when I do the decomposition of aggregate labor share changes in the time series.

Table 1.6 shows that the model generates a monotonically decreasing pattern in labor shares as productivity improves, as is also found in the data. Still, the labor share spread in the model is not as wide as the empirical spread. This is also true when using the empirically estimated markdowns in Table 1.7, and leaves room for other factors, such as differences in product market power or production technologies, to explain the remainder of the labor share spread. For example, assume a Cobb-Douglas production function in labor, $F(L) = AL^{1-\alpha}$, but also suppose the firm chooses output according to inverse demand $P(Q) = P_0 Q^{-1/\gamma}$, where γ is the price elasticity of demand. Also assume that because of labor market power wages are a markdown μ from the marginal revenue product of labor. Then the revenue function is $R(L) = P_0 A^{1-1/\gamma} L^{(1-1/\gamma)(1-\alpha)}$, where $1 - 1/\gamma$ gives the inverse price markup. The log labor share satisfies

$$\begin{aligned}\log(\text{Labor Share}) = \log\left(\frac{\mu R'(L) L}{R(L)}\right) &= \log(\mu) + \log(1 - 1/\gamma) + \log(1-\alpha) \qquad (1.30)\\ &= \log(\text{wage markdown}) - \log(\text{price markup}) + \log(\text{labor returns to scale})\end{aligned}$$

The difference in average log labor shares can then be ascribed to differences in average log wage markdowns, price markups, and returns to scale. What percentage of the average labor share differences between high- and low productivity firms can be attributed to differential wage markdowns? In Panel A of 1.9 I answer this question using the wage markdown estimates from Table 1.6 and my two measures of firm labor shares. I show estimated log markdowns and average log labor shares. The difference in empirical log markdowns from Table 1.7 is $\log(0.62) - \log(0.94) \approx -.41$, which I report in the bottom row of the third column of Panel A. The differences in the average log labor shares between high and low productivity firms if -0.72 for my baseline and -0.65 for the $\widehat{XLR}$ based measure; markdown differences can account for roughly three-fifths of the difference in average log labor shares (57% for my baseline labor share measure and 63.5% for the alternate measure).

My findings still leave an important quantitative role for product market power to play in fully explaining the cross-sectional gap in labor shares. If labor returns to scale are approximately the same between the two groups of firms then the remaining two-fifths could be explained by differences in price markups. The increasing product market power among dominant firms found by Autor et al. (2020) and De Loecker et al. (2020) and the widening markdowns that I find are likely to be jointly important phenomena quantitatively. Also note that my model implicitly forges a link between product and labor market power, because increases in product market power increase labor productivity, which in turn increases labor market power, both in my model and in the data.

I next compute the aggregate labor share using either $\widehat{XLR}$ or $LABEX$. The aggregate labor share is given by

$$\text{Agg LShare } (LABEX)_t = \frac{\sum_j LABEX_{j,t}}{\sum_j VA_{j,t}} \tag{1.31}$$

for $LABEX$, and similarly for $\widehat{XLR}$:

$$\text{Agg LShare } (\widehat{XLR})_t = \frac{\sum_j \widehat{XLR}_{j,t}}{\sum_j VA_{j,t}} \tag{1.32}$$

Here $VA_{j,t}$ is the value-added for firm j at time t, computed by adding either $LABEX_{j,t}$ or

$\widehat{XLR}_{j,t}$ to operating income adjusted for changes in inventory as in (1.34). Although the firm-level variation in log labor shares for the two measures is very highly correlated, Figure 1-7 shows that the downward trend in labor shares is more pronounced for Agg LShare ($\widehat{XLR}$). The average aggregate labor share drops from 0.579 to 0.524, when comparing the 1991-2002 and 2003-2014 subperiods; meanwhile Agg LShare ($LABEX$) shows a more modest decline of 0.506 to 0.485.

Since the trend in Agg LShare ($\widehat{XLR}$) is much closer to the Compustat labor share trends found by Hartman-Glaser et al. (2019), in the next exercise I focus primarily on changes in Agg LShare($\widehat{XLR}$), but show results for both versions. I compute the counterfactual labor share if markdowns remained at their 1991-2002 levels in the years 2003-2014. In particular, let $\eta_q^{1991-2002}$ denote the estimated wage markdown for productivity quartile q for the 1991-2002 period, and $\eta_q^{2003-2014}$ the estimated wage markdown for 2003-2014. The counterfactual labor share is given by

$$\text{Agg } \widetilde{\text{LShare}}\ (\widehat{XLR}) = \frac{\sum_j \widehat{XLR}_{j,t} \times \eta_{q(j,t)}^{1991-2002} \times \left(\eta_{q(j,t)}^{2003-2014}\right)^{-1}}{\sum_j VA_{j,t}} \tag{1.33}$$

It's important to be clear about what this counterfactual represents and what it does not. Similar to my prior exercise in section 1.5.1, this counterfactual asks the question: "Holding the distribution of workers across firms constant, how would the observed aggregate labor share change in the 2003-2014 period if enough final output was reallocated to workers to keep the gap between wages and marginal products constant at 1991-2002 levels?"

I compute Agg $\widetilde{\text{LShare }(}LABEX)_t$ analogously. I then compute the counterfactual log change in the average labor share across the two periods. Panel B of Table 1.9 shows the results. In the first two columns I show the actual average labor shares for the two measures, and in the third column (labeled $\widetilde{2003-2014}$) I give the counterfactual aggregate labor share. The first row shows that if markdowns had held constant at their 1991-2002 levels in the latter period, the average labor share Agg $\widetilde{\text{LShare}}$ $(\widehat{XLR})_t$ would have only declined from 0.579 to 0.554 instead of the observed level of 0.524. The actual log change, in the column Δ log(Lshare), was -0.10, whereas the counterfactual log change ($\Delta\, \widetilde{\log(\text{Lshare})}$) would have been -.044. This implies that about 56% percent of the decline ($\frac{0.10-0.044}{0.10} \approx 0.558$) can be explained by the change in wage markdowns. In the second row of panel B I compute the same counterfactual for $\widetilde{\text{Agg LShare}}\ (LABEX)_t$. Because the labor share decline from this measure is less stark to begin with, the counterfactual actual predicts an *increase* in the labor share of about 1.8% instead of the observed 4.2% decrease.

In either case, I estimate that the change in wage markdowns can explain the bulk of the decline in average labor shares over the two periods. The results using my preferred measure $\widetilde{\text{Agg LShare}}\ (\widehat{XLR})_t$ suggest that there is still some room for other proposed factors, such as increased price markups or labor-substituting technological change, to explain the remaining 44%.

1.6 Discussion

In this section I discuss potential causes of the rise in labor market power, economic and policy implications of my findings, and how my results can be distinguished with alternative explanations.

1.6.1 What Has Caused the Rise In Labor Market Power?

I uncover two key features of the rise in labor market power. I document the rising spread in labor productivity in Figure 1-1, and similar patterns have been documented in numerous other papers.[23] The rise in productivity dispersion has led highly productive firms to mark wages down by more relative to low productivity firms. At the same time, in appendix Table 1.15 I find decreases in the supply elasticities for firms of all productivity levels between the 1991-2002 and 2003-2014 periods. This also has the tendency to push wages down.

The rise in productivity dispersion could be driven by a number of phenomena, such as changing economies of scale arising from the shift towards intangible capital and information technology, increased global competition which causes firms to exit, larger barriers to entry,

[23] See Autor et al. (2020), Hartman-Glaser et al. (2019), De Loecker et al. (2020), Kehrig and Vincent (2021), and Gouin-Bonenfant (2020) for example.

or shifts in consumer taste towards specific brands or products. Whatever the causes, the dispersion in productivity allows productive firms to operate on a more inelastic portion of their labor supply curve, generating monopsony rents in equilibrium.

The overall decrease in supply elasticities may be caused by very different economic forces than the cross-sectional spread in elasticities. On a fundamental level, the supply elasticity a firm faces captures the willingness or ability of workers to substitute away from that firm and towards other firms. This suggests that workers have found it more difficult to substitute away from particular employers. Consistent with this Molloy, Smith, Trezzi, and Wozniak (2016) find that the US labor market has become less dynamic over the past several decades, with transitions between firms to and from employment being less frequent. Similarly, in appendix Table 1.15 I show that the average firm-level separations rate has declined. My calibration for the 1991-2002 and 2003-2014 periods suggests reduced outside options (model parameter b) and bargaining power (parameter β). Song, Price, Guvenen, Bloom, and von Wachter (2018) show that the 1990-2015 period has seen increased sorting of productive workers to high wage firms; I find a similar result in appendix Figure 1-9, which shows that the spread in average worker skill between productive and unproductive firms has widened, especially in the latter half of my sample period. Skill-biased technological change has induced an increasing share of individuals to delay entry into the workforce in favor of obtaining higher education college, potentially resulting in more specialized human capital in the process. Appendix Table 1.15 shows that the wage premium for incumbent workers has increased, suggesting that firms have also found it more difficult to substitute workers from outside the firm.

All these patterns are consistent with specific human capital rising in importance. There are two potential opposing effects from increased human capital specificity: 1) employers find it harder to replace skilled workers, which tends to raise the wages of incumbent workers; and 2) workers are less able (or less willing) to substitute their labor across particular employers, which leads to a larger gap between wages and marginal products. My model calibration and empirical estimates suggest that the second effect has dominated.

Berger et al. (2021), Rinz (2018) and Lipsius (2018) all find that local labor market concentration has been decreasing, which initially seems at odds with increasing aggregate labor market power. The above discussion could also help explain why monopsony power has increased even as local labor market concentration has decreased. Figure 1-4 and Table 1.4 show the importance of skilled workers in driving firm-level supply elasticities. The overall downward trend in supply elasticities is driven in large part by skilled workers, and the productive firms with the most labor market power also hire more skilled workers. This in turn means that local labor market concentration may be less important for determining monopsony power for skilled workers, because for them labor markets are far less local. For example, although skilled workers have lower rates of separation, Malamud and Wozniak (2012) and Amior (2020) document that conditional on moving, more educated workers make job moves that are far more geographically distant on average, and Diamond (2016) shows that educated workers tend to live in larger metropolitan areas which naturally have less local concentration. Supporting a human capital specificity explanation, Nimczik (2020) finds that more skilled workers tend to be employed in labor markets that are geographically dispersed but more concentrated in a set of particular industries.

The overall rise of non-compete agreements, which now affect a large portion of the labor force and are focused on skilled workers (Starr, Prescott, and Bishara, 2021), could also be playing a role in reducing supply elasticities. These appear to be successful in reducing both worker mobility and wages, especially among skilled workers.[24]

1.6.2 Economic and Policy Implications of Labor Market Power

My findings suggest an important subtlety to consider for policy interventions intended to curtail labor market power: in the cross-section, firms who exercise more labor market power tend to pay higher wages, not lower wages, and they are likely to be firms where many workers would want to be employed. It is precisely these firms' success in terms of productivity—an amalgam of high product demand, innovative success, productive efficiency, etc.—which grants them their monopsony rents. While I find that firms which have become productive

[24] See Garmaise (2011), Balasubramanian, Chang, Sakakibara, Sivadasan, and Starr (2020), Jeffers (2019), Johnson, Lavetti, and Lipsitz (2020), for example.

earn substantial rents from labor market power ex post, many of these firms have likely made risky investments ex ante which allowed them to become productive. The prospect of earning rents may compensate firms for undertaking risky R&D or other types of investments earlier on in their life cycle. These investments are also beneficial for workers because they generate growth and can improve wages in absolute terms, even if the gap between the marginal product of labor and wages widens. Accordingly, it is important to consider whether policy interventions geared towards reducing labor market power may deter firms from making these types of investments, or more generally act as a deterrent to potential market entrants.

Although I do not explicitly model product market power in this paper, my results also imply that labor and product market power are likely to comove. Because firms with high product market power charge high prices, part of what I measure as labor productivity (value added per worker) could in practice reflect firms' product market power. This implies that policy interventions, such as antitrust enforcement, that are meant to curb product market power could also be effective in reducing labor market power at the same time. However, it's not necessarily the case that this is guaranteed to make workers better off. For example, it's possible that breaking up productive firms could reduce productivity or devalue firm-specific human capital, leading to lower wages for workers.

Perhaps of most concern from the perspective of worker welfare is the aggregate decline in the overall supply elasticity across productivity and skill levels, which reflects workers' diminished ability to substitute employment across firms. As I argue regarding human capital specificity in the previous subsection, some of the forces that have led to this change may also have made it more costly for firms to replace existing workers. These labor adjustment costs create a deadweight loss. Consequently, policies that are geared towards reducing the costliness of hiring new workers while helping workers substitute their labor across potential employers may be mutually beneficial for both workers and firms. This could include the subsidization of new skill acquisition or easing regulatory burdens which make hiring employees costly.

Working to reduce artificial barriers to worker mobility stands to be directly beneficial to workers, and targets the types of forces generating labor market power that are more clearly negative from a welfare perspective. This could include reducing the scope of or banning entirely the use of non-compete agreements. Pursuing antitrust action to prevent employers from colluding to reduce worker mobility could also be helpful to workers. There is some precedent for such legal action already, as the US Department of Justice Antitrust Division filed a suit against a number of prominent Silicon Valley tech firms in 2010 for colluding from 2005 to 2009 to not poach employees from one another and a settlement forcing the firms to end this practice was quickly reached.[25]

Another interesting economic implication of my findings relates to the rise in wage inequality and the polarization of the US labor market driven by technological change (Acemoglu and Autor, 2011; Autor and Dorn, 2013). Since I find the lowest and most decreasing supply elasticities for skilled workers, who also have high earnings, this suggests that skilled workers' wages are marked down relative to a very high and increasing marginal product of labor relative to low skill workers. Thus my estimates imply an even larger role for technological change in generating wage inequality. However, it is also necessary to note that in the model I estimate dynamic markdown wedges for a single labor input, and not separately for different skill levels. If I added workers of different skill levels to the model the dynamic markdown wedge would be larger for the most skilled workers, who are most costly to replace, meaning that wage markdown differences between high- and low-skill workers are not as large as the simple supply elasticity difference would imply. Still, the elasticity difference is wide enough that it is highly unlikely this would reverse the of the skilled workers' lower supply elasticities entirely.

1.6.3 Alternative Explanations for My Findings

Some previous papers find similar patterns in firm-level labor shares but argue for different economic mechanisms. Here I discuss two specific examples. Hartman-Glaser et al. (2019) also focus on Compustat firms, and they similarly show that the share of output accruing to

[25] See court documents for the case "U.S. V. Adobe Systems, Inc., Et Al." found at https://www.justice.gov/atr/case/us-v-adobe-systems-inc-et-al.

capital owners has increased by the most for highly productive firms. They also show that firm level idiosyncratic risk rose substantially over the 1960-2014 period and propose a model based on optimal contracting to explain the capital share dynamics. In their model skilled workers demand wage contracts with embedded insurance for bearing firm-specific risk, and productive firms are able to allocate more output to capital. Increasing idiosyncratic risk amplifies this mechanism, leading to a drop in the aggregate labor share driven by productive firms in the right tail of the productivity distribution.

They calibrate the model to match changes in the labor share between the 1960-1970 and 1990-2014 periods. In appendix Figure 1-10 I plot time trends in their idiosyncratic risk measures before and during my sample period. Although firm-specific risk has indeed risen over that time horizon, idiosyncratic risk has actually been flat or slightly declining over the 1991-2014 period that my data covers. Therefore while their mechanism does speak to broad patterns in labor shares over a longer time horizon, it is incomplete as an explaination for the change in the aggregate labor share *within* the 1991-2014 period that is the focus of this paper.

Kehrig and Vincent (2021) document that large and productive manufacturing firms with low labor shares have driven the aggregate decline in the manufacturing sector labor share. They argue that this is more consistent with demand side forces than with wage markdowns because in their framework monopsony should primarily depress labor shares by reducing wages. However, I find that monopsony power is larger for productive firms precisely because they can pay high wages, which allows them to operate on a relatively inelastic portion of their labor supply curve. From the perspective of my findings, the demand side forces they find exaggerate productivity advantages and increase the spread in labor market power. Thus the results in this paper and those in Kehrig and Vincent (2021) are compatible explanations for the aggregate labor share decline.

Finally, I stress that my findings still leave plenty of room for other forces, such as increasing price markups or technological change, to explain cross-sectional differences and time series changes in labor shares, and these could also speak to the evidence on firm profitability, valuations, investment, and economic rents that I presented in section 1.2. I do argue, however, that labor market power has played a more important role in these phenomena than prior literature may suggest.

Monopsony Through Upward Sloping Labor Supply Curves or Wage Bargaining?

Models of rent-sharing via wage bargaining can also explain passthrough of firm-specific productivity shocks to wages, even without upward sloping firm-specific labor supply curves. A few patterns suggest monopsony-based explanations over a bargaining. For example, in most models of monopsony a shock to the wages of a competitor acts like a firm-specific labor supply shock to the firm. The result is that the firm must raise wages in order to retain workers, meaning competitor wage growth should be expected to raise wages without raising employment at the firm. This is exactly what I find in appendix Table 1.16. In the fourth column of this table I find that competitor wage growth is indeed strongly associated with own firm wage growth, but has no relation to employment growth. In Table 1.3 I find that skilled recruits also face lower supply elasticities compared to lower skilled recruits, suggesting that the low supply elasticity caused by the wage response is related to monopsony power rather than, say, rent sharing due to bargaining with skilled insiders who have a better bargaining position. Finally, my simple dynamic wage posting monopsony model with labor adjustment adjustment costs immediately can explain empirical patterns in recruit and incumbent wage passthroughs and supply elasticities; sorting patterns in supply elasticities across firms of different productivity types; and cross-sectional differences in firm level labor shares and profitability. Purely bargain based explanations would have a difficult time matching all the above patterns simultaneously.

1.7 Conclusion

In this paper I find evidence for a "superstar firms" view of labor market power: firms with productivity advantages face much lower supply elasticities and hence earn higher monopsony rents from wage markdowns. Differences in supply elasticities have widened over time, leading to an increased gap in labor shares between productive and unproductive firms in the cross section and contributing to the decline in the aggregate labor share. The value of wage markdowns is substantial. The average firm earns about a third of its capital income from wage markdowns, and wage markdowns are worth about 40% of capital income in aggregate. Consistent with economic rents earned from labor market power, productive firms have higher valuation ratios but not significantly higher investment or hiring rates.

Although my findings suggest that labor market power generates substantial value to productive firms in the form of economic rents, my results suggest some caution in interpreting welfare or policy implications. In my model and in the data, productive firms actually pay higher wages despite having larger markdowns, since they operate on a steeper portion of the labor supply curve. Because of the opportunities they afford their employees–especially in the form of higher wages—highly productive firms are likely to be coveted employers. Any policy recommendations to curb labor market power ought to consider that it is exerted by the most economically successful firms who actually tend to pay above average wages to their employees.

Skilled workers play an important role in creating these patterns, as they face the lowest supply elasticities of any skill group. There is some evidence in increased role for firm-specific human capital could be driving these patterns, but artificial barriers to mobility—such as the rising prevalence of non-compete agreements—may also play a role. Exploring these possibilities in more depth could be a particularly useful avenue for further research.

Bibliography

Abel, A. B. and J. C. Eberly (2011, 03). How Q and Cash Flow Affect Investment without Frictions: An Analytic Explanation. *The Review of Economic Studies 78*(4), 1179–1200.

Abowd, J. A. and T. Lemieux (1993). The effects of product market competition on collective bargaining agreements: The case of foreign competition in canada. *The Quarterly Journal of Economics 108*(4), 983–1014.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Abowd, J. M., B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. Woodcock (2009). *The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators.* University of Chicago Press.

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics, Volume 4. Amsterdam: Elsevier-North*, pp. 1043–1171.

Amior, M. (2020). Education and geographical mobility: The role of the job surplus.

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2020, 02). The Fall of the Labor Share and the Rise of Superstar Firms. *The Quarterly Journal of Economics 135*(2), 645–709.

Autor, D. H. and D. Dorn (2013, August). The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review 103*(5), 1553–97.

Balasubramanian, N., J. W. Chang, M. Sakakibara, J. Sivadasan, and E. Starr (2020). Locked in? the enforceability of covenants not to compete and the careers of high-tech workers. *Journal of Human Resources.*

Balke, N. and T. Lamadon (2020, November). Productivity shocks, long-term contracts and earnings dynamics. Working Paper 28060, National Bureau of Economic Research.

Barkai, S. (2020). Declining labor and capital shares. *The Journal of Finance 75*(5), 2421–2463.

Bassier, I., A. Dube, and S. Naidu (2020, August). Monopsony in movers: The elasticity of labor supply to firm wage policies. Working Paper 27755, National Bureau of Economic Research.

Belo, F., J. Li, X. Lin, and X. Zhao (2017, 07). Labor-Force Heterogeneity and Asset Prices: The Importance of Skilled Labor. *The Review of Financial Studies 30*(10), 3669–3709.

Belo, F., X. Lin, and S. Bazdresch (2014). Labor hiring, investment, and stock return predictability in the cross section. *Journal of Political Economy 122*(1), 129–177.

Benmelech, E., N. Bergman, and H. Kim (2018, February). Strong employers and weak employees: How does employer concentration affect wages? Working Paper 24307, National Bureau of Economic Research.

Berger, D. W., K. F. Herkenhoff, and S. Mongey (2021, May). Labor market power. Working Paper 25719, National Bureau of Economic Research.

Bloom, N., M. Schankerman, and J. Van Reenen (2013). Identifying technology spillovers and product market rivalry. *Econometrica 81*(4), 1347–1393.

Boudoukh, J., R. Feldman, S. Kogan, and M. Richardson (2018, 07). Information, Trading, and Volatility: Evidence from Firm-Specific News. *The Review of Financial Studies 32*(3), 992–1033.

Brardsen, G. and H. Lütkepohl (2011). Forecasting levels of log variables in vector autoregressions. *International Journal of Forecasting 27*(4), 1108–1115.

Caldwell, S. and E. Oehlsen (2018). Monopsony and the gender wage gap: Experimental evidence from the gig economy.

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018). Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics 36*(S1), S13–S70.

Card, D., J. Heining, and P. Kline (2013, 05). Workplace Heterogeneity and the Rise of West German Wage Inequality. *The Quarterly Journal of Economics 128*(3), 967–1015.

Chan, M., S. Salgado, and M. Xu (2021). Heterogeneous passthrough from tfp to wages.

Cho, D. (2018). The labor market effects of demand shocks: Firm-level evidence from the recovery act.

Cohen, L. and A. Frazzini (2008). Economic links and predictable returns. *The Journal of Finance 63*(4), 1977–2011.

Corhay, A., H. Kung, and L. Schmid (2020). Q: Risk, rents, or growth?

Covarrubias, M., G. Gutiérrez, and T. Philippon (2020). From good to bad concentration? us industries over the past 30 years. *NBER Macroeconomics Annual 34*, 1–46.

Crouzet, N. and J. Eberly (2021). Rents and intangible capital: A q+ framework.

Daniel, K., D. Hirshleifer, and L. Sun (2019, 06). Short- and Long-Horizon Behavioral Factors. *The Review of Financial Studies 33*(4), 1673–1736.

Davis, S. J., J. Haltiwanger, R. Jarmin, J. Miranda, C. Foote, and v. Nagypál (2006). Volatility and dispersion in business growth rates: Publicly traded versus privately held firms [with comments and discussion]. *NBER Macroeconomics Annual 21*, 107–179.

De Loecker, J., J. Eeckhout, and G. Unger (2020, 01). The Rise of Market Power and the Macroeconomic Implications. *The Quarterly Journal of Economics 135*(2), 561–644.

Diamond, R. (2016, March). The determinants and welfare implications of us workers' diverging location choices by skill: 1980-2000. *American Economic Review 106*(3), 479–524.

Donangelo, A. (2014). Labor mobility: Implications for asset pricing. *The Journal of Finance 69*(3), 1321–1346.

Donangelo, A. (2020, 02). Untangling the Value Premium with Labor Shares. *The Review of Financial Studies 34*(1), 451–508.

Donangelo, A., F. Gourio, M. Kehrig, and M. Palacios (2019). The cross-section of labor leverage and equity returns. *Journal of Financial Economics 132*(2), 497–518.

Dube, A., A. Manning, and S. Naidu (2018, September). Monopsony and employer misoptimization explain why wages bunch at round numbers. Working Paper 24991, National Bureau of Economic Research.

Eisfeldt, A., A. Falato, and M. Xiaolan (2021). Human capitalists.

Eisfeldt, A. L. and D. Papanikoloau (2013). Organization capital and the cross-section of expected returns. *The Journal of Finance 68*(4), 1365–1406.

Engbom, N. and C. Moser (2020). Firm pay dynamics.

Ewens, M. and M. Marx (2017). Founder replacement and startup performance. *The Review of Financial Studies 31*(4), 1532–1565.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Farhi, E. and F. Gourio (2018, November). Accounting for macro-finance trends: Market power, intangibles, and risk premia. Working Paper 25282, National Bureau of Economic Research.

Friedrich, B., L. Laun, C. Meghir, and L. Pistaferri (2019, April). Earnings dynamics and firm-level shocks. Working Paper 25786, National Bureau of Economic Research.

Garin, A. and F. Silverio (2020). How responsive are wages to demand within the firm? evidence from idiosyncratic export demand shocks.

Garmaise, M. J. (2011). Ties that truly bind: Noncompetition agreements, executive compensation, and firm investment. *Journal of Law, Economics, and Organization 27*(2), 376–425.

Gouin-Bonenfant, E. (2020). Productivity dispersion, between-firm competition, and the labor share.

Greenwald, D., M. Lettau, and S. Ludvigson (2021). How the wealth was won: Factor shares as market fundamentals.

Grullon, G., Y. Larkin, and R. Michaely (2019, 04). Are US Industries Becoming More Concentrated?*. *Review of Finance 23*(4), 697–743.

Guiso, L., L. Pistaferri, and F. Schivardi (2005). Insurance within the firm. *Journal of Political Economy 113*(5), 1054–1087.

Gutierrez, G. and T. Philippon (2017). Investmentless growth: An empirical investigation. *Brookings Papers on Economic Activity*, 89–169.

Hartman-Glaser, B., H. Lustig, and M. Z. Xioalan (2019). Capital share dynamics when firms insure workers. *The Journal of Finance 74*(4), 1707–1751.

Herschbein, B., C. Macaluso, and C. Yeh (2020). Monopsony in the us labor market.

Jager, S. and J. Heining (2019). How substitutable are workers? evidence from worker deaths.

Jarosch, G., J. S. Nimczik, and I. Sorkin (2019, September). Granular search, market structure, and wages. Working Paper 26239, National Bureau of Economic Research.

Jeffers, J. (2019). The impact of restricting labor mobility on corporate investment and entrepreneurship.

Johnson, M. S., K. Lavetti, and M. Lipsitz (2020). The labor market effects of legal restrictions on worker mobility.

Karabarbounis, L. and B. Neiman (2013, 10). The Global Decline of the Labor Share*. *The Quarterly Journal of Economics 129*(1), 61–103.

Kehrig, M. and N. Vincent (2021, 03). The Micro-Level Anatomy of the Labor Share Decline. *The Quarterly Journal of Economics 136*(2), 1031–1087.

Kim, H. (2020). How does labor market size affect firm capital structure? evidence from large plant openings. *Journal of Financial Economics 138*(1), 277–294.

Kline, P., N. Petkova, H. Williams, and O. Zidar (2019, 03). Who Profits from Patents? Rent-Sharing at Innovative Firms. *The Quarterly Journal of Economics 134*(3), 1343–1404.

Knepper, M. (2020, 03). From the fringe to the fore: Labor unions and employee compensation. *The Review of Economics and Statistics 102*(1), 98–112.

Kogan, L., D. Papanikolaou, L. D. W. Schmidt, and J. Song (2020, April). Technological innovation and labor income risk. Working Paper 26964, National Bureau of Economic Research.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017, 03). Technological Innovation, Resource Allocation, and Growth*. *The Quarterly Journal of Economics 132*(2), 665–712.

Kroft, K., Y. Luo, M. Mogstad, and B. Setzler (2020, June). Imperfect competition and rents in labor and product markets: The case of the construction industry. Working Paper 27325, National Bureau of Economic Research.

Kuehn, L.-A., M. Simutin, and J. J. Wang (2017). A labor capital asset pricing model. *The Journal of Finance 72*(5), 2131–2178.

Lachowska, M., A. Mas, R. D. Saggio, and S. A. Woodbury (2020, January). Do firm effects drift? evidence from washington administrative data. Working Paper 26653, National Bureau of Economic Research.

Lamadon, T., M. Mogstad, and B. Setzler (2019, June). Imperfect competition, compensating differentials and rent sharing in the u.s. labor market. Working Paper 25954, National Bureau of Economic Research.

Lee, D. S. and A. Mas (2012, 01). Long-run impacts of unions on firms: New evidence from financial markets, 1961–1999. *The Quarterly Journal of Economics 127*(1), 333–378.

Lipsius, B. (2018). Labor market concentration does not explain the falling labor share.

Liu, Y. (2019). Labor based asset pricing.

Lucking, B., N. Bloom, and J. Van Reenen (2019). Have r&d spillovers declined in the 21st century? *Fiscal Studies 40*(4), 561–590.

Malamud, O. and A. Wozniak (2012). The impact of college on migration: Evidence from the vietnam generation. *The Journal of Human Resources 47*(4), 913–950.

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

Matsa, D. A. (2010). Capital structure as a strategic variable: Evidence from collective bargaining. *The Journal of Finance 65*(3), 1197–1232.

McFadden, D. L. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics*, pp. 105–42. Academic Press: New York.

Molloy, R., C. L. Smith, R. Trezzi, and A. Wozniak (2016). Understanding declining fluidity in the u.s. labor market. *Brookings Papers on Economic Activity*, 183–237.

İmrohoroğlu, A. and S. Tüzel (2014). Firm-level productivity, risk, and return. *Management Science 60*(8), 2073–2090.

Mueller, H. M., P. P. Ouimet, and E. Simintzi (2017, 05). Within-Firm Pay Inequality. *The Review of Financial Studies 30*(10), 3605–3635.

Neuhierl, A., A. Scherbina, and B. Schiusene (2013). Market reaction to corporate press releases. *The Journal of Financial and Quantitative Analysis 48*(4), 1207–1240.

Nimczik, J. (2020). Job mobility networks and data-driven labor markets.

Peters, R. H. and L. A. Taylor (2017). Intangible capital and the investment-q relation. *Journal of Financial Economics 123*(2), 251–272.

Ransom, M. R. and D. P. Sims (2010). Estimating the firm's labor supply curve in a "new monopsony" framework: Schoolteachers in missouri. *Journal of Labor Economics 28*(2), 331–355.

Rinz, K. (2018). Labor market concentration, earnings inequality, and earnings mobility.

Robinson, J. (1969). *The Economics of Imperfect Competition*. Macmillan.

Rouwenhourst, K. G. (1995). Asset pricing implications of equilibrium business cycle models. In *Frontiers of Business Cycle Research*, pp. 294–330. Princeton Univesity Press.

Schubert, G., A. Stansbury, and B. Taska (2020). Employer concentration and outside options.

Shen, M. (2021, 02). Skilled Labor Mobility and Firm Value: Evidence from Green Card Allocations. *The Review of Financial Studies*.

Sokolova, A. and T. Sorensen (2021). Monopsony in labor markets: A meta-analysis. *ILR Review 74*(1), 27–55.

Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. von Wachter (2018, 10). Firming Up Inequality. *The Quarterly Journal of Economics 134*(1), 1–50.

Sorkin, I. (2018, 01). Ranking Firms Using Revealed Preference. *The Quarterly Journal of Economics 133*(3), 1331–1393.

Stansbury, A. and L. H. Summers (2020). The declining worker power hypothesis: An explanation for the recent evolution of the american economy. *Brookings Papers on Economic Activity*, 1–77.

Starr, E. P., J. Prescott, and N. D. Bishara (2021). Noncompete agreements in the us labor force. *The Journal of Law and Economics 64*(1), 53–84.

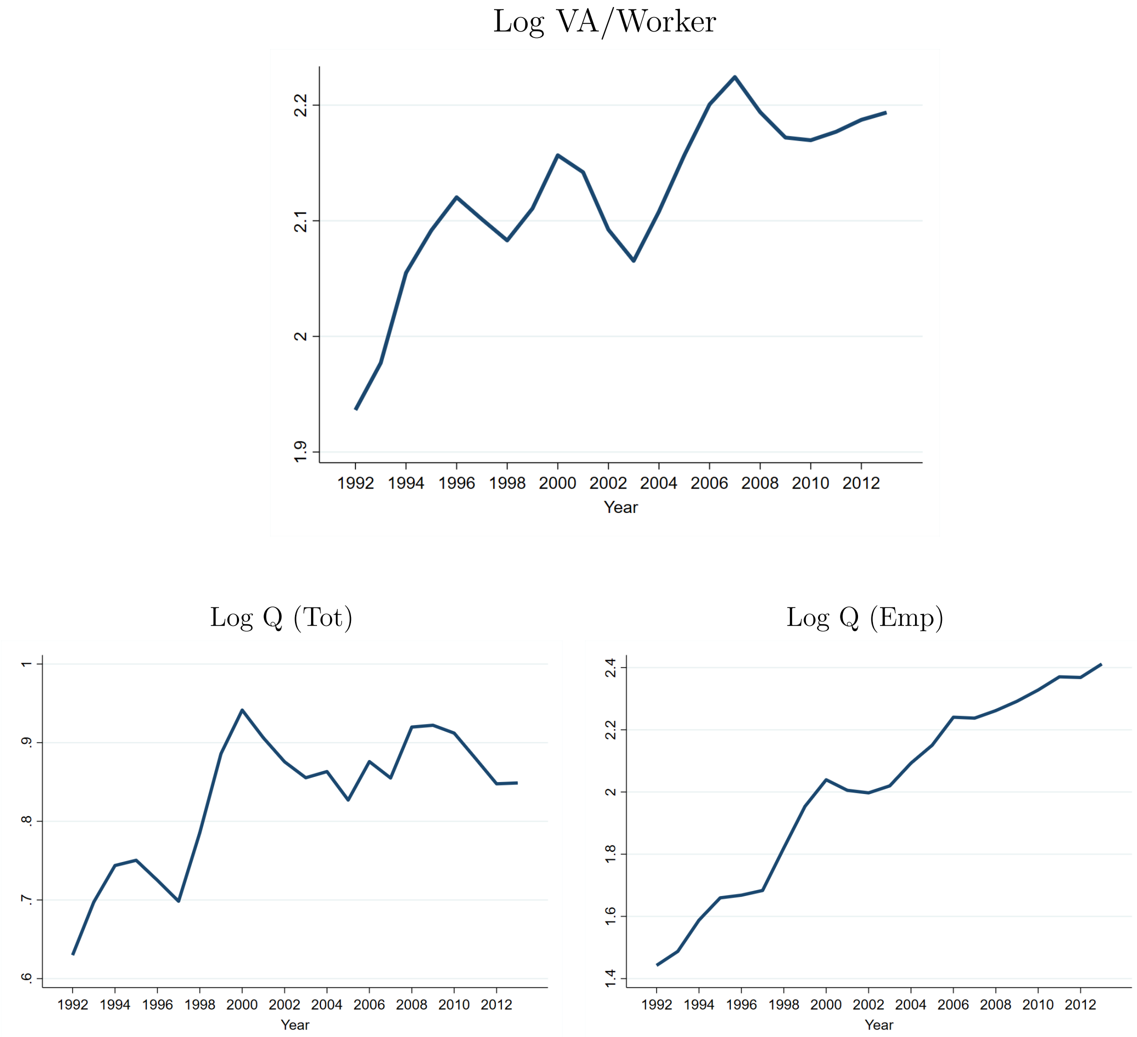
Van Reenen, J. (1996, 02). The Creation and Capture of Rents: Wages and Innovation in a Panel of U.K. Companies. *The Quarterly Journal of Economics 111*(1), 195–226.

Vasicek, O. A. (1973). A note on using cross-sectional information in bayesian estimation of security betas. *The Journal of Finance 28*(5), 1233–1239.

Vuolteenaho, T. (2002). What drives firm-level stock returns? *The Journal of Finance 57*(1), 233–264.

Figures

Figure 1-1: High- minus Low-Labor Productivity Firm Spreads in Average Log VA/Worker, AKM Worker Skill, Log Q (Total), and Log Q (Emp) Trend Upward Over Time

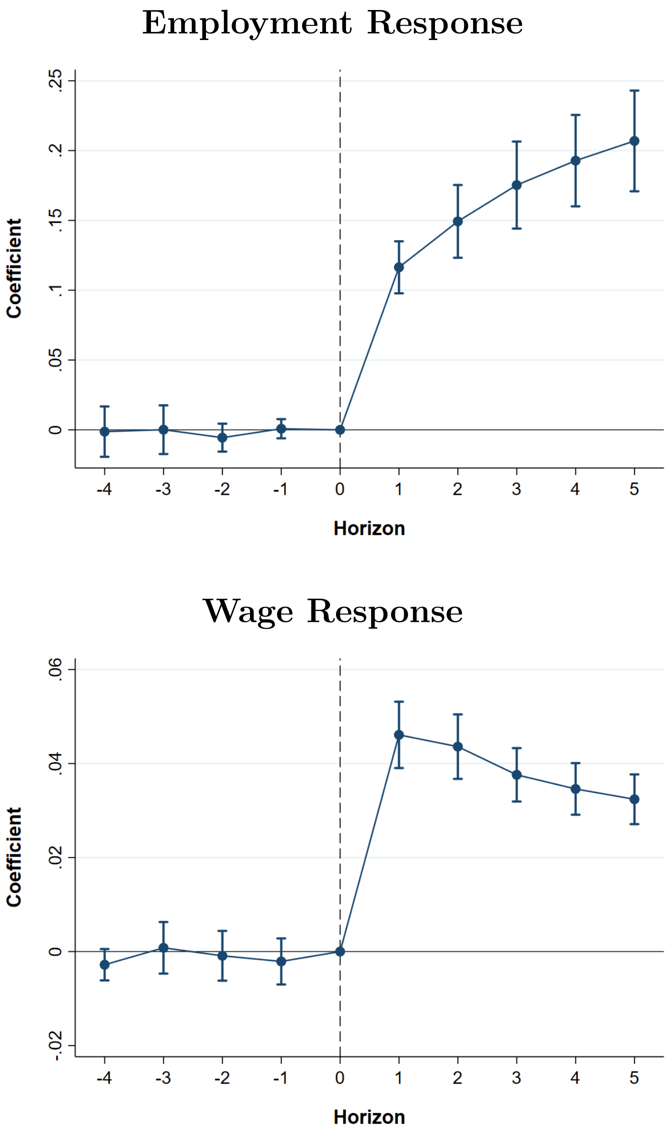

Log VA/Worker

Log Q (Tot)

Log Q (Emp)

Note: This figure shows the differences between firms in the top- and bottom-quartiles of productivity for the average log value-added per worker; firm worker skill level from (1.41) in the text; log total Tobin's Q ratio from Peters and Taylor (2017); and log employment-based Q ratio from (1.42) . Sample period spans 1991-2014.

Figure 1-2: Employment and Wage Responses to a Stock Return Shock

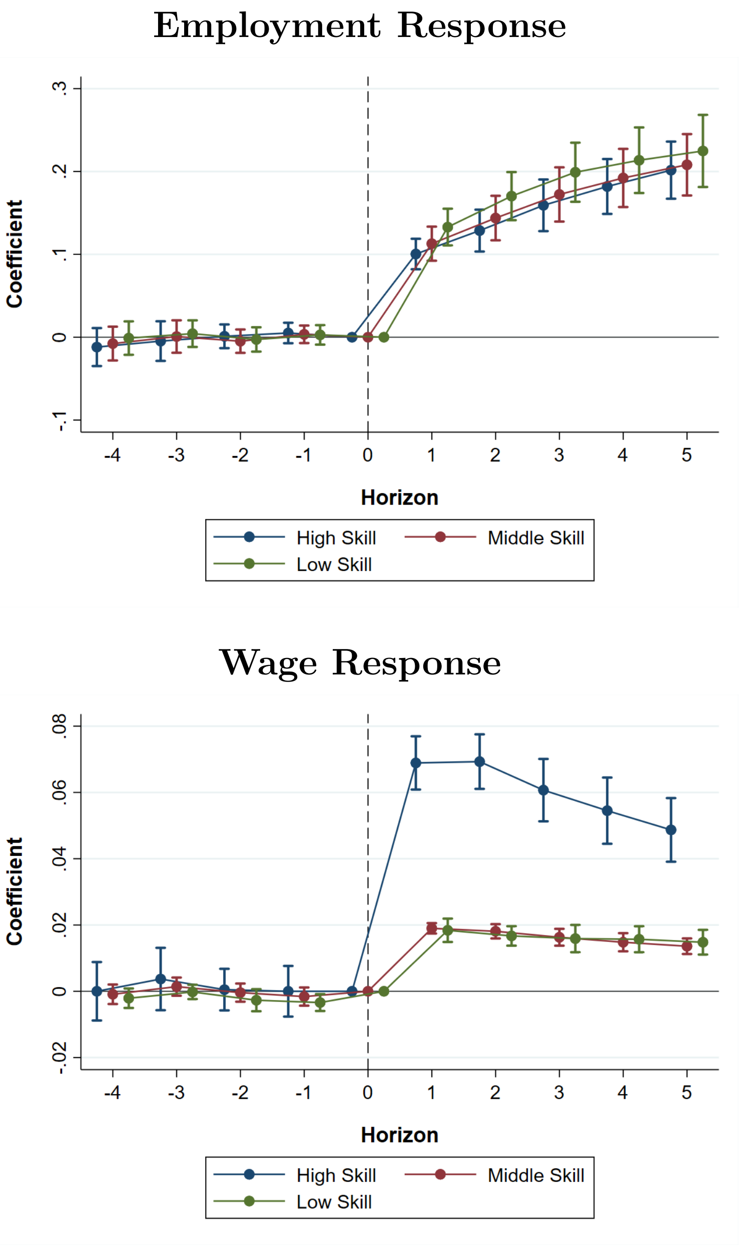


Employment Response

Wage Response

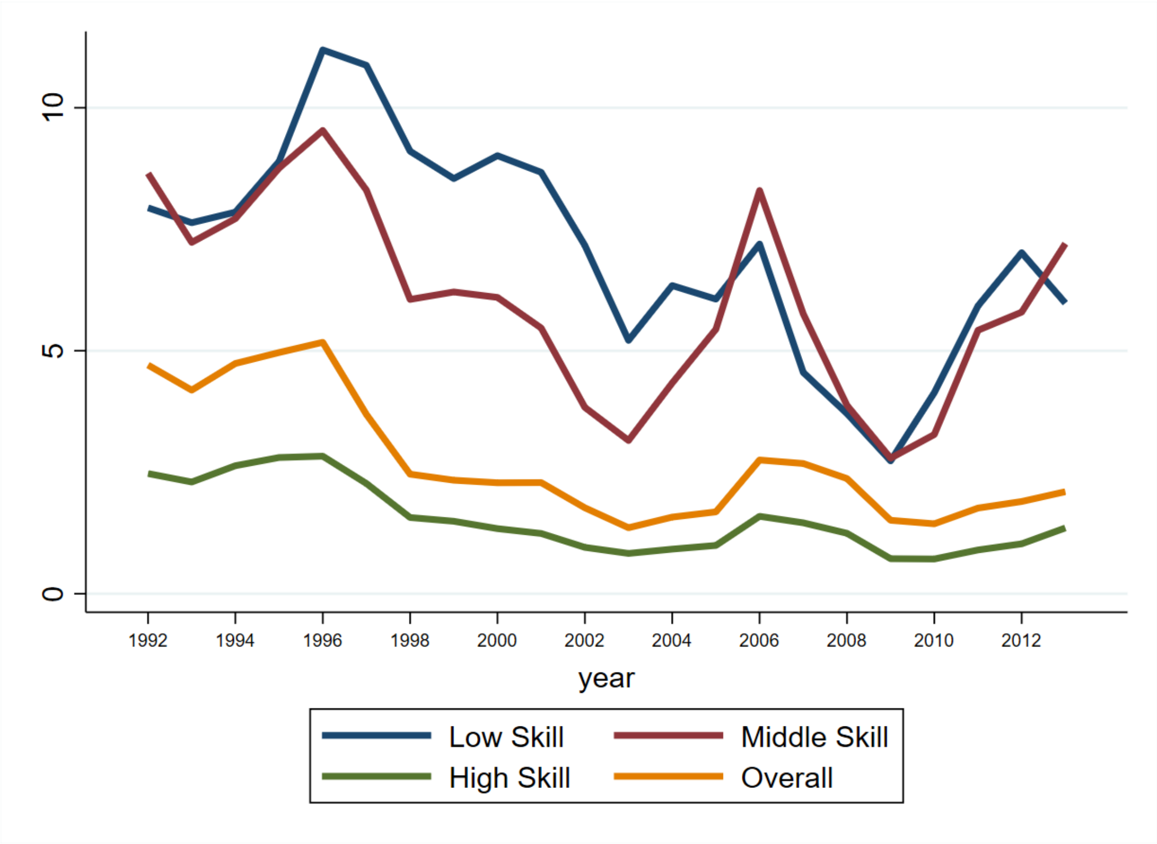
Note: This figures shows the employment and wage responses from estimating (1.7) in the main text for $h = -5$ to 5 years. Confidence intervals are based off standard errors double clustered by industry and year.

Figure 1-3: Employment and Wage Responses to Stock Return Shock, by Worker Skill



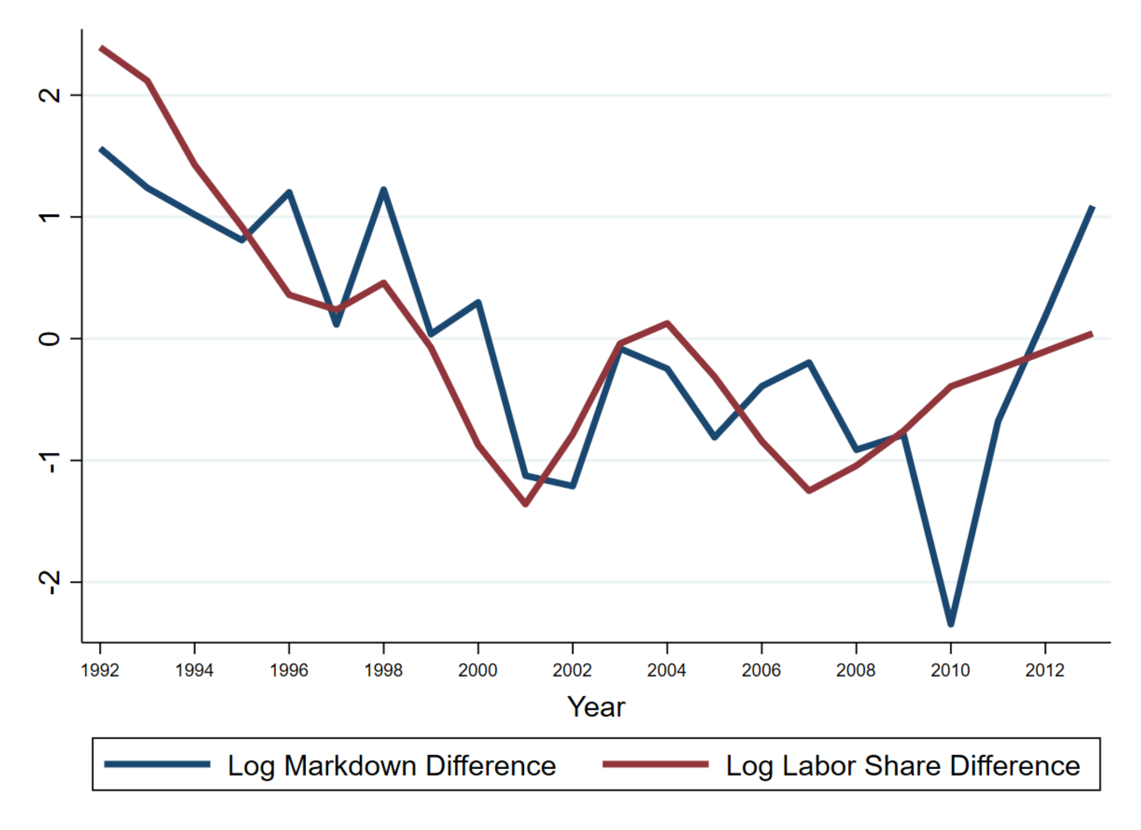
Note: This figures shows the employment and wage responses from estimating (1.7) in the main text for $h = -5$ to 5 years for workers of different skill levels. Individuals in the bottom two quintiles of the cross-sectional distribution of worker effects are considered low-skilled, the third and fourth quintiles middle-skilled, and the top quintile high-skilled. Confidence intervals are based off standard errors double clustered by industry and year.

Figure 1-4: Estimated Supply Elasticities Trend Downward Over Time For All Skill Groups



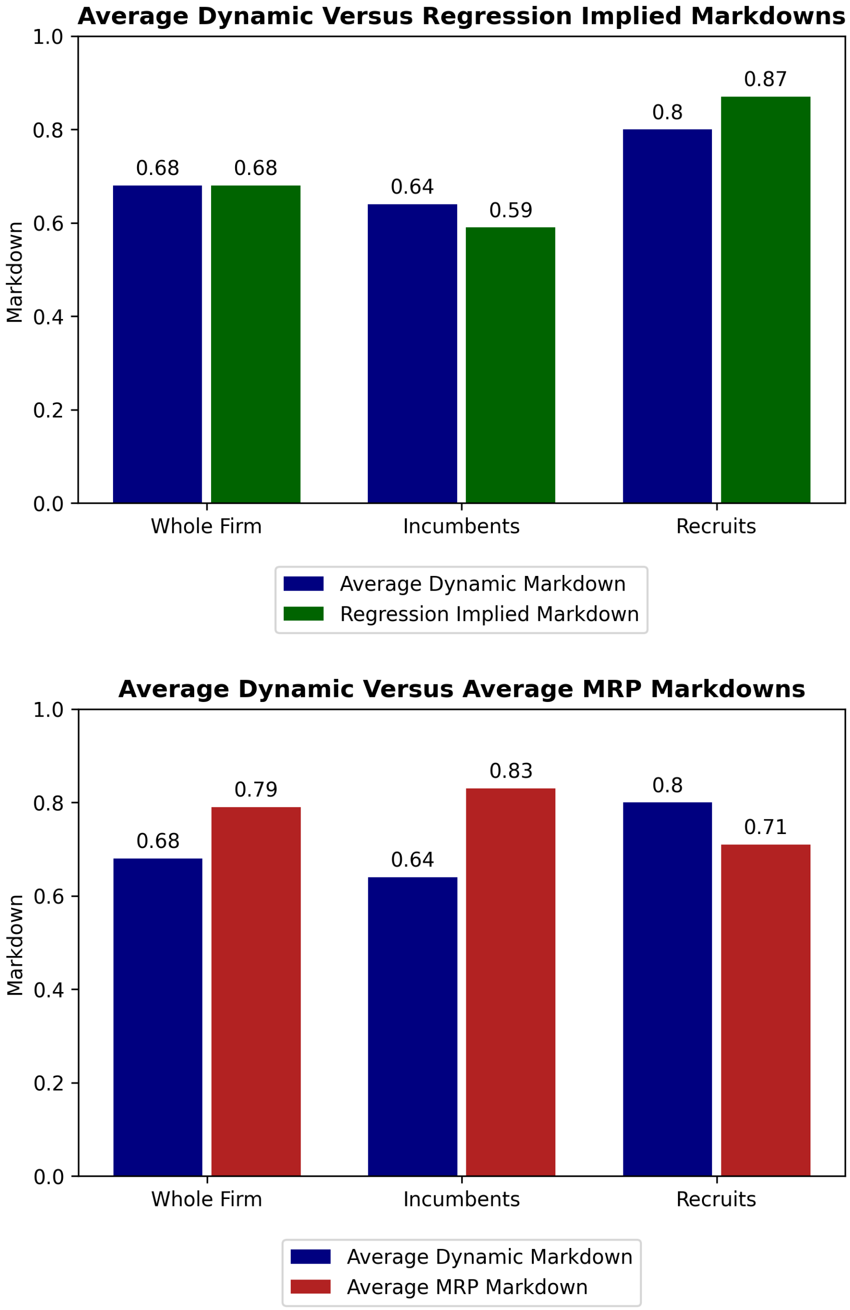
Note: This figure shows estimates of supply elasticities implied by estimating employment and wage responses from (1.6) for workers of different skill levels and for overlapping moving 3-year windows. The elasticity for a given year is for the moving window centered at that year (except for the start and end years, which are respectively the first or last year of 2-year window). The sample spans 1992-2013. High skill workers are in the top quintile of individual fixed effects from a wage of individual earnings into worker- and firm-specific components; low skill workers are in the bottom two quintiles and middle skill the third and fourth quintiles.

Figure 1-5: The Spread in Labor Shares and Estimated Markdowns Between High- and Low-Labor Productivity Firms Trends Downward Over Time



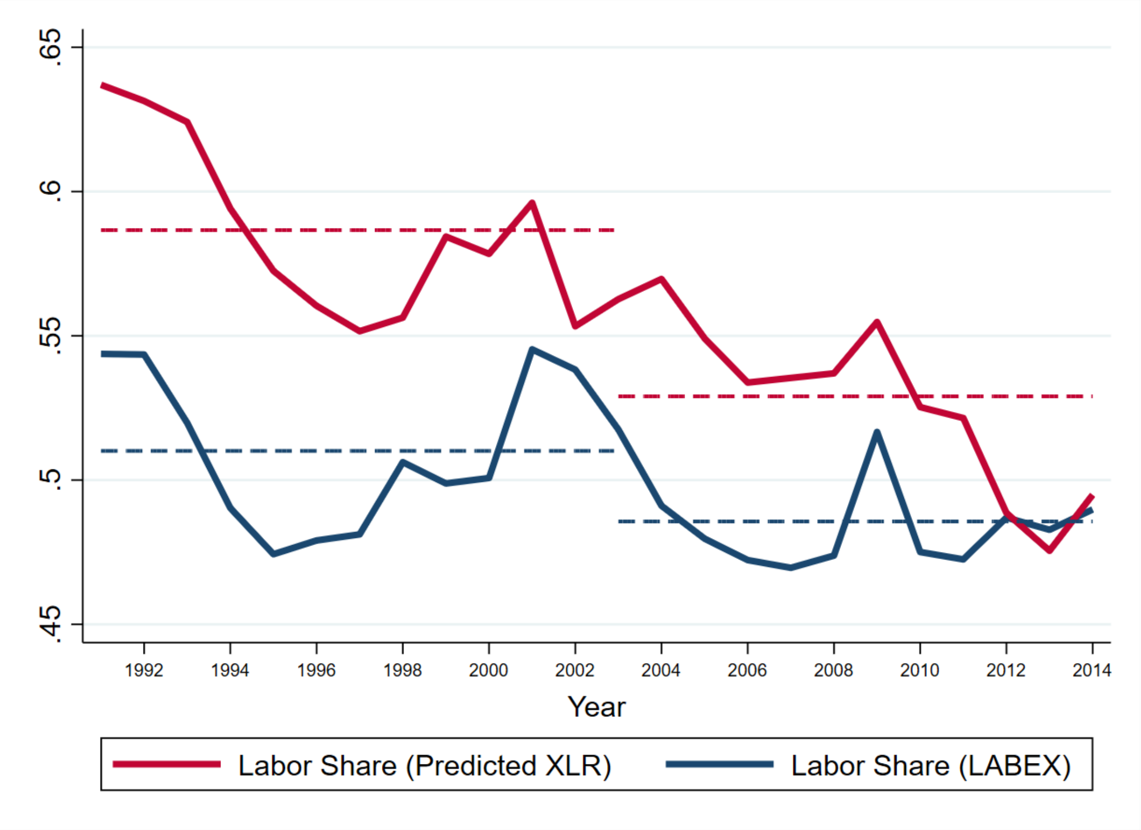
Note: This figure shows differences elasticity implied log markdowns between top- and bottom-quartile labor productivity firms from (1.9) in the main text, and log labor share differences following (1.10). The elasticity for a given year is for the moving window centered at that year (except for the start and end years, which are respectively the first or last year of 2-year window). Both series are standardized to unit standard deviation and zero mean. The sample period spans 1992-2013. The two series have a correlation of 0.73.

Figure 1-6: Model Implied Markdowns



Note: This figure shows markdowns implied by the calibration of the model. The dynamic markdown is the markdown adjusted for model dynamics due to adjustment costs, and is given by (1.20) in the main text. The regression implied markdown is markdown implied by estimating the supply elasticity by running the regression of model employment/wage growth on stock returns and taking the ratio of the two coefficients. The MRP markdown is the markdown from marginal revenue product. See section 3.1 in text for details.

Figure 1-7: Aggregate Labor Shares Over Time



Note: This figure shows the aggregate labor shares for firms in my Compustat-LEHD matched sample. The dashed lines give the mean aggregate labor share for the given measure over the 1991-2002 and 2003-2014 subperiods. I report aggregate labor shares using two measures of firm labor expenses. Labor Share (LABEX) uses my LEHD-based measure of firm labor expenses, while Labor Share (Predicted XLR) is constructed by imputing Compustat staff and labor expenses. I describe my imputation of XLR in section 1.7.1 of the appendix. Labor shares are computed as the ratio of total labor expenses to total value-added, where I define valued-added following Donangelo et al. (2019). See section 1.5.2 in main text for further details.

Tables

Table 1.1: Productive Firms Have Lower Labor Shares, Higher Valuations, and Better Operating Performance

Panel A: Log Labor Share

log VA/Worker	-0.343	-0.389	-0.514	-0.555
	(0.050)	(0.050)	(0.025)	(0.036)
Size Controls		X		X
Industry X Year FE			X	X
N	57500	57500	57500	57500
R^2 (within)	0.297	0.335	0.458	0.476

Panel B: Log Valuation Ratios

	log Q (Tot)	log Q (Emp)	log Mkcap/Book	log Mkcap/Sales
log VA/Worker	0.524	0.540	0.254	0.412
	(0.048)	(0.060)	(0.029)	(0.033)
Size Controls	X	X	X	X
Industry X Year FE	X	X	X	X
N	48500	47500	55000	57000
R^2 (within)	0.125	0.276	0.069	0.315

Panel C: Operating Performance

	ROA_t	ROA_{t+1}	ROE_t	ROE_{t+1}
log VA/Worker	0.107	0.098	0.277	0.2456
	(0.006)	(0.006)	(0.016)	0.0221
Size Controls	X	X	X	X
Industry X Year FE	X	X	X	X
N	57500	53500	55500	51500
R^2 (within)	0.168	0.128	0.085	.0665

Note: This table displays coefficients on log value added per worker for predicting firm log labor shares in panel A; firm log valuation ratios in panel B; and current and future firm operating performance in panel C. Controls for size include log employment, assets, and sales. Log Q (Tot) is the log of intangible-adjusted total Q from Peters and Taylor (2017). Log Q (Emp) is an employment-based analogue of Tobin's Q, where the denominator is the skill-weighted workforce of the firm. ROA is the return on assets (income before extraordinary items plus deprecation over total assets) and ROE is return on equity (income before extraordinary items plus deprecation over book equity). Operating performance and valuation ratios are winsorized at the 1% level by year. See section 1.2 in main text for more details. Standard errors double clustered by year and industry in parentheses.

Table 1.2: Average Investment Rates/Tobin's Q and Investment-Q Relation for Firms Sorted on Labor Productivity

Panel A: Investment Rates and Tobin's Q

Productivity:	Quartile 1	Quartile 2	Quartile 3	Quartile 4	P-val 4-1
Inv Rate (Total)	0.18	0.16	0.16	0.20	0.33
Q (Total)	0.73	0.79	1.04	1.70	0.00
Hiring Rate	0.40	0.33	0.29	0.29	0.00
Q (Emp)	0.78	0.97	1.67	4.83	0.00

Panel B: Investment-Q Relation

Productivity:	Quartile 1	Quartile 2	Quartile 3	Quartile 4	P-val 4-1	R-sq (Within)	N
			Dep Variable: Investment Rate (Total)				
Q (Tot)	0.039	0.030	0.026	0.022			
	(0.004)	(0.003)	(0.002)	(0.001)	0.000	0.287	48000
Cash Flow	0.177	0.182	0.204	0.251			
	(0.028)	(0.020)	(0.021)	(0.027)	0.000		
			Dep Variable: Hiring Rate				
Q (Emp)	0.067	0.077	0.048	0.032			
	(0.009)	(0.011)	(0.007)	(0.004)	0.000	0.073	43500
Cash Flow	0.260	0.225	0.120	-0.026			
	(0.053)	(0.048)	(0.024)	(0.035)	0.000		

Note: Panel A of this table shows average investment rates in capital and in new hires for firms of different productivity quartiles. Panel B includes results from estimating (1.1) in the main text. Q (Total) is the intangible-adjusted total Q from Peters and Taylor (2017) and the total investment rate is the dollar investment in total (physical plus intangible) capital divided by lagged replacement value of total capital. Q (Emp) is an employment-based analogue of Tobin's Q, where the denominator is the skill-weighted workforce of the firm. "P-val 4-1" gives the p-value from a test that the coefficients for firms in the top and bottom quartiles have equal values. All specifications include firm, year, and productivity quartile fixed effects. The sample period spans 1991-2014. Standard errors double clustered by industry and year in parentheses.

Table 1.3: Baseline Elasticity Estimates, By Worker Skill and Incumbent or Recruit Status

	All	Low Skill	Middle Skill	High Skill
Panel A: Whole Firm				
Employment	0.116	0.128	0.110	0.102
	(0.01)	(0.011)	(0.011)	(0.01)
Wages	0.046	0.016	0.019	0.084
	(0.004)	(0.002)	(0.001)	(0.005)
Implied Elasticity	2.53	7.97	5.92	1.22
Panel B: Incumbents				
Employment	0.091	0.082	0.083	0.099
	(0.009)	(0.009)	(0.01)	(0.01)
Wages	0.053	0.018	0.020	0.082
	(0.004)	(0.002)	(0.001)	(0.004)
Implied Elasticity	1.73	4.55	4.10	1.20
Panel C: Recruits				
Employment	0.203	0.220	0.200	0.152
	(0.016)	(0.017)	(0.016)	(0.015)
Wages	0.028	0.022	0.018	0.038
	(0.002)	(0.002)	(0.002)	(0.003)
Implied Elasticity	7.19	9.98	11.32	3.99

Note: This table shows estimates of the supply elasticities implied by the employment and wage responses to stock returns from estimating (1.6) in the text. Controls include 3-digit NAICS industry by year and productivity quartile fixed effects; lagged growth rates in wages, employment, and total assets; and the contemporaneous change in average worker skill level at the firm (see (1.41) for definition). Workers are grouped into skill groups based on their estimated worker effects from a modified Abowd et al. (1999) style wage decomposition with time-varying firm fixed effects. Individuals in the bottom two quintiles of the cross-sectional distribution of worker effects are considered low-skilled, the third and fourth quintiles middle-skilled, and the top quintile high-skilled. Changes in average worker skill are computed within the population of workers considered in the specification. Panel A estimates (1.6) for the whole firm and for workers of different skill levels. Panels B and C restrict (1.6) to incumbents (workers who were employed at the firm the previous year) and recruits (workers who joined the firm in the given year). Standard errors clustered by industry and year are in parentheses.

Table 1.4: Productive Firms Face Lower Supply Elasticities For Workers of All Skill Levels

Productivity:	Quartile 1	Quartile 2	Quartile 3	Quartile 4	P-val 4-1	R-sq	N
			Whole Firm				
Employment	0.119	0.089	0.120	0.106			
	(0.013)	(0.010)	(0.011)	(0.011)	0.249	0.127	43500
Wages	0.028	0.033	0.051	0.091			
	(0.003)	(0.003)	(0.004)	(0.010)	0.000	0.423	43500
Elasticity	4.324	2.708	2.336	1.166			
			Low Skill Workers				
Employment	0.128	0.102	0.143	0.139			
	(0.015)	(0.010)	(0.013)	(0.015)	0.522	0.110	43500
Wages	0.012	0.016	0.021	0.030			
	(0.002)	(0.003)	(0.003)	(0.003)	0.000	0.289	43500
Elasticity	10.510	6.369	6.768	4.663			
			Middle Skill Workers				
Employment	0.116	0.084	0.118	0.103			
	(0.016)	(0.010)	(0.010)	(0.013)	0.307	0.085	43500
Wages	0.014	0.016	0.022	0.029			
	(0.002)	(0.002)	(0.002)	(0.003)	0.000	0.182	43500
Elasticity	8.199	5.298	5.305	3.546			
			High Skill Workers				
Employment	0.102	0.078	0.106	0.087			
	(0.014)	(0.010)	(0.009)	(0.010)	0.148	0.076	43500
Wages	0.050	0.056	0.076	0.119			
	(0.006)	(0.004)	(0.006)	(0.010)	0.000	0.454	43500
Elasticity	2.042	1.388	1.389	0.735			

Note: This table contains supply elasticity estimates for firms sorted on log value-added/worker quartiles as in (1.8) in the text. Controls include 3-digit NAICS industry by year and productivity quartile fixed effects; lagged growth rates in wages, employment, and total assets; and the contemporaneous change in average worker skill level at the firm (see (1.41) for definition). Workers are grouped into skill groups based on their estimated worker effects from a modified Abowd et al. (1999) style wage decomposition with time-varying firm fixed effects. Individuals in the bottom two quintiles of the cross-sectional distribution of worker effects are considered low-skilled, the third and fourth quintiles middle-skilled, and the top quintile high-skilled. Changes in average worker skill are computed within the population of workers considered in the specification. “P-val 4-1" gives the p-value from a test that the coefficients for firms in the top and bottom quartiles have equal values. Wage data are from the LEHD, and the sample period spans 1991-2014. Standard errors double clustered by industry and year in parentheses. See section 1.3 in main text for more details.

Table 1.5: Model Calibration

Parameter	Explanation	Value
Externally Calibrated		
L	Labor Force Size	1.0
r	Discount rate	0.02
κ	Supply scale factor	0.741
ρ_z	Productivity persistence	0.9
$\bar{z}$	Fixed Labor Productivity	1.0
σ_c	Convex adjustment cost volatility	1.0
Internally Calibrated		
α	(One minus) Labor returns to scale	0.22
β	Supply Elasticity Shifter	0.38
b	Reservation Wage	0.545
σ_z	Labor productivity volatility	0.145
$\bar{C}$	Convex adjustment cost level	0.57

Note: This table shows the parameter calibration for the model in section 3.1 of the text. Externally calibrated parameters are calibrated beforehand without trying to match target model moments, while internally calibrated parameters are explicitly set in order to match target moments.

Table 1.6: Model vs Data Moments

Panel A: Targeted Moments

Moment Name	Model	Data
Separations Rate	0.31	0.30
Incumbent Premium	0.15	0.15
Supply Elasticity	2.09	2.52
Elasticity (Inc.)	1.44	1.73
Elasticity (Rec.)	6.66	7.19
Inc/Rec Wage Pass. Ratio	1.93	1.86
Log Labor Share	-0.50	-0.51

Panel B: Sorting on Productivity

Moment Name	Model	Data
Elasticity (Q1)	4.85	4.32
Elasticity (Q2)	2.53	2.71
Elasticity (Q3)	1.80	2.34
Elasticity (Q4)	1.36	1.17
Log Labor Share (Q1)	-0.31	-0.22
Log Labor Share (Q2)	-0.46	-0.40
Log Labor Share (Q3)	-0.56	-0.54
Log Labor Share (Q4)	-0.68	-0.94

Note: This table compares model implied moments with their empirical counterparts. “Inc" and “Rec" denote incumbents and recruits, respectively. In order to mimic the empirical estimates, the model supply elasticities are computed by running a regression of model-generated stock returns on employment and wage growth and taking the ratio of the responses. See section 3.1 in text for details.

Table 1.7: Adjusted Markdowns for Firms Sorted on Labor Productivity Implied By Empirical Elasticity Estimates and Markdown Wedge ν From Model

	Elasticity	Unadjusted Markdown	Adjusted Markdown
Overall	2.52	0.72	0.83
Quartile 1	4.32	0.81	0.94
Quartile 2	2.71	0.73	0.85
Quartile 3	2.34	0.70	0.81
Quartile 4	1.17	0.54	0.62

Note: This table compares markdowns implied by my empirical supply elasticity estimates before adjusting for the model markdown wedge ν estimated from the model and after. For supply elasticity ϵ, the unadjusted markdown is $\epsilon/(1+\epsilon)$; the adjusted markdown is $\nu \times \epsilon/(1+\epsilon)$.

Table 1.8: Valuing Labor Market Power—Wage Markdowns as a Fraction of Operating Income

Panel A: Full Sample (1991-2014)

	Overall	Low Productivity	High Productivity
Mean Firm-Level Operating Income Share	0.34	0.17	0.43
Median Firm-Level Operating Income Share	0.30	0.08	0.49
Mean Aggregate Operating Income Share	0.40	0.10	0.45

Panel B: By Subperiod (1991-2002 and 2003-2014)

	Overall	Low Productivity	High Productivity
1991-2002			
Mean Firm-Level Operating Income Share	0.34	0.19	0.42
Median Firm-Level Operating Income Share	0.29	0.10	0.45
Mean Aggregate Operating Income Share	0.39	0.12	0.43
2003-2014			
Mean Firm-Level Operating Income Share	0.35	0.14	0.44
Median Firm-Level Operating Income Share	0.32	0.07	0.52
Mean Aggregate Operating Income Share	0.41	0.07	0.46

Note: This table shows the dollar value of wage markdowns as a fraction of operating income; these are broken down by the full sample period in panel A and by subperiod in panel B. See section 1.5.1 of the main text for more details.

Table 1.9: Differences in Labor Shares Explained by Markdowns, Cross-Section and Time Series

Panel A: High Minus Low Productivity Cross-Sectional Average Labor Share Spread

	Low Productivity	High Productivity	High - Low	% Due to ΔMarkdown
Log Markdown	-0.05	-0.46	-0.41	
Avg Log LShare ($\widehat{\textsc{xlr}}$)	-0.18	-0.83	-0.65	63.5
Avg Log LShare ($\textsc{labex}$)	-0.22	-0.94	-0.72	57.1

Panel B: Time Series Change in Aggregate Labor Share

	$1991-2002$	$2003-2014$	$\widetilde{2003-2014}$	$\Delta\log(\text{Lshare})$	$\Delta\widetilde{\log(\text{Lshare})}$	% Due to ΔMarkdown
Agg LShare ($\widehat{\textsc{xlr}}$)	0.579	0.524	0.554	-0.10	-0.044	55.8
Agg LShare ($\textsc{labex}$)	0.506	0.485	0.515	-0.042	0.018	143.2

Note: Panel A of this table decomposes the fraction cross-sectional average labor share differences between high- and low-productivity firms that can be attributed to wage markdowns. In panel A I use markdown estimates for the full 1991-2014 period reported in Table 1.8. Panel B of this table decomposes the time series change in the aggregate labor share between the 1991-2002 and 2003-2014 period using markdowns estimated separately for the two subperiods. The column $\widetilde{2003-2014}$ denotes the counterfactual average labor share for the 2003-2014 period if markdowns were held constant at their 1991-2002 estimates, as defined in (1.33) in the main text. The column $\Delta\log(\text{Lshare})$ gives the actual log change in the average labor share between the two periods, while $\Delta\widetilde{\log(\text{Lshare})}$ gives the counterfactual log labor share change holding markdowns constant. The last column of panel B gives the fraction of the observed change in the labor share that is attributable to wage markdown changes. See section 1.5.2 in main text for further details.

1.7.1 Data Appendix

Constructing Labor Productivity, Wages, Labor Shares, and Other Firm-Level Variables

Firm Value-Added and Labor Shares

I follow Donangelo et al. (2019) in defining value-added for Compustat firms following as the sum of operating income before depreciation, changes in inventories, and labor expenses.

$$\mathrm{VA}_{j,t} = \mathrm{OIBDP}_{j,t} + \Delta\mathrm{INVFG}_{j,t} + \mathrm{LABEX}_{j,t} \tag{1.34}$$

Changes in inventories are set to zero when missing. Here

$$\mathrm{LABEX}_{j,t} = \widetilde{W}_{j,t} \times (\mathrm{EMP}_{j,t} + \mathrm{EMP}_{i,t-1})/2 \tag{1.35}$$

Instead of imputing wages as industry-size cell averages of Compustat item XLR as in Donangelo et al. (2019), I create a measure of the average wage ($\widetilde{W}_{j,t}$) paid to workers using my LEHD-Compustat match. I detail my computation of the average wage later on in this section. Because coverage of the LEHD varies by year and may not contain all establishments from a given Compustat gvkey in a given year, I multiply the inferred LEHD wage by the average of the firm's employment in years t and $t-1$, rather than directly summing up LEHD wage compensation. Because Compustat employment is reported at year-end, I also follow Donangelo et al. (2019) in taking the average employment in adjoining years. Labor productivity is given by

$$\log(\mathrm{VA}/\mathrm{Worker})_{j,t} = \log(\mathrm{VA}_{j,t}/\mathrm{EMP}_{j,t}) \tag{1.36}$$

Unless otherwise specified, I refer hereafter to labor productivity, productivity, and log value-added per worker interchangeably. Finally, I define the labor share of firm i at time t by

$$\mathrm{LSHARE}_{j,t} = \frac{\mathrm{LABEX}_{j,t}}{\mathrm{VA}_{j,t}} \tag{1.37}$$

AKM Wage Decomposition

I now detail how I decompose wages into worker- and firm-specific heterogeneity in the tradition of Abowd et al. (1999) (AKM). I start with a modification of the AKM decomposition proposed by Lachowska et al. (2020) and Engbom and Moser (2020) that allows for the firm-specific component of wages to vary by time. Let i index individual workers; $j(i,t)$ a function indicating the firm j that employs individual i at time t; and $X_{i,t}$ a third-degree polynomial in worker age that is flat at age 40, as in Sorkin (2018) and Card, Heining, and Kline (2013). I then estimate the wage decomposition

$$\log(w)_{ijt} = \alpha_i + \phi_{j(i,t),t} + \beta X_{i,t} + \epsilon_{i,t} \tag{1.38}$$

I estimate (1.38) for matched Compustat firms for overlapping 5-year moving windows. Because my analysis is at the annual frequency and LEHD earnings are quarterly, each wage w_{ijt} represents a full-year equivalent real wage for individual i in year t, as in Sorkin (2018). In order to be included in the sample for estimating (1.38), firm j must be worker i's primary employer for that year (the firm with highest earnings), worker i must have been employed at that firm for at least two consecutive quarters within the year, and the worker must have earned more than \$3250 in 2011 US dollars. Papers performing AKM wage decompositions on LEHD earnings data set a lower threshold on annual earnings because earnings because hours worked are not observed in the LEHD. Due to the large size of the LEHD data, I estimate (1.38) for four disjoint 25% subsamples of my original sample of individuals found in the LEHD. All firm level aggregates taken from these estimates represent averages across these four 25% subsamples. Estimating (1.38) also requires that firm-years and individual worker pairs must belong to a set satisfying connectedness conditions in order for the fixed effects α_i and $\phi_{j(i,t),t}$ to be separately identified. I provide more details on this connectedness requirement and the sampling procedure in appendix 1.7.1.

I use the sample of workers in (1.38) to create my inferred firm average wage $\widetilde{W}_{j,t}$. Let $N^{ins}_{j,t}$ denote the total number of workers mapped to firm j that are in the sample of (1.38) in year t. Let $N^{not-ins}_{j,t}$ denote the number of unique individuals that show up on the payrolss of firm j in year t that are not included in the sample (either due to earnings below the required threshold or firm j not being the primary employer for that year). Let $\widetilde{w}_{ijt}$ denote the actual (not full-year adjusted) year t earnings of worker i at firm j. Then I compute the firm-specific wage as

$$\widetilde{W}_{j,t} = \frac{\sum_i \widetilde{w}_{ijt}}{N_{j,t}^{ins} + N_{j,t}^{not-ins}} \tag{1.39}$$

I contrast this unadjusted wage $\widetilde{W}_{j,t}$ with a full-time full year equivalent adjusted wage $W_{j,t}$. Let $\Gamma_{j,t}$ denote the set of workers that are in the sample of (1.38) for firm j in year t:

$$W_{j,t} = \frac{\sum_{i \in \Gamma_{j,t}} w_{ijt}}{N_{j,t}^{ins}} \tag{1.40}$$

I use the actual average earnings $\widetilde{W}_{j,t}$ multiplied by Compustat employment for computing firm level value-added, labor expenses, and labor shares, as in (1.34), (1.35), and (1.37). Meanwhile, I focus on the employment response to log changes in the firm's full-time, full-year equivalent adjusted wage $W_{j,t}$ when I estimate supply elasticities in Section 1.3. I make these choices in order to ensure that my computation of labor expenses represents actual spending on labor per Compustat employee, while estimated supply elasticities correspond to the employment response induced by a change in the wage offer for a consistently defined period of employment.

Firm-Level Measures Derived From AKM Estimates

I introduce three measures that I derive from the wage decomposition in (1.38). The first is a measure of the firm's skill:

$$\alpha_{j,t} = \log\left(\frac{\sum_{i \in \Gamma_{j,t}} \exp(\widehat{\alpha}_{i,t})}{N_{j,t}^{ins}}\right) \tag{1.41}$$

Thus $\alpha_{j,t}$ is the log of the average component of wages that is due to the worker-specific heterogeneity of the firm's labor force. Firms with more skilled workers will have a higher $\alpha_{j,t}$. Next, I introduce an employment based analogue of the Tobin's Q valuation ratio. Let

$V_{j,t}$ denote the total enterprise value of firm j in year t. Then define $Q_{j,t}^{Emp}$ as

$$Q_{j,t}^{Emp} = \frac{V_{j,t}}{\sum_{i \in \Gamma_{j,t}} \exp(\widehat{\alpha}_{i,t})} \tag{1.42}$$

I compute firm value $V_{j,t}$ following Peters and Taylor (2017). Finally, I introduce a skill-adjusted measure of the firm hiring rate. Denote by $\Gamma_{j,t}^{Recruit}$ the set of workers i that are recruited to firm j in year t, where a recruit is any worker whose primary employer is j in year t, but not in year $t-1$. Then define

$$\text{Inv Rate}_{j,t}^{Hire} = \frac{\sum_{i \in \Gamma_{j,t}^{Recruit}} \exp(\widehat{\alpha}_{i,t})}{\sum_{i \in \Gamma_{j,t-1}} \exp(\widehat{\alpha}_{i,t-1})} \tag{1.43}$$

Hence Inv Rate$_{j,t}^{Hire}$ represents the ratio of the skill-weighted number of workers hired in year t relative to the skill-weighted number of workers employed by the firm in the previous year.

Sampling/Cleaning LEHD Wage Data and Estimation of Wage Decomposition

The basic person-level identifier variable in the LEHD data is the PIK, which has a one-to-one correspondence with and individual's Social Security number. My baseline LEHD sample comes from the list of unique PIK identifiers obtained from the union of all individuals found in the 2000 Decennial Census; the SIPP and Current Population Survey; and a 10% subsample of the Numident sample. I further generate four disjoint 25% subsamples of this list of PIKs intersected with the list of PIKs in the LEHD.

I repeat the following steps for each of the four disjoint 25% subsamples. All estimates are taken separately across these disjoint subsamples; all firm level aggregates represent averages across these four subsamples. For each year in the LEHD 1990-2015 I retain all individuals who were found to be employed at a Compustat-linked firm in that year. This forms my LEHD-Compustat match. I then clean LEHD earnings data using a procedure that follows very closely with Sorkin (2018). Because days and hours worked in the LEHD are not observed, these steps are meant to convert quarterly LEHD earnings to their full-year wage equivalents. I categorize quarterly earnings observations into three groups: full, continuous, and discontinuous. Quarter q is a full earnings occur when individual i is linked to firm j at quarters q, $q-1$ and $q+1$; quarter q is a continuous is when individual i is linked to firm j at quarters q and on of $q-1$ and $q+1$, but not both; finally, discontinuous quarters occur when individual i is linked to firm j at during quarter q but not at $q-1$ or $q+1$.

Because full quarters are most likely to represent full-time employment, they are prioritized as follows. If individual i has any full quarters of employment at firm i in the given year, then annual wages are taken to be

$$4 \times \text{Total Earnings in Full Quarters/Number of Full Quarters.}$$

If there are no full quarters of employment, then annual earnings are

$$8 \times \text{Total Earnings in Continuous Quarters/Number of Continuous Quarters.}$$

Finally, if there are no continuous quarters, annual earnings are given by

$$12 \times \text{Total Earnings in Discontinuous Quarters/Number of Discontinuous Quarters.}$$

Assuming separations occur uniformly within a quarter, a continuous quarter represents a half quarter of employment at the firm and a discontinuous quarter represents a third of a quarter of employments, so this adjusts wages to full-year terms. Full quarters require no such adjustment, which is the reason for prioritizing full quarters of employment over others. Earnings are further adjusted to real equivalents. I follow Sorkin (2018) in using the CPI from the 4th quarter of 2011 as the baseline for this real wage adjustment.

I link each individual to their primary in each year. The primary employer is the one where their unadjusted earnings are highest. I define the employer at the gvkey level for worker-firm-years linked to a Compustat firm, and at the EIN level for worker-firm-years with no such link. Since I restrict to persons linked to Compustat in the current year, every individual in my sample will have at least one job in year t linked to a Compustat gvkey, although this will not always be their primary employer. Using the adjusted earnings I estimate the modified AKM decomposition for individuals' primary employers:

$$\log(w)_{ijt} = \alpha_i + \phi_{j(i,t),t} + \beta X_{i,t} + \epsilon_{i,t} \tag{1.44}$$

As in Sorkin (2018), I require individuals to have earned more than \$3250 in real earnings at their employer in the year in order to be considered part of the sample for estimating (1.44). Since hours or days worked are not observed, this drops individuals likely to have had a minimal attachment to the firm in the year due to part-time employment. I estimate (1.44) for overlapping 5-year periods starting in 1991 and ending in 2010. Because I have estimates for overlapping 5-year intervals, I take estimates for the final year from the sample (i.e. firm and person effects in 2013 come from the 2009-2013 subsample, 2014 comes from 2010-2014). The exception is for 1991-1994, which do not have a full 5-year period ending in the given year, so estimates for these years come from the 1991-1995 subsample. This gives me estimates of model parameters for the 1991-2014 period (though the LEHD covers 1990 and 2015, I use these years to determine full/continuous/discontinuous quarters in 1991 and 2014, respectively, leading my sample period to span 1991 to 2014).

Fixed effects in (1.44) are only defined for the collection of firm-years connected by worker flows across firms and time. Because I have time-varying firm effects, the connectivity requirements are slightly different than in the classic AKM decomposition with time-invariant firm effects (see Lachowska et al. (2020) and Engbom and Moser (2020) for a detailed discussion on the connectivity requirements in order for parameters to be identified). Accordingly, when estimating (1.44) I restrict the sample to the largest connected set of worker-firm-year observations following these two papers. Because Compustat firms are large, this connectivity restriction drops a tiny fraction of the data.

Imputing Compustat Staff and Labor Expense (XLR)

Here I describe my imputation of Compustat variable XLR. I use $LABEX_{j,t}$ from (1.35) in the main text. I use $LABEX_{j,t}$ to predict XLR out-of-sample (I also use predicted XLR even when actual XLR is available for consistency). To ensure the predicted XLR is positive

I run the following regression in logs:

$$\log(XLR_{j,t}) = \alpha + \alpha_{I(j),t} + \beta_t \log(LABEX_{j,t}) + \epsilon_{j,t} \tag{1.45}$$

Here $\alpha_{I(j),t}$ are two-digit NAICS by year fixed effects and β_t are time-varying coefficients on $\log(LABEX_{j,t})$.

My predicted level of XLR is then just $\widehat{XLR}_{j,t} = \exp\left(\widehat{\alpha} + \widehat{\alpha}_{I(j),t} + \widehat{\beta_t}\log(LABEX_{j,t})\right)$. When estimation error is taken into consideration, simply taking the exponential of the predicted log of the variable in question often leads to better forecasts (Brardsen and Lütkepohl, 2011). A Jensen's inequality adjustment term for a scaling factor, using the variance of the residuals and assuming lognormality leads to a non-trivial overestimate of the actual XLR, while the average level of $\widehat{XLR}_{j,t}$ is a much closer to actual $XLR_{j,t}$. Because of this I use the exponential of the log to create a strictly positive predicted $XLR_{j,t}$. In-sample $\widehat{XLR}_{j,t}$ has a correlation of 0.95 with actual $XLR_{j,t}$.

K-Means Clustering to Estimate Empirical Labor Market Boundaries

In this section I describe my method for grouping firms into empirical labor market boundaries based on k-means clustering of flows of workers across firms. Let $R_{i,j,t}$ be the numer of workers firm i has hired from firm j in year t, and $L_{i,j,t}$ the number of workers firm i loses to firm j in year t. I then construct a matrix A_t of flows across firms by assigning the the i, jth entry as follows:

$$A_{i,j,t} = \log\left(1 + \sum_{\tau=t-2}^{t} R_{i,j,\tau} + L_{i,j,\tau}\right) \tag{1.46}$$

I then perform k-means clustering on the columns of A to group firms into clusters that hire from one another, using 1 minus the cosine similarity of column vectors of A_t as the distance metric between vectors. The cosine distance between column $A_{j,t}$ and $A_{j',t}$ is defined as

$$1 - \frac{\sum_i A_{i,j,t} \times A_{i,j',t}}{\sqrt{\sum_i A_{i,j,t}^2}\sqrt{\sum_i A_{i,j',t}^2}}, \tag{1.47}$$

which is one minus the uncentered correlation between the vectors $A_{j,t}$ and $A_{j',t}$. Thus the distance metric accounts for the differences in the size of the two comparison firms. I take the log of the sum of inflows and outflows in order to downweight extremely large firms but still allow entries to be increasing in the number of workers that flow between the two firms. Because k-means clustering has a random component, I perform the routine 5 times each year using different starting points for the clusters, selecting the cluster assignment which explains the largest share of workers flows across firms for that year. Because the number of clusters must be pre-specified, I perform the routine for $k = 10$ and $k = 20$ labor market clusters per year.

I find that the method explains labor empirical labor markets quite well. In appendix table 1.17 I examine how much variation labor market-by-year fixed effects explain in the levels of and changes in log wage and employment as well as the fraction of firm-to-firm worker transitions occurring within the given labor market boundary. I compare the $k = 10$ and $k = 20$ clusters with 2-digit and 3-digit NAICS industrys. The table shows that about half of worker flows occur within the $k = 10$ version of the empirical labor markets, while also explaining three-fifths of log wages and having very comparable explanatory power for wage and employment growth, and stock returns as the other empirical labor market boundaries. In the final column of the table I show that, while the empirical labor market boundaries do a good job of capturing empirical labor markets, they capture similar variation in stock returns, and employment/wage growth as my baseline 3-digit NAICs fixed effects. In particular, there is a very small change in explanatory power when both labor market-by-year and 3-digit NAICs-by-year fixed effects are included at the same time.

1.7.2 Controlling for Proxies of Labor and Capital Adjustment Costs in Investment-Q Regressions

The smaller investment response to Tobin's Q for productive firms are suggestive that Tobin's Q proxies relatively more for economic rents and not investment opportunities for these firms; however, the lower investment rates could also be due to higher adjustment costs for a given level of investment.[26]. Adjustment costs are likely different skilled and unskilled workers. For example, Belo et al. (2017) argue that adjustment costs are likely to be considerably higher for firms with a more skilled workforce, and Jager and Heining (2019) shows direct empirical evidence that firms have difficulty substituting for skilled workers. In appendix Table 1.12 I regress hiring on employment Q, and interact employment Q separately with labor productivity and estimated log firm average worker effects ($\alpha_{j,t}$ from (1.41)). The interaction term still has a strongly negative coefficient even after accounting for firm skill, suggesting that at least for hiring rates, the patterns are not likely to be driven purely by differences in adjustment costs. The coefficient on the interaction between average worker effects and employment Q is also negative, suggesting that these firms do face higher adjustment costs.

In appendix Table 1.13 I devise a similar exercise for capital investment based on the intangible share of capital, where I interact total Q with the log intangible share of total capital as proxy for variation in capital adjustment costs. To the extent that intangible and physical capital have differential convex adjustment costs, these findings are also not driven by productive firms facing disparate costs of adjusting their capital stock on account of having differing amounts of intangible capital.

1.7.3 Elasticity Estimate Bias in a Simple Model of the Labor Market

Similar to Card et al. (2018) or Lamadon et al. (2019), suppose that L workers choose employers among a market of N firms, and normalize $L = 1$ for simplicity. Worker i's utility from working at firm j is increasing in firm-specific amenities $a_{j,t}$, the log of the wage offer $w_{j,t}$, and an unobservable taste shock $\epsilon_{i,j,t}$.

$$u_{i,j,t} = \theta \log(w_{j,t}) + a_{j,t} + \epsilon_{i,j,t} \tag{1.48}$$

[26] In the case where convex adjustment costs are quadratic and returns to scale are constant so that Tobin's Q = marginal Q, the coefficient on Tobin's Q is precisely one divided by the multiplicative convex adjustment cost parameter.

Assuming the taste shocks follow a type I extreme value distribution, then standard results from McFadden (1973) imply the firm-specific labor supply curve:

$$L(w_{j,t}, a_{j,t}) = \lambda_t^{-1} \exp(a_{j,t}) w_{j,t}^{\theta} \tag{1.49}$$

Here I assume each firm views itself as atomistic in the market, and so they take the constant $\lambda_t = \left(\sum_{j'=1}^{N} \exp(a_{j',t}) w_{j',t}^{\theta}\right)$ as given. The parameter θ gives the firm-specific supply elasticity. Let $A_{j,t} = \lambda_t^{-1} \exp(a_{j,t})$ denote the level of the supply curve.

Firms choose the wage offer $w_{j,t}$ to maximize the following:

$$V_{j,t} = \max_{w_{j,t}} Z_{j,t}(L_{j,t})^{1-\alpha} - w_{j,t} L_{j,t} \tag{1.50}$$

subject to the functional form of the labor supply curve (1.49). Denote the wage markdown by $\mu \equiv \frac{\theta}{\theta+1}$ and define constant $c \equiv \frac{1}{1+\theta\alpha} \log\left(\mu(1-\alpha)\right)$. Solving (1.50) yields the expressions for the log optimal employment, wage, and firm value:

$$\log(L_{j,t}) = \theta c + \frac{\theta}{1+\theta\alpha} \log(Z_{j,t}) + \frac{1}{1+\theta\alpha} \log(A_{j,t}) \tag{1.51}$$

$$\log(w_{j,t}) = c + \frac{1}{1+\theta\alpha} \log(Z_{j,t}) - \frac{\alpha}{1+\theta\alpha} \log(A_{j,t}) \tag{1.52}$$

$$\log(V_{j,t}) = \log(1-(1-\alpha)\mu) + (1-\alpha)\theta c + \frac{1+\theta}{(1+\theta\alpha)} \log(Z_{j,t}) + \frac{1-\alpha}{1+\theta\alpha} \log(A_{j,t}) \tag{1.53}$$

For simplicity I look at the bias from running a regression of the levels of log employment/wages on log firm value. My empirical strategy of running the regression of stock returns on the growth rates in wages and employment is essentially the same except in differences instead of levels. The parameter estimate obtained from regressing log employment and wages on firm value and taking the ratio of the two coefficients is given by

$$\frac{\theta(1+\theta)\sigma_z^2 + \theta(1-\alpha)\sigma_{az} + (1+\theta)\sigma_{az} + (1-\alpha)\sigma_a^2}{(1+\theta)\sigma_z^2 + (1-\alpha)\sigma_{az} - \alpha(1+\theta)\sigma_{az} - \alpha(1-\alpha)\sigma_a^2} \tag{1.54}$$

Here σ_z^2 is the variance of $\log(Z_{j,t})$; σ_{az} is the covariance of $\log(Z_{j,t})$ and $\log(A_{j,t})$; and, σ_a^2

is the variance of $\log(A_{j,t})$. When there are no labor supply shocks, so that $\sigma_a^2 = \sigma_{az} = 0$, equation (1.54) collapses to the true supply elasticity θ.

For (1.54) to represent an upper bound on the supply elasticity—which implies a conservative estimate of the magnitude of wage markdowns—we must have

$$\rho_{az}\sigma_z > -\frac{(1-\alpha)}{1+\theta}\sigma_a \tag{1.55}$$

where ρ_{az} is the correlation of $\log(A_{j,t})$ and $\log(Z_{j,t})$. The intuition behind (1.55) is simple. Increases in the level of the supply curve $\log(A_{j,t})$, (increases in "amenities") allow firms to hire more workers at a given wage, which reduces wages, increases employment, and increases firm value, all else held constant. This tends to bias the wage response downward and the employment response upward, leading to an upward biased elasticity estimate. However, if $\log(A_{j,t})$ is sufficiently negatively correlated with firm productivity, then increases in $\log(A_{j,t})$ reduce firm value, and the elasticity estimate becomes downward biased.

The most obvious candidate for reversing the inequality (1.55) is market-specific productivity shocks. This is because $A_{j,t}$ is decreasing in the market-wide wage index $\lambda_t = \left(\sum_{j'=1}^{N} \exp(a_{j',t}) w_{j',t}^{\theta}\right)$:

$$A_{j,t} = \lambda_t^{-1}\exp(a_{j,t}) = \left(\sum_{j'=1}^{N} \exp(a_{j',t}) w_{j',t}^{\theta}\right)^{-1} \exp(a_{j,t}) \tag{1.56}$$

Suppose that firm-specific productivity $Z_{j,t}$ is given by an aggregate market component and an idiosyncratic component: $Z_{j,t} = \widetilde{X}_t X_{j,t}$. Equation (1.52) then implies that the wage offer can be expressed as

$$w_{j,t} = C_{j,t}\widetilde{X}_t^{\frac{1}{1+\alpha\theta}} \lambda_t^{\frac{\alpha}{1+\alpha\theta}} \tag{1.57}$$

where the firm-specific term $C_{j,t}$ depends on $\exp(a_{j,t})$, $X_{j,t}$, and model parameters. Define $\widetilde{C}_{j,t} = \exp(a_{j,t}) C_{j,t}^{\theta}$. Then

$$\lambda_t^{\frac{1}{1+\alpha\theta}} = \left(\sum_{j'} \widetilde{C}_{j',t}\right) \widetilde{X}_t^{\frac{\theta}{1+\alpha\theta}} \equiv \widetilde{C}_t \widetilde{X}_t^{\frac{\theta}{1+\alpha\theta}} \tag{1.58}$$

or

$$\lambda_t = \widetilde{C}_t^{1+\alpha\theta} \widetilde{X}_t^{\theta} \tag{1.59}$$

Hence $\log(A_{j,t}) = a_{j,t} - (1+\alpha\theta)\log\left(\widetilde{C}_t\right) - \theta \log\left(\widetilde{X}_t\right)$ is decreasing in $\log(\widetilde{X}_t)$. If enough of the variance in $\log(A_{j,t})$ is driven by $\log(\widetilde{X}_t)$ this can reverse the inequality in (1.55), causing downward biased supply elasticities. Aggregate productivity shocks do have meaningful variance and firm-specific amenities may be much more slow moving, so this case is feasible empirically. Since I assume firms are "small" in the market, the correlation of $\log(\widetilde{C}_t)$ and $\log(C_{j,t})$ is approximately zero and so I don't consider the impact of this term. The $\log(\widetilde{C}_t)$ term nets out with market-wide fixed effects along with $\log(\widetilde{X}_t)$. In summary, this means that without market-specific controls I am likely to overestimate the importance of labor market power.

After netting out market-specific productivity shocks (via market-by-year fixed effects or other controls) the condition (1.55) becomes highly plausible. When market shocks are netted out, (1.55) says that workers' perceptions of firm amenities should not *decrease* by too much when idiosyncratic productivity *improves*, or that firms should not cut their amenities by a lot on average when they experience a positive productivity shock. The more likely scenario is that firm amenities also improve when their productivity goes up, which guarantees that (1.55) holds, implying my estimates would be conservative if biased at all.

1.7.4 Simple Example: Comparing Firm Value in Monopsony and Counterfactual Competitive Equilibria

To see how firm value might change in a competitive equilibrium relative to the monopsony equilibrium, consider the following simple setting. There is a representative firm which solves the following:

$$V = \max_{w} AL(w)^{1-\alpha} - wL(w) \tag{1.60}$$

where $L(w) = w^{\varepsilon}$ and ε is the supply elasticity. The firm's optimal wage and employment are

$$w = \left(\frac{\varepsilon}{\varepsilon + 1}(1 - \alpha)A\right)^{1/(1+\alpha\varepsilon)}, \qquad L = \left(\frac{\varepsilon}{\varepsilon + 1}(1 - \alpha)A\right)^{\varepsilon/(1+\alpha\varepsilon)} \tag{1.61}$$

And the value of wage markdowns as a share of operating income is

$$\text{markdown share} = \frac{\left(\frac{\epsilon+1}{\epsilon} - 1\right) w^{\varepsilon+1}}{V} \tag{1.62}$$

Consider the calibration $A = 1$, $\varepsilon = 2.5$, and $\alpha = 0.3$. This yields an aggregate labor share of about 0.5 and a markdown share of operating income of 40%, both of which are very close to their empirical counterparts. The firm value in this calibration is $V = 0.25$.

Now suppose that the firm takes wages as given and equilibrium wages are determined by the same labor supply function:

$$V^c = \max_L AL^{1-\alpha} - wL \tag{1.63}$$

where $L = w^{\varepsilon}$ gives the market clearing condition for wages. The resulting competitive wages and employment are

$$w^c = ((1 - \alpha)A)^{1/(1+\alpha\varepsilon)}, \qquad L^c = ((1 - \alpha)A)^{\varepsilon/(1+\alpha\varepsilon)} \tag{1.64}$$

Both the wage and employment increase in the competitive equilibrium. In the same calibration as before with $A = 1$, $\theta = 2.5$, and $\alpha = 0.3$, the competitive firm value is $V^c = 0.21$. So there is a 16% reduction in firm value in the counterfactual competitive equilibrium, even though wage markdowns are worth 40% of firm income in the equilibrium where the firm takes advantage of its labor market power.

1.7.5 Supply Elasticity Robustness Checks: Empirical Specifications and Data

Different Controls for Market Level Shocks

Failing to account for common market shocks could bias my supply elasticity estimates downward, leading to an upward bias in the extent of labor market power. In Table 1.16 I present the elasticity estimate from my baseline specification (1.6) , as well as for a set of alternative specifications with different controls for common market shocks. The top panel of the table shows the employment responses to a stock return shock, while the bottom panel shows the wage responses. My estimate of the average elasticity—obtained by taking the ratio of the employment and wage responses—in the baseline specification (column 1) is about 2.5. In columns 2-6 of Table 1.16 I include different variations of controls for common market shocks. At the bottom of these columns I include a p-value for the test that the coefficient in the given column is different from my baseline estimate in column 1.

I find that each set of controls for common shocks yields estimates that are economically and statisically close. From this exercise I conclude that the main specification in column 1 of Table 1.16 does a reasonably good job of capturing the relevant variation in common market shocks that could bias my estimates downward. Consequently I focus on this baseline specification, noting that any quantitative changes from using a different specification would be minor.

In column 2 of Table 1.16 I drop the industry-by-year fixed effects. Consistent with the discussion above and the simple model in appendix 1.7.3, this reduces my supply elasticity estimate. The difference has a p-value of .052, but economically the difference is not that large (2.525 in column 1 versus 2.397 in column 2). In the third column I add empirical labor market-by-year fixed effects instead of the industry-by-year fixed effects from column 1. I obtain these empirical labor market boundaries by performing k-means clustering on the flows of workers between firms; I explain the method in more detail in section 1.7.1 of the appendix and show that it does a good job of capturing variation in worker flows between firms, as well as in wages and employment levels and growth rates. I choose $k = 10$ labor market clusters per year as my baseline.[27] See appendix table 1.17 for more details on the comparative performance of different definitions of labor market boundaries. In column 3 the elasticity is about 2.54 when controlling for these empirical labor markets.

I drop the labor market fixed effects in column 4 of Table 1.16, instead controlling for the weighted average employment and wage growth of a firms' labor market competitors. To create this measure I weight the employment or wage growth by the number of workers a competitor has hired from the given firm, divided by the total number of workers that have been hired away by other firms in the sample. I similarly do this for the number of workers a given firm hires from a candidate firm. I then take the average of the inflow- and outflow-weighted measures. This column demonstrates that wages and employment do respond to competitor labor demand shocks, but controlling for these variables do little to change employment or wage marginal effects, and hence do not move the estimated elasticity.

In column 5 I control for both industry-by-year and labor market-by-year fixed effects, and in column 6 I use $k = 20$ labor market clusters per year instead of $k = 10$. This does not substantially change the supply elasticity estimate in quantitative terms and again the difference is not statistically significant.

Alternative Labor Demand Shocks

I now perform a number of robustness checks in order to gauge how much unobserved firm-specific labor supply shocks may be biasing my estimates. I first estimate supply elasticities using several different proxies for firm labor demand shifters. Results are found in appendix Table 1.18. I report the estimated supply elasticity using the alternative measure, the supply elasticity from my baseline specification , and the p-value on the test that my baseline supply elasticity estimates are different from the given estimate. Because a couple of the measures are available only for selected subsamples of the data, I re-estimate my baseline specification for these same subsamples of the data before comparing elasticity estimates.

In the first column of 1.18 I use stock returns of firms' customers rather than the firms

[27] Using a different clustering method, Nimczik (2020) finds that $k = 9$ empirical labor markets does a good job of capturing empirical market boundaries in Austria.

themselves as shock to labor demand. In the next column I use the firm-specific R&D tax credits from Bloom, Schankerman, and Van Reenen (2013) (with data updated by Lucking, Bloom, and Van Reenen (2019)) as a labor supply shifter; since this is a level variable rather than a flow, I estimate (1.6) in levels rather than first differences in this column. To test if firm-specific exposure to common stock market risk factors matters, in the third column I use return residuals backed out from a regression on the 5 Fama and French (2015) factors, plus the momentum risk factor. In the fourth column I use patent-induced shocks from Kogan, Papanikolaou, Seru, and Stoffman (2017) as a shifter of firm labor demand; and finally, in the last column I follow Daniel, Hirshleifer, and Sun (2019) in using stock returns in excess of the market return in a four day window around earnings announcements, which isolates periods of time when information about fundamental firm cash flows is revealed.

Each different labor supply shifter in Table 1.18 yields elasticity estimates that are very close quantitatively and statistically indistinguishable from my baseline estimates. Because each of these measures can be expected to covary differently with possible unobserved firm-specific labor supply shocks, this suggests that any bias from such shocks is not likely to have large a quantitative effect on my stock-return based supply elasticity estimates.

In the first column of Table 1.18 I show the response to stock returns of firms' customers using cleaned Compustat customer-supplier data provided by Wharton Research Data Services. Cohen and Frazzini (2008) show that stock return shocks to customers eventually propagate to upstream to their suppliers as a demand shock, and consistent with this notion I find that the wages and employment at supplier firm both respond significantly positively to the stock returns to their customers. The identifying assumption is that these shocks to customers constitute a pure demand shock to the firm, and are orthogonal to firm-specific labor demand shocks after accounting for industry-by-year fixed effects. Customers' returns are likely much less affected by any idiosyncratic labor supply shocks of their suppliers than the suppliers' own returns. Still, a potential concern is that customers may also be labor market competitors, and so in unreported results I consider a version where I exclude customer-supplier links between firms who have hired from one another within a 5 year window centered at the current year, or who are in the same 3-digit NAICS industry. In both cases the estimated supply elasticity is nearly the exact same as the estimate based on own-firm stock return and is statistically indistinguishable, with both elasticity estimates for the customer return sample being close to 2.

Column 2 of Table 1.18 uses variation in labor demand induced by the federal tax treatment of R&D expenses from Bloom et al. (2013), who show that differential exposure causes firms to undertake more R&D. Firms with differing tax-induced incentives to engage in more or less R&D may also be induced to adjust their labor forces to meet this incentive, which constitutes a labor demand shock. Because this measure is a level rather than a flow, I estimate a version of (1.6) in levels rather than changes when using the tax credit measure, now controlling for lagged log employment, wages, and assets and contemporaneous firm skill composition. I continue to include industry by year fixed effects to soak up market level variation. Again we find an elasticity estimate that is similar to the baseline and statistically indistinguishable, with the tax-credit induced elasticity being a little lower than my baseline for this sample of firms (1.73 to 1.96).

In the third column of Table 1.18 I use the firms' cumulative abnormal log excess returns with respect to the Fama and French (2015) five factor model augmented with the momentum factor.[28] I construct stock return residuals as follows. In each year and for each firm i I regress daily log excess returns on the 6 risk factors:

$$\log(R_{i,t}) - \log(R_{f,t}) = \alpha_i^\tau + \beta_i^\tau F_t + \epsilon_{i,t} \tag{1.65}$$

Here τ denotes the year corresponding to day t and F_t is a vector of risk factors. I take the estimated $\hat{\beta}_i$ for year τ and take these as the risk exposures for year $\tau + 1$ to get cumulative abnormal log returns. A literature starting with Vasicek (1973) suggests that out-of-sample risk exposures are more accurately estimated when applying Bayesian shrinkage to deflate the estimates towards a common average. I shrink the $\hat{\beta}_i$ estimates towards the cross-sectional average $E_\tau[\hat{\beta}_i]$, using the cross-sectional variance of the $\hat{\beta}_i$ estimates, $\text{Var}_\tau(\hat{\beta}_i^\tau)$, and the

[28] All risk factor and risk-free rate data are obtained from Ken French's website.

standard error of the own firm's estimate, $SE^2_{\beta_i}$, to get the shrinkage weights:

$$\widetilde{\beta_i^\tau} = (1-\omega)\hat{\beta}_i + \omega E_\tau[\hat{\beta_i^\tau}] \tag{1.66}$$

where

$$\omega = \frac{SE^2_{\beta_i}}{SE^2_{\beta_i^\tau} + \text{Var}_\tau(\hat{\beta_i^\tau})} \tag{1.67}$$

I similarly deflate the intercepts α_i. Finally, the cumulative abnormal returns for year $\tau+1$ are given by

$$\text{Abnormal Return}_{i,\tau} = \sum_{t\in\tau} \log(R_{i,t}) - \log(R_{f,t}) - \widetilde{\beta_i^{\tau-1}} F_t - \widetilde{\alpha_i^{\tau-1}} \tag{1.68}$$

I use Abnormal Return$_{i,\tau}$ in place of the excess return and re-estimate (1.6). I follow the same timimg convention as in mys baseline spec so that abnormal returns are aggregated from July until June of the following year. My inclusion of this measure relates primarily to the need to control for market shocks discussed previously and demonstrated in Table 1.16. In particular, a long literature in asset pricing argues that these factors represent systematic shocks, exposure to which demands compensation with higher returns. My inclusion of industry-by-year fixed effects implicitly imposes constant betas within industry on these systematic shocks, but it's possible the heterogeneity in exposure to systematic factors could matter. This procedure accordingly allows firms to have differential exposures to common factors. Using this measure again yields a very similar elasticity estimate to my baseline (2.34 relative to a baseline of 2.52, difference statistically insignificant).

In the fourth column I show the supply elasticity implied by employment and wage responses to patent induced shocks to firm value from Kogan et al. (2017), who show that the measure predicts changes in firm productivity, employment, and sales, all consistent with a marginal revenue productivity shock. Patent values are estimated from stock price movements in a small window around patent grants, and capture information in price movements related to firm innovation. I follow Kogan et al. (2020) in looking at the response of valuable of patents from the year they are filed rather than granted and use patenting in year t as a shock to labor demand from year t to $t+1$. This specification is related in spirit to Kline et al. (2019), who estimate passthroughs of patent-induced shocks to worker and firm outcomes based on predicted Kogan et al. (2017) patent values. Again I find an elasticity estimate that is close to my baseline.

Finally, I use the response to own-firm stock return shocks, except I only use stock price responses in a small time window around earnings announcements. Following Daniel et al. (2019) I use cumulative daily returns in excess of the market starting over the four day period starting the day before the announcement, summing up all 4-day announcement returns over the July to June period. The idea here is to isolate stock price movements that are highly likely to be related to information about firm productivity unrelated to information about labor supply. Stock price movements around earnings announcement are driven primarily by firms announcing unexpectedly high or low income; thus restricting to these small windows is more likely isolates price movements related to information about firm demand that is unrelated to firm-specific labor supply shocks. Though the measures are different, this follows a similar intuition to my use of Kogan et al. (2017) patent induced shocks to the firm—both measures isolate movements in prices due to information revealed to the market that is highly related to firm revenue productivity, and hence instruments for shifts in labor demand.

Controlling for Observable Labor Supply Shocks (Union Elections and Changes in Non-Compete Enforceability)

While these findings are useful in establishing the plausibility of my baseline supply elasticity estimates, it would still be helpful to control for potential firm-specific labor supply shocks, if they can be made observable. Although one can never definitively say that every firm-specific labor supply shock has been accounted for, I now include specifications where I control separately for two labor market shocks that have featured prominently in prior literature: union elections and changes in non-compete contract enforceability. Results are in appendix Table 1.19, which shows that bias from excluding these more salient observable firm-specific labor supply shocks is not important quantitatively. This lends credence to my argument that this sort of confounding variation is not likely to substantially bias my estimates in general.

I control for unionizations using data on union elections from the National Labor Relations Board and matched to Compustat records by Knepper (2020).[29] Changes in non-compete enforceability come from the lists of changes compiled by Ewens and Marx (2017) and Jeffers (2019). Following Ewens and Marx (2017) I assign non-compete changes by firms' headquarters.

Union elections may be one of the single best candidates for the type of firm-specific labor supply shock that could bias my estimates. For example, Lee and Mas (2012) find that union elections wins induce significant negative stock return responses; I verify the same result in the last row of Panel A of Table 1.19. However, even among the firms who have experienced large union elections, the unionization event accounts for a small amount of variation in firm-specific stock returns, so that controlling for unionizations leads to negligible changes in my estimates. Meanwhile, Jeffers (2019) argues that firms use non-competes successfully to diminish the mobility of their skilled workers, and so changes in non-compete enforceability could also in theory constitute a labor supply shock that bias my estimates. .

In Panel A of appendix Table 1.19 I re-estimate supply elasticities based off (1.6) for firms who ever experienced a union election during my sample; the first column is from the baseline specification without controlling for unionizations, and the others include different sets of unionization controls. Consistent with Lee and Mas (2012), I find a significantly negative stock return response in the year a union wins an election. Despite this fact, allowing for this supply shock to be observable has no effect on my estimated elasticities, despite the fact that I restrict the sample to only firms who have experienced a sufficiently large union election.

Following Knepper (2020), I focus on firms experiencing union elections where at least 20 employees voted in the election. In order to give maximal explanatory power to the union elections, I restrict the sample to just the set of firms identified at some point to have experienced a sufficiently large union election.

In the second column of Table 1.19 I include dummies for whether a union election win occurs in year t, $t+1$, or $t-1$ (as well as dummies for whether any election is occurs); in the

[29] Thanks to Matthew Knepper for generously sharing his data.

third column I ascribe all variation in stock returns in a union election year to the election by adding interactions of stock returns with the full set of union election dummies in the second columnn. In all cases the elasticity is quantitatively close to the baseline elasticity and statistically indistinguishable, even when assuming all variation in stock returns during in the 3 years surrounding surrounding a union election are due to the election.

Non-compete agreements make it more difficult to move to a competing employer; Jeffers (2019) shows that they are quite common and especially prevalent among skilled workers. Changes in non-compete enforceability are therefore another good candidate for firm-specific labor supply shocks that could bias my estimates. Following Ewens and Marx (2017) I assign non-compete changes by firms' headquarters. Panel B of Table 1.19 shows the resulting supply elasticities when controlling for non-compete changes. In the first column I include indicators for whether a non-compete increases or decreases in enforceability in years t, $t+1$, or $t-1$. In the next columns I add interactions of all non-compete dummies with the stock return in that year. As was the case with union elections, after controlling non-compete changes, supply elasticity estimates are quantitatively very close and statistically indistinguishable.

Robustness Checks for Labor Productivity Sorted Supply Elasticity Estimates

Since production methods and labor markets vary from industry to industry, one concern with my Table 1.4 could be reliance on unconditional labor productivity sorts that use between rather than within industry variation. In Table 1.20 I instead sort firms into productivity quartiles within their 2-digit NAICS industry and re-estimate (1.8) for employment and wages to back out supply elasticities for workers of all skill levels. All my basic findings from Table 1.4 are unchanged for the within industry sorts in 1.20, and all the elasticity point estimates are very similar.

In appendix Table 1.21 I address a few more potential concerns with elasticity estimates for firms sorted on productivity. One possibility is that wage responses are larger in the short-run for productive firms because they have more immediate flexibility in adjusting their wages, and so the monotonically decreasing pattern in elasticities sorted on productivity could be driven by the horizon. For example, unproductive firms may be more constrained in their ability to adjust wages in the short horizon. To address this, I re-estimate elasticities from (1.8) for the 3-year horizon. Specifically, I replace stock returns and employment/wage growth (as well as the control for contemporaneous changes in firm skill composition) with their 3-year equivalents:

$$\log(Y_{j,t+3}) - \log(Y_{j,t}) = \alpha + \alpha_{q(j,t)} + \alpha_{I(j),t} + \sum_{q=1}^{4} \mathbf{1}(q(i,t) = q) \times \beta_q \text{Stock Ret}_{j,t\to t+3} + \Gamma X_{j,t} + \epsilon_{j,t} \tag{1.69}$$

Panel A of Table 1.21 shows that productive firms similarly have much lower supply elasticities at this horizon. Thus the cross-sectional sorting in elasticities is not merely driven by the time horizon. Elasticities in general are also a little higher for this horizon, implying more elastic labor supply over the long run.

Another concern may be that results are driven by the equity-based compensation of the most skilled workers. For example, Eisfeldt, Falato, and Xiaolan (2021) document a large rise in equity compensation over my sample period, which may be due to contracting issues unrelated to a wage posting, monopsonistic model of the labor market. The LEHD includes all compensation that is immediately taxable, which includes equity based pay upon exercise. However, note also that such equity-based incentive pay is not incompatible with a monopsony framework, as it intrinsically ties the compensation of employees with their marginal revenue productivity, as would be implied by a monopsony model. Another issue is that a non-trivial fraction of equity-based compensation may not show up in LEHD earnings because they are taxed as capital gains, in which case the wage responses for supply elasticities may be mismeasured for skilled workers. That being said, Table 1.4 already alleviates these concerns in part, because productive firms face significantly lower supply elasticities even for the least skilled workers, whose compensation is far less tied to equity-based incentive pay.[30] Another factor that helps alleviate this concern is the fact that elasticity estimates are by far the lowest for skilled workers, and this is due to the larger wage passthrough of firm-specific

[30] Eisfeldt et al. (2021) find that 97% of equity-based incentive pay is accrues to the top 10% of workers in the manufacturing sector.

stocks to skilled workers' compensation. Deferred compensation that does not immediately show up as taxable earnings would tend to bias down the wage passthrough of skilled workers, but I find that skilled workers have by far the highest wage passthroughs of all. To address whether differences in deferred compensation are likely to drive the sorting patterns in supply elasticities, in Panel B of Table 1.21 I re-estimate (1.8) with changes in the time-varying firm wage fixed effects from (1.38), instead of the log average firm wage. This is the component of the log wage that is entirely firm-specific and is paid to all workers, regardless of their skill, incumbent status, or worker-firm match quality, and hence is less likely to be attached to firm- or worker skill-specific tendencies towards higher equity-based compensation. Again the most productive firms face by far the lowest supply elasticites. These elasticity estimates are not surprisingly a little higher than in Table 1.4, because they aren't able to take into account that the most highly paid workers make up a large share of the overall firm average wage and also have the least elastic labor supply.

Value-added per worker may be a noisy productivity proxy. In panel C I instead sort on estimates of firm total factor productivity from İmrohoroğlu and Tüzel (2014). Findings similarly go through in this case, as sorting on TFP also generates a decreasing pattern in supply elasticities. In unreported results I find a strong negative relationship between estimated TFP and firm labor shares. Hence my findings aren't driven by my choice of the log value-added per worker measure, but are instead driven by productivity advantages in general.

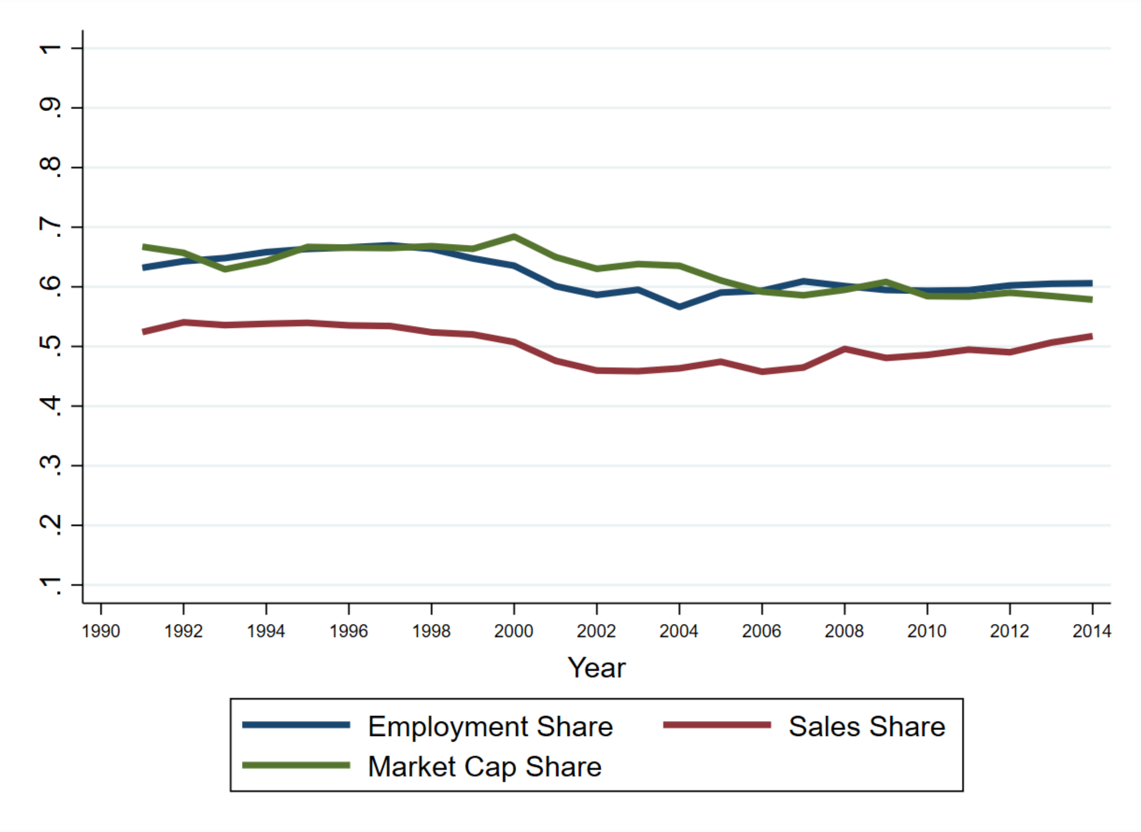
Finally, due to concerns some have raised regarding employment reported in Compustat (Davis, Haltiwanger, Jarmin, Miranda, Foote, and Nagypál, 2006), in panel D I replace the Compustat-based employment with Longitudinal Business Database employment. I obtain LBD employment by aggregating reported employment across all LBD establishments linked to a given Compustat gvkey in that year.[31] I still find strongly monotonically decreasing supply elasticities across productivity types, but the LBD employment is not as responsive

[31] LBD employment is collected in March while Compustat employment is almost always reported in December, and so I use LBD figures from the March nearest to the Compustat December employment report date to compute employment growth. I find that this yields larger employment responses than taking the previous March observation or an average of the two.

to stock return shocks and so I get slightly lower elasticity point estimates. This suggests that LBD-based supply elasticities would if anything increase the quantitative magnitude of my findings.

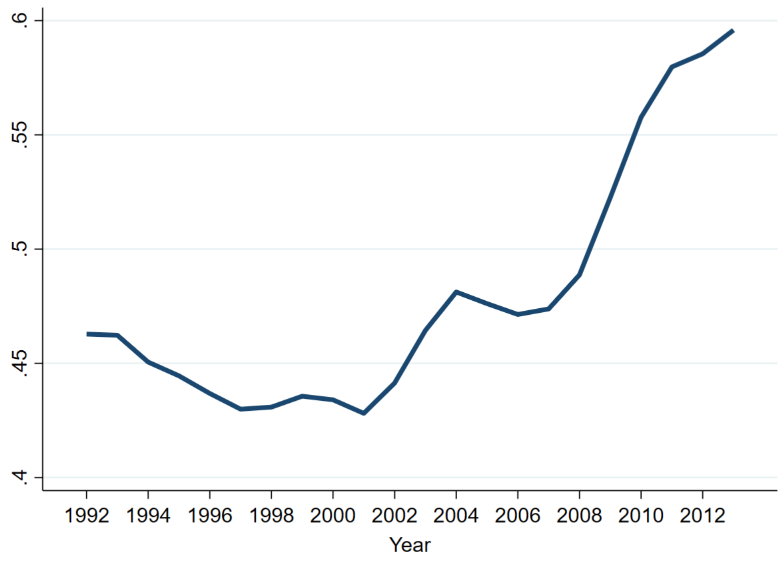
1.7.6 Appendix Figures and Tables

Figure 1-8: Compustat-LEHD Matched Sample: Shares of Employment, Market Cap and Sales by Year



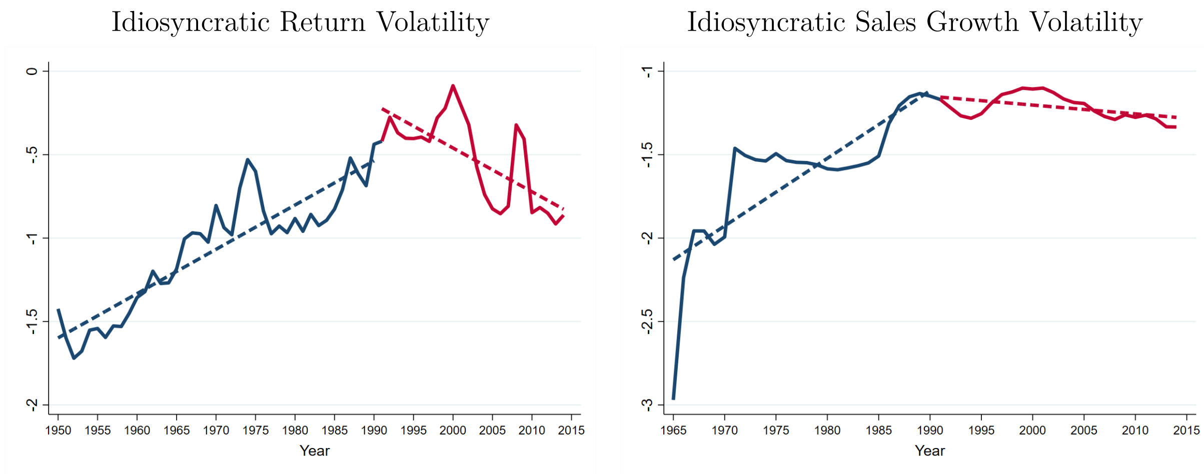
Note: This figure gives the shares of total Compustat sales, employment, and stock market cap that are represented in my Compustat-LEHD matched sample. The sample period spans 1991-2014.

Figure 1-9: Spread in Average Worker Skill Between Productive and Unproductive Firms Trends Upward



Note: This figure shows the differences for the average firm worker skill level (from (1.41) in the text) between firms in the top- and bottom-quartiles of labor productivity. The sample period spans 1991-2014.

Figure 1-10: Idiosyncratic Risk Has Increased Over the Long Term, But is Declining or Flat Within the 1991-2014 Period


Idiosyncratic Return Volatility

Idiosyncratic Sales Growth Volatility

Note: This figure plots the log average idiosyncratic stock return and sales growth volatility from Hartman-Glaser et al. (2019). The data span 1950-2014 for returns and 1965-2014 for sales. The 1991-2014 period covered in my paper is shown in red and the period before is shown in blue. Dashed lines give a separate linear time trend for the 1991-2014 and pre-1991 periods.

Table 1.10: Compustat Matched Sample and Overall Compustat Summary Stats

Panel A: LEHD-Compustat Matched Sample

	N	Mean	SD	P5	P50	P95
Log Assets	64000	5.862	1.946	2.887	5.752	9.327
Log Market Cap	63000	12.690	2.070	9.457	12.610	16.280
Log Sales	63500	0.646	1.861	-2.254	0.588	3.845
Log Employment	64000	5.872	2.009	2.769	5.843	9.260
Log Phys Capital	64000	4.914	2.205	1.542	4.808	8.734
Log Int Capital	63500	5.058	1.981	2.099	4.924	8.501
Excess Return	61500	0.119	0.542	-0.717	0.108	0.982

Panel B: Full Compustat Sample

	N	Mean	SD	P5	P50	P95
Log Assets	110000	5.396	2.202	1.991	5.246	9.334
Log Market Cap	164000	12.170	2.076	8.930	12.060	15.790
Log Sales	133000	-0.136	2.231	-3.689	-0.211	3.648
Log Employment	142000	5.180	2.408	1.380	5.133	9.200
Log Phys Capital	122000	4.428	2.643	0.366	4.284	8.992
Log Int Capital	137000	4.318	2.287	0.858	4.172	8.312
Excess Return	159000	0.104	0.518	-0.700	0.092	0.924

Note: This table provides basic summary stats for the full Compustat sample and my LEHD-Compustat merged sample. The sample period spans 1991-2014.

Table 1.11: Productive Firms Hire More Skilled Workers on Average

Dep Var: Log Firm Average AKM Worker Effects				
log VA/Worker	0.189	0.147	0.163	0.110
	(0.011)	(0.018)	(0.011)	(0.014)
Size Controls		X		X
Industry X Year FE			X	X
N	57500	57500	57500	57500
R^2 (within)	0.344	0.386	0.217	0.263

Note: This table shows regressions of log firm average AKM worker fixed effects (defined in (1.41)) on firm labor productivity as proxied by log value-added per worker. Controls for size include log employment, assets, and sales. Standard errors double clustered by year and industry in parentheses.

Table 1.12: Highly Productive Firms Have Lower Hiring Response to Employment Q After Accounting for Worker Skill

Dep Var: Hiring Rate	(1)	(2)	(3)	(4)
Q (Emp)	0.119	0.116	0.112	0.109
	(0.017)	(0.018)	(0.016)	(0.017)
Q (Emp) × log VA/Worker	-0.016	-0.014	-0.014	-0.013
	(0.003)	(0.004)	(0.003)	(0.004)
log VA/Worker	-0.033	-0.034	-0.039	-0.041
	(0.011)	(0.011)	(0.012)	(0.012)
Q (Emp) × Avg Skill		-0.01		-0.012
		(0.009)		(0.009)
Avg Skill		-0.017		-0.006
		(0.08)		(0.085)
Cash Flow			0.783	0.819
			(0.123)	(0.149)
Cash Flow × log VA/Worker			-0.155	-0.172
			(0.027)	(0.04)
Cash Flow × Avg Skill				0.124
				(0.117)
R-sq (within)	0.074	0.075	0.078	0.078
N	44000	44000	44000	44000

Note: This table shows regressions of investment in new hires on employment Q. The employment Q measure is interacted with a proxy for the costliness of labor/capital adjustment, the skill level of the firm's workforce (average worker AKM fixed effect). The last two columns add controls for cash flows as defined in Peters and Taylor (2017). All specifications include firm and year fixed effects. Standard errors double clustered by industry and year are in parentheses.

Table 1.13: Highly Productive Firms Have Lower Investment Response to Tobin's Q After Accounting for Intangible Capital Share

Dep Var: Total Invest Rate	(1)	(2)	(3)	(4)
Q (Tot)	0.066	0.059	0.054	0.048
	(0.009)	(0.008)	(0.008)	(0.009)
Q (Tot) × log VA/Worker	-0.007	-0.006	-0.006	-0.006
	(0.002)	(0.002)	(0.002)	(0.002)
log VA/Worker	0.011	0.01	-0.009	-0.01
	(0.003)	(0.003)	(0.002)	(0.002)
Q (Tot) × Log Intangible Share		-0.006		-0.005
		(0.003)		(0.002)
Log Intangible Share		0.008		0.008
		(0.004)		(0.004)
Cash Flow			0.223	0.203
			(0.045)	(0.044)
Cash Flow × log VA/Worker			-0.003	-0.002
			(0.01)	(0.01)
Cash Flow × Log Intangible Share				-0.021
				(0.01)
R-sq (within)	0.23	0.235	0.286	0.293
N	48500	48500	48500	48500

Note: This table shows regressions of investment in total capital on total Q from Peters and Taylor (2017) (panel B). The total Q is interacted with the firm's log intangible capital share of total capital as a potential proxy for the costliness of capital adjustment. The last two columns add controls for cash flows as defined in Peters and Taylor (2017). All specifications include firm and year fixed effects. Standard errors double clustered by industry and year are in parentheses.

Table 1.14: Calibration by Subperiod

Parameter	Explanation	1991-2002	2003-2014
α	(One minus) Labor returns to scale	0.22	0.22
β	Supply Elasticity Shifter	0.5	0.34
b	Reservation Wage	0.549	0.515
σ_z	Labor productivity volatility	0.141	0.149
$\bar{C}$	Convex adjustment cost level	0.38	0.745

Note: This table gives two alternative calibrations of the dynamic wage-posting monopsony model from section 3.1 in the main text. The calibrations are made to fit data moments for the 1991-2002 and 2003-2014 subperiods instead of the full sample moments from 1991-2014.

Table 1.15: Model Versus Data Moments by Subperiod

	1991-2002		**2003-2014**	
Moment	Model	Data	Model	Data
Separations Rate	0.36	0.32	0.29	0.27
Incumbent Premium	0.13	0.13	0.17	0.17
Inc/Rec Wage Pass. Ratio	1.67	2.10	2.11	1.59
Elasticity (Overall)	2.48	2.92	1.66	1.88
Elasticity (Inc.)	1.82	2.04	1.21	1.20
Elasticity (Rec.)	5.68	8.17	5.13	5.99
Elasticity (Q1)	5.04	4.86	3.16	3.25
Elasticity (Q2)	2.84	3.00	1.93	2.18
Elasticity (Q3)	2.20	2.95	1.51	1.55
Elasticity (Q4)	1.65	1.34	1.14	0.85
Log Labor Share	-0.49	-0.49	-0.52	-0.53
Log Labor Share (Q1)	-0.35	-0.21	-0.31	-0.23
Log Labor Share (Q2)	-0.46	-0.40	-0.48	-0.38
Log Labor Share (Q3)	-0.53	-0.50	-0.57	-0.53
Log Labor Share (Q4)	-0.64	-0.90	-0.72	-0.98

Note: This table compares model and data moments for my subsample calibrations of the dynamic wage-posting monopsony model from section 3.1 in the main text. "Inc" and "Rec" denote incumbents and recruits, respectively. The calibrations are made to fit data moments for the 1991-2002 and 2003-2014 subperiods instead of the full sample moments from 1991-2014. The label Q1 denotes the bottom quartile of labor productivity, Q4 denotes the top quartile, etc. The calibration of 5 model parameters explicitly targets the 7 moments that are not labor productivity quartile-specific, while the remaining 8 productivity quartile moments are not targeted.

Table 1.16: Baseline Elasticity Estimates Are Insensitive to Additional Controls for Common Market Shocks

Specification:	(1)	(2)	(3)	(4)	(5)	(6)
Employment						
Excess Return	0.116	0.115	0.116	0.115	0.116	0.116
	(0.009)	(0.010)	(0.009)	(0.010)	(0.009)	(0.009)
Competitor Emp Growth				0.043		
				(0.015)		
Competitor Wage Growth				0.002		
				(0.038)		
N	45000	45000	45000	45000	45000	45000
R^2	0.115	0.069	0.077	0.069	0.120	0.127
Wages						
Excess Return	0.046	0.048	0.046	0.046	0.045	0.045
	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)
Competitor Emp Growth				0.022		
				(0.004)		
Competitor Wage Growth				0.155		
				(0.030)		
N	45000	45500	45500	45000	45000	45000
R^2	0.413	0.383	0.398	0.389	0.421	0.425
Base Controls	X	X	X	X	X	X
Year FE		X		X		
Industry × Year FE	X				X	X
Labor Market × Year FE			X		X	X*
Implied Elasticity	2.525	2.397	2.536	2.492	2.568	2.585
	(0.305)	(0.291)	(0.289)	(0.300)	(0.307)	(0.303)
P-value (Difference)		0.052	0.850	0.527	0.202	0.114

Note: This table shows estimates of (1.6) in the main text when different sets of controls are considered. Baseline control variables include the contemporaneous change in AKM worker effects (worker skill), and lagged wage, employment, and asset growth at the firm level. Industry fixed effects are defined at the 3-digit NAICS level. Labor market fixed effects are from empirically defined labor market clusters (estimation described in section 1.7.1 of the appendix) with $K = 10$ labor market clusters per year. The "X*" in the 6th column indicates that I alternatively use $K = 20$ labor market clusters per year. Competitor emp and wage growth are an average of the growth rates of competitor firms that either hire from or whose employees are hired by the given firm. Local market controls include average changes in local labor market concentration, unemployment rates, and stock returns of firms operating in the same labor market. Wage data are from the LEHD, and the sample period spans 1991-2014. See section 1.3 in text for more details. "P-value (Difference)" denotes the p-value from the test that the given supply elasticity estimate is different from the baseline estimate in column 1. Standard errors double clustered by industry and year are in parentheses.

Table 1.17: Explanatory Power of Labor Market Proxies for Worker Flows, Wages, Stock Returns, and Employment

	K = 10	K = 20	NAICS2	NAICS3	K = 10 + NAICS3
Worker Flow Share	0.50	0.42	0.34	0.23	
Log Wage	0.62	0.65	0.51	0.60	0.70
Wage Growth	0.03	0.04	0.03	0.03	0.04
Excess Return	0.14	0.15	0.11	0.16	0.18
Log Emp	0.19	0.26	0.10	0.18	0.26
Emp Growth	0.03	0.03	0.03	0.03	0.04

Note: This table shows the explanatory power of different candidate labor market boundaries for several different variables. The first row reports the fraction of worker transitions between Compustat firms that occur within the same candidate market definition. The remaining rows report the adjusted R^2 from a regression of the given variable on market × year fixed effects. The first two columns show results for labor market clusters estimated with 10 or 20 clusters and the next two columns instead use either 2- or 3-digit NAICS codes. The last column reports adjusted R^2 values for labor market × year FEs and 3-digit NAICS × year FEs included simultaneously. Sample spans 1992-2013.

Table 1.18: Baseline Supply Elasticity Estimates are Robust to Alternative Shocks to Labor Demand

	Customer Ret	R&D Tax Credit	FF Resid	Patents	Earnings Ret
Elasticity	2.01	1.73	2.34	2.77	2.67
	(0.53)	(0.54)	(0.29)	(1.11)	(0.42)
Baseline Elasticity	2.12	1.96	2.53	2.53	2.53
	(0.33)	(0.18)	(0.30)	(0.30)	(0.30)
P-Value (Diff)	0.79	0.65	0.25	0.83	0.55
N	11500	21500	45000	45000	44500

Note: This table gives supply elasticity estimates using different labor demand shifters. All specifications have the baseline controls from (1.6) in main text, including industry × year fixed effects. “Customer Ret" uses the stock returns of the firm’s customers instead of the firm itself. “R&D Tax Credit” uses federal treatment of R&D expenses from Bloom et al. (2013) and Lucking et al. (2019) as a labor demand shifter. “FF Resid” uses cumulative log abnormal returns relative to the Fama-French 5-factor model augmented with momentum. “Patents” uses patent induced shocks from Kogan et al. (2020). Finally, “Earnings Ret” follows Daniel et al. (2019) in using stock returns in excess of the market return in a 4 day window around earnings announcement instead of stock returns. Elasticities computed as the ratio of the employment and wage responses to the given shock. The baseline elasticity is the elasticity estimate for the my main method using annual excess returns as a labor demand shock. In the first two columns I report baseline elasticities for the selected subsample for which the given shock is available. “P-Value (Diff)" gives the p-value on the differences between the elasticity using the given shock and for my baseline estimate. See appendix section 1.7.5 for further details. Standard errors double clustered by industry and year are in parentheses, and are computed by estimating the elasticity via two-stage least squares where wages are predicted in the first stage and employment is then regressed on predicted wages.

Table 1.19: Labor Supply Shocks from Union Elections and Changes in Non-Compete Enforceability Do Not Affect Supply Elasticity Estimates

Panel A: Union Elections

	(1)	(2)	(3)
Elasticity	1.65	1.64	1.72
	(0.33)	(0.32)	(0.34)
P-Val (Diff)		0.57	0.35
N	4200	4200	4200
Baseline Controls	X	X	X
Union Dummies		X	X
Union Dummies × Excess Return			X
Industry × Year FE	X	X	X
Union Win Excess Return	-0.08***		

Panel B: Non-Compete Changes

	(1)	(2)	(3)
Elasticity	2.37	2.36	2.4
	(0.50)	(0.50)	(0.52)
P-Val (Diff)		0.29	0.70
N	11000	11000	11000
Baseline Controls	X	X	X
Non-Compete Dummies		X	X
Non-Compete Dummies × Excess Return			X
Industry × Year FE	X	X	X

Note: Panel A of this table shows how elasticity estimates change when including controls for a union election occuring and for a union win in years t, $t-1$, or $t+1$. Elections are taken from a list compiled by Knepper (2020). The third column adds interactions of union dummies with the excess return in that year. Panel B of this table uses non-compete changes from the lists compiled by Jeffers (2019) and Ewens and Marx (2017). Following Ewens and Marx (2017) firms are considered treated if their headquarters state changes non-compete laws in a given year; non-compete controls include separate dummies for non-compete increases and decreases in the years t, $t-1$, and $t+1$, and the third column interacts these dummies with stock returns. “P-Value (Diff)" gives the p-value on the differences the elasticity using the given shock and for my baseline estimate. The sample is restricted to only those firms who ever had a unionization/non-compete event, and the baseline estimate is computed for the same sample. Standard errors double clustered by industry and year are in parentheses, and are computed by estimating the elasticity via two-stage least squares where wages are predicted in the first stage and employment is then regressed on predicted wages.

Table 1.20: Productive Firms Face Lower Supply Elasticities For Workers of All Skill Levels—Within Industry Productivity Sorts

Productivity:	Quartile 1	Quartile 2	Quartile 3	Quartile 4	P-val 4-1	R-sq	N
Whole Firm							
Employment	0.114	0.103	0.106	0.109			
	(0.012)	(0.012)	(0.010)	(0.011)	0.641	0.127	43500
Wages	0.028	0.039	0.047	0.088			
	(0.004)	(0.003)	(0.004)	(0.010)	0.000	0.421	43500
Elasticity	4.028	2.620	2.241	1.234			
Low Skill Workers							
Employment	0.123	0.118	0.131	0.138			
	(0.013)	(0.013)	(0.016)	(0.014)	0.285	0.109	43500
Wages	0.014	0.017	0.020	0.028			
	(0.003)	(0.003)	(0.002)	(0.004)	0.016	0.289	43500
Elasticity	8.786	6.837	6.660	4.950			
Middle Skill Workers							
Employment	0.111	0.098	0.107	0.103			
	(0.014)	(0.012)	(0.010)	(0.013)	0.444	0.084	43500
Wages	0.015	0.018	0.019	0.029			
	(0.002)	(0.002)	(0.002)	(0.003)	0.005	0.182	43500
Elasticity	7.283	5.281	5.743	3.580			
High Skill Workers							
Employment	0.099	0.092	0.094	0.086			
	(0.013)	(0.012)	(0.008)	(0.011)	0.266	0.076	43500
Wages	0.049	0.063	0.075	0.116			
	(0.005)	(0.004)	(0.006)	(0.011)	0.000	0.453	43500
Elasticity	2.047	1.458	1.250	0.738			

Note: This table contains supply elasticity estimates for firms sorted on log value-added/worker quartiles as in Table 1.4, except I now sort firms on labor productivity within 2-digit NAICS industry. Controls include 3-digit NAICS industry by year and productivity quartile fixed effects; lagged growth rates in wages, employment, and total assets; and the contemporaneous change in average worker skill level at the firm (see (1.41) for definition). Workers are placed into skill groups based on their estimated worker effects from a modified Abowd et al. (1999) style wage decomposition with time-varying firm fixed effects. Individuals in the bottom two quintiles of the cross-sectional distribution of worker effects are considered low-skilled, the third and fourth quintiles middle-skilled, and the top quintile high-skilled. Changes in average worker skill are computed within the population of workers considered in the specification. "P-val 4-1" gives the p-value from a test that the coefficients for firms in the top and bottom quartiles have equal values. Wage data are from the LEHD, and the sample period spans 1991-2014. Standard errors double clustered by industry and year in parentheses. See section 1.3 in main text for more details.

Table 1.21: Elasticities for Firms Sorted on Productivity—Robustness

Productivity:	Quartile 1	Quartile 2	Quartile 3	Quartile 4	P-val 4-1	R-sq	N
Panel A: 3-Year Horizon							
Employment	0.203	0.173	0.202	0.193			
	(0.016)	(0.016)	(0.017)	(0.017)	0.563	0.202	34000
Wages	0.026	0.036	0.046	0.079			
	(0.002)	(0.002)	(0.004)	(0.008)	0.000	0.506	34000
Elasticity	7.778	4.793	4.418	2.441			
Panel B: Wage Growth Using Changes in AKM Firm Effects							
Employment	0.119	0.089	0.120	0.106			
	(0.013)	(0.010)	(0.011)	(0.011)	0.240	0.127	43500
Wages	0.017	0.021	0.029	0.043			
	(0.002)	(0.002)	(0.002)	(0.005)	0.000	0.252	43500
Elasticity	6.953	4.238	4.068	2.474			
Panel C: TFP Sort							
Employment	0.109	0.087	0.102	0.120			
	(0.011)	(0.008)	(0.008)	(0.015)	0.482	0.139	37500
Wages	0.034	0.040	0.053	0.085			
	(0.004)	(0.004)	(0.007)	(0.011)	0.000	0.427	37500
Elasticity	3.212	2.194	1.906	1.424			
Panel D: LBD Employment Growth							
Employment	(0.085)	(0.084)	(0.095)	(0.092)			
	0.013	0.009	0.013	0.013	0.683	0.082	40000
Wages	(0.028)	(0.033)	(0.051)	(0.091)			
	0.003	0.003	0.004	0.010	0.000	0.422	43500
Elasticity	3.105	2.552	1.854	1.011			

Note: This table shows robustness checks for the elasticity estimates obtained from estimating variants of (1.8) in the main text. All equations use the baseline set of controls from (1.8), which includes industry × year fixed effects. Panel A estimates a variant of (1.8) where stock returns and employment/wage growth are at the 3-year horizon. In Panel B changes in AKM firm effects $\phi_{j,t}$ from estimating (1.38) replace growth in the firm-level average wage in estimating the supply elasticity. Panel C sorts on firm total-factor productivity from İmrohoroğlu and Tüzel (2014) instead of log value-added per worker. Panel D uses employment growth from Longitudinal Business Database employment figures instead of Compustat. "P-val 4-1" gives the p-value from a test that the coefficients for firms in the top and bottom quartiles have equal values. Wage data are from the LEHD, and the sample period spans 1991-2014. Standard errors double clustered by industry and year in parentheses. See section 1.3 in main text for more details.

Chapter 2

Technology-Skill Complementarity and Labor Displacement: Evidence from Linking Two Centuries of Patents with Occupations

Economists and workers alike have long worried about the employment prospects of occupations whose key tasks can be easily performed by a machine, robot, software, or some other form of capital that substitutes for labor.[1] These concerns have been exacerbated by recent breakthroughs in automation technologies (e.g., software, artificial intelligence, robotics) which have expanded the set of manual and cognitive tasks which can performed by machines and have occurred contemporaneously with an increase in income inequality and a fall in

[0] This chapter is joint work with Leonid Kogan, Dimitris Papanikolaou, and Larry Schmidt.

[1] Fear of technological unemployment is not new. In 350 BCE, Aristotle wrote: "[If] the shuttle would weave and the plectrum touch the lyre without a hand to guide them, chief workmen would not want servants, nor masters slaves." In 1811, skilled weavers and textile workers (known as Luddites) worried that mechanizing manufacturing (and the unskilled laborers operating the new looms) would rob them of their means of income. In 1930, Keynes described this type of potential labor market risk when he said, "We are being afflicted with a new disease of technological unemployment...due to our discovery of means of economising the use of labor outrunning the pace at which we can find new uses for labor." More recently, a McKinsey report estimated that between 400 million and 800 million jobs could be lost worldwide due to robotic automation by the year 2030.

the labor share of aggregate output.[2] Yet, despite the importance of these issues, systematic evidence for technological displacement remains elusive.[3] Our goal is to fill this gap: we leverage over a century and a half of data to propose and validate new metrics of workers' exposure to technological innovation and relate them to workers' labor market outcomes, both at the aggregate as well as the individual level.

To quantify workers' exposures to technical change we measure the similarity between the textual description of the tasks performed by an occupation and that of major technological breakthroughs. We identify the latter through the textual analysis of patent networks using the methodology of Kelly, Papanikolaou, Seru, and Taddy (2020). To estimate the distance between a breakthrough innovation and workers' task descriptions, we leverage recent advances in natural language processing that allow us to compute a measure of the similarity between documents that accounts for synonyms. By exploiting the timing of patent grants we can identify the extent to which certain worker groups (occupations) are exposed to major technological breakthroughs at a given point in time.

In sum, our indices capture the extent to which specific occupations are exposed to breakthrough innovations in a given year. We emphasize that, a priori, we are agnostic on whether innovations that are similar to tasks certain occupations perform are likely to be substitutes or complements. For that, we need to examine how our indicators correlate with labor market outcomes. A key advantage of our methodology is that it relies only on document text; as such, we are able to construct time-series indices of occupation exposures that span the last two centuries. For example, our technology exposure for "molders, shapers, and casters, except metal and plastic"—an occupation category which includes glass blowers as a sub-occupation—takes a relatively high value in the early 1900s because of similarity with patents such as US patent number 814,612, entitled "Method of Making glass sheets."

[2] For instance, one of the leading explanations for the increase in the skill premium is skill-biased technical change, whereas the decline in the labor share has been attributed to capital-embodied technical change. See Goldin and Katz (2008); Krusell, Ohanian, Ríos-Rull, and Violante (2000); Karabarbounis and Neiman (2013); Acemoglu and Restrepo (2020, 2018, 2021)

[3] Due to the difficulty of constructing broad measures of labor-displacive innovations, existing work has focused on analyzing specific instances in which the impact of a specific technology on workers can be identified (Atack, Margo, and Rhode, 2019; Feigenbaum and Gross, 2020; Akerman, Gaarder, and Mogstad, 2015; Humlum, 2019).

This patent relates to a technology for making glass called the cylinder machine, which allowed glass manufacturers to replace the labor of skilled hand glass blowers in favor of a highly mechanized and capital-intensive production process.[4]

Examining our technology exposure measure, we find that, prior to 1980, innovation was consistently associated with manual physical tasks; by contrast, the innovations of the late 20th/early 21st century have become relatively more related to cognitive tasks. This pattern is partly driven by the increased prevalence of breakthrough patents related to computers and electronics. Last, occupations that are associated with interpersonal tasks have consistently low exposures to innovation throughout the entire sample period.

Our analysis of technical change and labor market outcomes delivers several new findings. First, we find that workers most exposed to technological breakthroughs have experienced consistently negative labor market outcomes. Indeed, a higher rate of innovation in a given industry is associated with a decline in the labor share of output even as labor productivity rises. Comparing workers across occupations differentially exposed to technology improvements, we find that an acceleration of technical change is associated with declines in employment and wage earnings for affected workers. The negative correlation between employment and our technology exposure measure is largely consistent over time—starting from the Second Industrial Revolution of the late 19th century to the present. We find some evidence that the negative relation between our measure and employment is stronger in recessions—consistent with the literature on job polarization (Jaimovich and Siu, 2018). These negative relations are estimated using a combination of public-use Census micro-data, which are available over longer horizons, and confidential administrative data from the Census which include individual tax records starting in the early 1990s.[5]

[4] Jerome (1934) documents a dramatic transformation in the production process of the glass making industry as a result of the cylinder machine: "By 1905 many hand plants had gone out of business, wages of blowers and gatherers were reduced 40 per cent, and the new machine may be said to have achieved commercial success . . . in the quarter century following the introduction of machine blowing, the window-glass industry, one of the last strongholds of specialized handicraft skill, has undergone a technological revolution resulting in the almost complete disappearance of the hand branch of the industry and the elimination of two skilled trades and one semiskilled, and also the partial elimination of the skilled flatteners."

[5] Such a negative relation is not obvious ex-ante, since our approach could in principle identify both labor-saving as well as labor-enhancing technologies. If we are primarily interested in identifying labor-saving technologies, it is possible that our measure is diluted by mixing labor-saving and productivity-enhancing

Second, we exploit the richness of administrative data to examine how these relations vary with observable characteristics, by studying a unique panel of administrative earnings records from the US Social Security Administration which are linked with information on occupation and education from the Current Population Survey. Thus, relative to the literature which has mostly studied repeated cross-sections, we are able to measure a worker's occupation prior to the development of related technologies, then estimate how her earnings evolve in future years even if she switches employers, industries, and/or occupations. This analysis allows us to study the link between innovation and subsequent worker-level earnings growth rates and to address a number of potential concerns about composition effects driving our results. Our empirical analysis leverages the granularity of our patent-occupation measures to exploit variation at the industry-occupation level, i.e., in relative differences in the rate at which firms in different industries develop new technologies which are similar to a given occupation at the same point in time.

Across specifications, we find that the labor earnings of older or less educated workers are significantly more responsive to technological innovation than the average worker. More importantly, however, we find that the highest paid workers are also relatively more exposed. Specifically, workers at the top end of the earnings distribution—relative to their peers in the same occupation and in the same industry—also experience significantly larger declines than the average workers. Specifically, a one-standard deviation increase in our exposure measure is associated with a 2.5% decline in subsequent earnings compared to a 1.3% decline for the average worker. Our results are unchanged if we further control for common shocks to labor demand at the industry and occupation levels via industry-time and occupation-time fixed effects, respectively.

The fact that earnings of highly paid workers respond more to our technology exposure

innovations. To address this possibility, we exploit recent advances in topic modeling to construct a composite predictor from patent text whose purpose is to maximize the in-sample predictability of employment declines–i.e., to identify language consistent with labor saving innovations. After conducting this analysis, we find that our baseline innovation measures have a correlation of approximately 75% with this statistical factor. Comparing the performance of our measure to this benchmark, we found that both approaches lead to quantitatively similar negative outcomes in terms of both earnings and employment. On this basis, we conclude that our methodology primarily identifies labor-displacing innovations.

measure appears at first to be at odds with the canonical model of capital-skill complementarity. Indeed, a common proxy for worker skill is past income, so the fact that more highly paid workers experience larger declines seems surprising. We argue that it is not: though (a subset of) skilled workers as a group may benefit, if technological innovation is associated with skill displacement individual workers whose prior skills become obsolete may be left behind. Thus, higher paid (skilled) workers have more to lose. Consistent with this view, we find that, for the highest-paid workers (top 5% in past earnings relative to their peers in the same occupation and industry), a one-standard deviation increase in technology exposure is associated with a 1.26 percentage point increase in the probability of falling to the bottom decile of wage growth—compared to a 0.41% percentage point increase for the average worker.

To formalize this intuition, we introduce skill displacement in the canonical model of capital-skill complementarity (see, e.g. Krusell et al., 2000). We allow individual workers to supply both skilled and unskilled labor services; the quantity of skilled labor a given worker can supply depends on her skill. Improvements in technology are associated with increased likelihood of skill loss. Thus, even though skilled workers as a group (specifically, those that retain their skill) experience higher wage earnings following improvements in technology, unlucky individual workers can be left behind. On average, top workers may experience lower earnings growth following periods of technological advances if the increase in the likelihood of displacement is sufficiently high.

The calibrated model quantitatively replicates our new facts. In the model, increases in technological innovation lead to an increase in labor productivity and the skill premium—yet the labor share of output falls. On average, exposed workers in the model experience declines in wage earnings relative to peers whose skills are not related to the new technologies, and these differences are the largest for the highest paid workers. Importantly, these patterns emerge even though technology is more complementary to skilled that unskilled labor services. Following an innovation, high income workers whose skills are not displaced benefit from two forces: 1) complementarities with the more productive technology and 2) the fact that displacement of other high skilled workers' skills makes their expertise even more scarce and thus more valuable. Our model replicates our empirical result that workers with lower earnings also are hurt by the emergence of new technologies; specifically, this result obtains not because specific skills are displaced, but rather because of an increase in the supply of workers performing unskilled tasks which lowers wages.

With the model in hand, we also conduct some simple comparative statics exercises to consider the potential implications of an acceleration of the rate of innovation in the economy, consistent with the observed increase in the arrival rate of breakthrough patents which began in 1980. We consider two potential experiments. In the first case, we increase the arrival rate of new technologies but hold fixed the rate at which workers accumulate human capital. In the latter case, we also increase the rate at which workers acquire new skills so that the overall number of efficiency units of skilled human capital stays constant. In both model scenarios, such a shift generates increases in output, declines in the labor share, and increases in the skill premium in both the short and long run, all of which are consistent with trends in recent data from the US. In the former case, income inequality increases over the medium term but declines over the longer run because the higher rate of skill displacement eventually compresses the skill distribution by enough to offset the impact of a higher skill premium. In the latter case, this equalizing force is neutralized and thus income inequality increases in both the short and long run.

In sum, we provide and validate a new measure of workers' exposure to technological change that is based on the similarity between patent documents and worker job descriptions. Overall, we document a robustly negative relation between out technology exposure measure and subsequent labor market outcomes, results which are consistent with a model with capital-skill complementarity and skill displacement.

Our work contributes to the voluminous literature seeking to understand the determinants of rising inequality and the fall in the labor share. Existing work emphasizes the complementarity between technology and certain types of worker skills (Goldin and Katz, 1998, 2008; Autor, Levy, and Murnane, 2003; Autor, Katz, and Kearney, 2006; Goos and Manning, 2007; Autor and Dorn, 2013); or the substitution between workers and capital (Krusell et al., 2000;

Hornstein, Krusell, and Violante, 2005, 2007; Karabarbounis and Neiman, 2013; Acemoglu and Restrepo, 2020; Hemous and Olsen, 2021). Many models in this literature treat a worker skill as a fixed characteristic and study how demand for technologies affects differences in wages between groups with different ex ante skill levels. Our contribution is to provide a direct measure of technology exposure of specific workers and examine the extent to which advances in technology are associated with differences in their labor market outcomes. Motivated by our empirical evidence, our model allows for the possibility that gains from new technologies can displace the demand for specific expertise of workers skilled at tasks associated with older vintages, similar to a literature on vintage specificity of human capital (Chari and Hopenhayn, 1991; Violante, 2002; Deming and Noray, 2020) and models which seek to explain earnings losses from job displacement via obsolescence/loss of specific human capital (Neal, 1995; Kambourov and Manovskii, 2009; Huckfeldt, 2021; Braxton and Taska, 2020).

We are not the first to analyze the differential exposure of certain occupations to technical change. Autor and Dorn (2013); Acemoglu and Autor (2011); Autor et al. (2003) document the secular decline in occupations specializing in routine tasks, starting in the late 20th century. The key idea is that routine tasks can be easily codified into a sequence of instructions. Hence, such tasks are relatively more prone to labor-saving technological change than other more complex tasks. Despite the success that this literature has had in explaining which occupations have been exposed to technologies, and what have been the effects, it is still an open question how this exposure changes over time, which technologies relate to which types of tasks and which occupations, and whether or not technological unemployment is a robust phenomenon in other time periods. More recently, Webb (2019) also analyzes the similarity between patents and occupation task descriptions. Our work differs in both scope and aim. Webb (2019) focuses on automation and the future of work, and thus restricts attention to patents identified as being related to robots, AI, or software. As a result, the analysis in Webb (2019) is largely cross-sectional in nature as he focuses on a single technological episode—the rise of AI and robots. In contrast, we construct time-series indicators to understand the relation between innovation and employment over different technological episodes and its impact on workers with different characteristics. Further, focusing on the more recent period for which wage earnings data is available (after 1980), we show that the predictability of our measure for worker earnings is complementary to the information contained in the routine-task intensity measure of Autor and Dorn (2013), the AI and robotics occupation exposure measure of Webb (2019) and is not driven by industry-specific trends.[6]

A significant contribution of our work lies in its scope: we provide a measure of occupational exposure to technical change that spans the period from 1850 to 2010. An important advantage of our analysis is that it allows us to draw broad conclusions regarding the relation between technical change and worker outcomes over a long time period. Further, by constructing measures at the patent-occupation level, our approach allows us to study technological change at a highly granular level. Patents also have the advantage of being associated with specific timing (filing and approval dates) and are linkable to specific firms. To this end, our empirical analysis uses this granular information to compute measures of technological change at the industry-occupation level. Our results thus complement some earlier studies which, by narrowing their scope, are able to analyze the impact of worker earnings associated with specific technologies. For example, Atack et al. (2019) analyze how workers' task transitioned from hand to machine production in the late 19th century. Recently, Feigenbaum and Gross (2020) show that incumbent telephone operators were more likely to be in lower-paying occupations following the adoption of mechanical switching technology by AT&T. Akerman et al. (2015) and Humlum (2019) provides an in depth analyses of impacts of adoption of broadband internet and industrial robots, respectively, leveraging microdata on affected workers and firms.

Though labor income risk is not the primary focus of our study, we reach a similar conclusion as Kogan, Papanikolaou, Schmidt, and Song (2020): higher-paid workers face considerably greater risk in their labor income as a result of technological innovation. Though some of the conclusions are similar, these two papers ask different questions. Kogan et al.

[6] In related work, Mann and Püttmann (2018) and Dechezleprêtre, Hémous, Olsen, and Zanella (2021) use patent text with different classification algorithms to identify automation patents in more recent periods, though they do not relate these patents with specific occupations performing related tasks.

(2020) examine the dynamics of wage earnings in response to innovation by the workers' own firm or its competitors in the product market. Kogan et al. (2020) are interested in the extent to which profit-sharing motives transfer the risk of creative destruction from the firm owners to its workers. By contrast, we examine outcomes for all workers in the same industry, differentiated by their occupation (and its exposure to major innovations). Since our goal is to capture not only innovation by a firm but also the overall adoption of a technology in a given sector, the exact origin of these innovations are not particularly relevant.

2.1 Motivation: Technology, Productivity, and the Labor Share

We begin with a set of facts regarding the joint dynamics of aggregate measures of innovation, measured productivity, and the labor share that serve to motivate the remainder of our analysis. To do so, we obtain data on industry-level measures of output (value added), employment and the labor share from the NBER manufacturing database—which cover the 1958 to 2018 period.

Measuring the degree of technological innovation that takes place in a particular industry at a particular point in time is considerably more challenging. We do so by relying on patent data and closely follow the methodology of Kelly et al. (2020). In particular, Kelly et al. (2020) identify breakthrough innovations as those that are both novel (whose descriptions are distinct from their predecessors) and impactful (they are similar to subsequent innovations). They measure a patent's novelty as its dissimilarity with the existing patent stock at the time it was filed. In particular, they construct a measure of 'backward similarity'

$$BS_j^{\tau_b} = \sum_{i \in \mathcal{B}_{j,\tau_b}} \rho_{j,i}, \tag{2.1}$$

where $\rho_{i,j}$ is the pairwise cosine similarity (using TF-IDF weights) of patents i and j and $\mathcal{B}_{j,\tau_b}$ denotes the set of "prior" patents filed in the τ_b calendar years prior to j's filing. Patents with low backward similarity deviate from the state of the art and are therefore novel. Similarly, they measure a patent's impact by its 'forward similarity' as

$$FS_j^{\tau_f} = \sum_{i \in \mathcal{F}_{j,\tau_f}} \rho_{j,i}, \tag{2.2}$$

where $\mathcal{F}_{j,\tau_f}$ denotes the set of patents filed over the next τ_f calendar years following patent j's filing. The forward similarity measure in (2.2) estimates of the strength of association between the patent and future technological innovation over the next τ years.

The Kelly et al. (2020) patent-level measure combines forward and backward similarity to identify patents that are both novel and impactful,

$$q_j^{\tau} = \log FS_j^{\tau_f} - \log BS_j^{\tau_b}. \tag{2.3}$$

To create a time-series index, Kelly et al. (2020) remove calendar year fixed effects from (2.3) in order to adjust for shifts in language over time. After defining a 'breakthrough' patent as one that falls in the top 10% of the unconditional distribution of importance, they then construct a time series index as the number of breakthrough inventions granted in each year.

In brief, their measure of innovation in industry j in year t is defined as

$$\psi_{j,t} = \frac{1}{\kappa_t} \sum_{p \in \Gamma_t} \alpha_{j,p}. \tag{2.4}$$

To construct (2.4), we need to determine the set of breakthrough patents that are relevant to a given industry. We first identify the set Γ_t of 'breakthrough' patents as those that fall in the top 10% of the unconditional distribution of patent importance (based on their ratio of 10-year forward to 5-year backward similarity). We map patents to industries based on their CPC technology class using the probabilistic mapping constructed by Goldschlag, Lybbert, and Zolas (2020). Here, $\alpha_{j,p}$ denotes the probability of breakthrough patent p being assigned to industry j. Last, we scale (2.4) by US population κ_t and normalize to unit standard deviation.

We then estimate the following specification

$$\log X_{j,t+k} - \log X_{j,t} = \alpha(k) + \beta(k)\psi_{j,t} + \delta(k) Z_{j,t} + \varepsilon_{j,t}, \qquad k = 1 \ldots T \text{ years.} \tag{2.5}$$

We focus on four outcome variables: output (value-added); employment; labor productivity (value-added per worker); and the labor share. We examine horizons of up to $T = 6$ years. Controls include the lagged 5-year growth rate of the outcome variable and year fixed effects.

Figure 2-6 plots the estimated impulse response coefficients $\beta(k)$. Panel A illustrates that an one-standard-deviation increase in the degree of innovation $\psi_{j,t}$ in a given industry is associated with an approximately 2% increase in output over the next six years. However, this increase in output primarily reflects an increase in productivity: as we see from Panels B, the overall level of employment in the industry weakly falls. As a result, we see in Panel C that, in response to a one-standard-deviation increase in $\psi_{j,t}$, labor productivity in the industry rises sharply—by approximately 3% over the next six years. Panel D illustrates that the labor share of output in the industry falls.

In brief, an increase in our technology measure is associated with higher measured labor productivity and a decline in the labor share. This pattern strongly suggests that, on average, our technology measure $\psi_{j,t}$ captures innovations that likely act as substitutes, rather than complements to labor. That is, even though improvements in technology are associated with an increase in the measured productivity of labor, they are associated with declines in the labor share. The remainder of the paper builds on this idea and aims to provide further refinement to our technology measure by identifying which set of innovations are more likely to substitute for labor inputs. In the process of doing so, we also broaden the coverage from just manufacturing to all sectors in the economy.

2.2 Measuring Workers' Technology Exposure

Here, we construct a measure of technological innovation occurring at a particular point in time that is relevant for a particular set of worker tasks (occupation). Our interpretation is that this measure can be used to proxy for a worker's exposure to technological innovation. To do so, we add an additional source of cross-sectional variation to the Kelly et al. (2020) time-series measure of innovation discussed above. Specifically, we recognize that breakthrough innovations at a given point in time may be differentially related to particular occupations. To construct our measure, we rely on measuring the 'distance' between the description of the technology (from the patent document) to the description of the tasks that a given occupation performs (from the Dictionary of Occupational Titles, DOT). The remainder of this section describes our methodology.

2.2.1 Data and Methodology

We identify technologies that are relevant to specific worker groups as those that are similar to the descriptions of the tasks performed by a given occupation. We do so by analyzing the textual similarity between the description of the innovation in the patent document and the worker's job description.

We obtain text data for measuring patent/job task similarity from two sources. Job task descriptions come from the revised 4th edition of the Dictionary of Occupation Titles (DOT) database. We use the patent text data parsed from the USPTO patent search website in Kelly et al. (2020), which includes all US patents beginning in 1976, comprising patent numbers 3,930,271 through 9,113,586, as well as patent text data obtained from Google patents for pre-1976 patents. Our analysis of the patent text combines the claims, abstract, and description section into one patent-level corpus for each patent. Since the DOT has a very wide range of occupations (with over 13,000 specific occupation descriptions) we first crosswalk the DOT occupations to the considerably coarser and yet still detailed set of 6-digit occupations in the 2010 edition of O*NET. We then combine all tasks for a given occupation at the 2010 O*NET 6-digit level into one occupation-level corpus. See Appendix 2.6 for further details on cleaning and preparing the text files for numerical representation.

To identify the similarity between a breakthrough innovation and an occupation, we need to identify meaningful connections between two sets of documents that account for differences in the language used. The most common approach for computing document similarity is to create a matrix representation of each document, with columns representing document counts for each term (or some weighting of term counts) in the dictionary of all terms contained in the set of documents, and with rows representing each document. Similarity scores could then be computed simply as the cosine similarity between each vector of weighted or unweighted term counts:

$$\rho_{i,j} = \frac{V_i}{||V_i||} \cdot \frac{V_j}{||V_j||} \tag{2.6}$$

Here V_i and V_j denote the vector of potentially weighted terms counts for documents i and j.

This approach is often referred to as the 'the bag-of-words' approach, and has been used successfully in many settings. For example, Kelly et al. (2020) use a variant of this approach to construct measures of patent novelty and impact based on pairwise distance measures between patent documents. Since patent documents have a structure and a legalistic vocabulary that is reasonably uniform, this approach works quite well for patent-by-patent comparisons. However, this approach is less suited for comparing patent documents to occupation task descriptions. These two sets of documents come from different sources and often use different vocabulary. If we were to use the bag-of-words approach, the resulting vectors V_i and V_j would be highly sparse with most elements equal to zero, which would bias the distance measure (2.6) to zero.

The root cause of the problem is that the distance measure in (2.6) has no way of accounting for words with similar meanings. For example, consider a set of two documents, with the first document containing the words 'dog' and 'cat' and the other containing the words 'puppy' and 'kitten'. Even though the two documents carry essentially the same meaning, the bag of words approach will conclude that they are distinct: the representation of the two documents is $V_1 = [1, 1, 0, 0]$ and $V_2 = [0, 0, 1, 1]$, which implies that the two documents are orthogonal, $\rho_{1,2} = 0$.

To overcome this challenge, we leverage recent advances in natural language processing that allow for synonyms. The main idea behind this approach is to represent each word as a dense vector. The distance between two word vectors is then related to the likelihood these words capture a similar meaning. In our approach, we use the word vectors provided by Pennington, Socher, and Manning (2014), which contains a vocabulary of 1.9 million word meanings (embeddings) represented as (300-dimensional) vectors.[7] Appendix Section 2.6.2 contains a brief discussion of how the Pennington et al. (2014) word embeddings are constructed.

The next step consists of using these word vectors to construct measures of document similarity. To begin, we first construct a weighted average of the word embeddings with a document (a patent or occupation description). Specifically, we represent each document as a (dense) vector X_i, constructed as a weighted average of the set of word vectors $x_k \in A_i$ contained in the document,

$$X_i = \sum_{x_k \in A_i} w_{i,k} x_k. \tag{2.7}$$

A key part of the procedure consists of choosing appropriate weights $w_{i,k}$ in order to emphasize important words in the document.

In natural language processing, a common approach to emphasize terms that are most diagnostic of a document's topical content is the 'term-frequency-inverse-document-frequency' (TF-IDF). We follow the same approach: in constructing (2.7), we weigh each word vector by

$$w_{i,k} \equiv TF_{i,k} \times IDF_k. \tag{2.8}$$

The first component of the weight, term frequency (TF), is defined as

$$TF_{i,k} = \frac{c_{i,k}}{\sum_j c_{i,j}}, \tag{2.9}$$

[7] The basis for this word vector space is arbitrary; distances between word embeddings are only well-defined in relation to one another and a different training instance of the same data would yield different word vectors but very similar pairwise distances between word vectors. The two most popular approaches are the "word2vec" method of Mikolov, Sutskever, Chen, Corrado, and Dean (2013) and the global vectors for word representation introduced by Pennington et al. (2014). These papers construct mappings from extremely sparse and high-dimensional word co-occurence counts to dense and comparatively low-dimensional vector representations of word meanings called word embeddings. Their word vectors are highly successful at capturing synonyms and word analogies (vec(king) − vec(queen) ≈ vec(man) − vec(woman) or vec(Lisbon) − vec(Portugal) ≈ vec(Madrid) − vec(Spain), for example). Thus they are well-suited for numerical representations of the "distance" between words. The word vectors provided by Pennington et al. (2014) are trained on 42 billion word tokens of web data from Common Crawl and are available at `https://nlp.stanford.edu/projects/glove/`.

where $c_{i,k}$ denotes the count of the k-th word in document i—a measure of its relative importance within the document.

The inverse-document frequency is

$$IDF_k = \log\left(\frac{\#\text{ of documents in sample}}{\#\text{ of documents that include term } k}\right). \qquad (2.10)$$

Thus, IDF_k measures the informativeness of term k by under-weighting common words that appear in many documents, as these are less diagnostic of the content of any individual document.

In brief, $TFIDF_{i,k}$ overweighs word vectors for terms that occur relatively frequently within a given document and underweighs terms that occur commonly across all documents. We compute the inverse-document-frequency for the set of patents and occupation tasks separately, so that patent document vectors underweight word embeddings for terms appearing in many patents and occupation vectors underweight word embeddings for job task terms that appear in the task descriptions of many other occupations.

Armed with a vector representation of the document that accounts for synonyms, we next use the cosine similarity to measure the similarity between patent i and occupation j,

$$\text{Sim}_{i,j} = \frac{X_i}{||X_i||} \cdot \frac{X_j}{||X_j||} \qquad (2.11)$$

This is the same distance metric as the bag of words approach, except now X_i and X_j are dense vectors carrying a geometric interpretation akin to a weighted average of the semantic meaning of all nouns and verbs in the respective documents.

In sum, we use a combination of word embeddings and TF-IDF weights in constructing a distance metric between a patent document (which includes the abstract, claims, and the detailed description of the patented invention) and the detailed description of the tasks performed by occupations. Our methodology is conceptually related, though distinct, to the method proposed by Webb (2019), who also analyzes the similarity between a patent and

O*NET job tasks.[8]

2.2.2 Examples

To illustrate the effectiveness of our methodology in identifying links between technology and occupation task descriptions, we consider a few representative examples, some of which are summarized in Figure 2-1.

A key advantage of our measure is that it is available over long periods of time, and thus allows us to study very different technologies from three distinct periods of technological change–the Second Industrial Revolution of the late 1800's, the period spanning the from 1920s to around 1940, and the information technology revolution spanning the end of the 20th and beginning of the 21st centuries. For example, consider three patents in the list of breakthrough patents identified by Kelly et al. (2020). Patent 276,146, titled "Knitting Machine", was issued in the height of the Second Industrial Revolution in 1883. The occupation that is most closely related to this patent is "Textile Knitting and Weaving Machine Setters, Operators, and Tenders"; the next most similar occupation is "Sewing Machine Hand Operators", followed by "Sewers, hand". Next consider the patent for "Metal wheel for vehicles (1,405,358), which is issued in 1922. The occupation most closely related to this patent is "Automotive Service Technicians and Mechanics", with other production and metal machine workers following. Finally, we examine a patent from a very different era and representing a very different technology. The patent, entitled "System for managing financial accounts by a priority allocation of funds among accounts," is U.S. patent number 5,911,135 and was issued in 1999. The top occupations related to this patent are Financial

[8] Webb (2019) focuses on similarity in verb-object pairs in the title and the abstract of patents with verb-object pairs in the job task descriptions and restricts his attention to patents identified as being related to robots, AI, or software. He uses word hierarchies obtained from WordNet to determine similarity in verb-object pairings. By contrast, we infer document similarity by using geometric representations of word meanings (GloVe) that have been estimated directly from word co-occurence counts. Furthermore, we use not only the abstract but the entirety of the patent document—which includes the abstract, claims, and the detailed description of the patented invention. In addition to employing a different methodology, we also have a broader focus: we are interested in constructing time-series indices of technology exposures. As such, we compute occupation-patent distance measures for all occupations and the entire set of USPTO patents since 1836.

managers, credit analysts, loan interviewers and clerks.

We next perform the reverse exercise, where we fix a particular occupation, and list the most relevant innovations. The occupations we choose are cashiers, loan interviewers and clerks, and railroad conductors. Table 2.11 lists the top five patents that are linked to each of these occupations. Examining the patent tiles, we see that each one of these patents is directly related to the work performed by the given occupation. For example, one of the top patents for cashiers is “Vending type machine dispensing a redeemable credit voucher upon payment interrupt” (patent 5,055,657); the top patent for loan interviewers and clerks is titled “Automatic business and financial transaction processing system” (patent number 6,289,319). And finally, for rail road conductors, titled “Automatic train control system and method” (patent 5,828,979) is the top patent. In general the patents showing up on this list represent technologies that (1) relate to the work performed by individuals in that the occupation; and (2) if adopted, appear likely to be able to change the way that an occupation performs its core work functions and/or substitute for work done by that occupation.

In sum, these examples illustrate the ability of our method in identifying technologies that are related to a particular occupation. However, it is not immediately obvious whether these technologies benefitted workers in these occupations or whether they led to the displacement of workers. As a concrete example, consider US patent number 6,289,319, titled “Automatic business and financial transaction processing system”, and which as shown in Table 2.10 is the most similar patent to the “Loan Interviewers and Clerks” occupation. The DOT task description indicates that a person with this occupation “calls or writes to credit bureaus, employers, and personal references to check credit and personal references.” The description of this patent states that “Loan processing has traditionally been a labor-intensive business...the principal object of this invention is to provide an economical means for screening loan applications.” We interpret this innovation as an example of a technology which has high potential to be labor saving because it is intended to perform the same tasks performed manually by a worker in a more efficient manner.

However, there are also many counter-examples of new technologies that improve the productivity of tasks that workers are currently performing. Our exposure measure potentially also picks up these instances. For instance consider the occupation "Database Administrators" (SOC code 151141). According to the DOT, a database administrator "coordinates physical changes to computer databases." According to our distance measure, one of the most similar patents to this occupation is US patent number 5,093,782, entitled "Real time event driven database management system." This patent indicates that it provides "a database management system which is capable of supporting processes requiring the updating and retrieval of data elements at a high rate." This is likely to make the work of database administrators more efficient and hence looks more likely to be labor productivity-enhancing for this occupation.

Most likely, some of these technologies benefitted some workers at the expense of others. To illustrate the potential for such differential effects across workers of different skill levels, we consider two examples of labor saving technologies from Jerome (1934). First, consider two key innovations in the textile weaving industry during the early 20th century, the Barber-Colman warp-tying machine (patent 1,115,399) and the drawing-in machine (patent 1,364,091). Both of these technologies benefitted skilled workers at the expense of unskilled labor. Jerome (1934) notes that, the Barber-Colman warp-tying machine "will do the work of about 15 hand operators" while "it can be run by one tender." Similarly, he notes that "It is estimated that each (drawing-in machine) machine, requiring ordinarily the attention of one operator and half the time of an assistant, replaces from 5 to 6 hand drawers-in." Both of these patents are identified as breakthrough patents by Kelly et al. (2020). In terms of related occupations, our methodology identifies various types of textile workers as being the some of the most relevant.

However, not all labor-saving technologies benefit skilled labor. For instance, consider two major innovations in the window glass industry during the late 19th century—the Colburn sheet machine (patent 840,833) and the cylinder machine (patent 814,612). Following their introduction, the manufacturing process for window glass switched from being hand-made to being entirely mechanized by 1925. The displacement of skilled workers was rapid: by 1905 many hand plants had gone out of business, wages of blowers and gatherers were reduced 40 per cent. Jerome (1934) summarizes their impact thus: “In the quarter century following the introduction of machine blowing, the window-glass industry, one of the last strongholds of specialized handicraft skill, has undergone a technological revolution resulting in the almost complete disappearance of the hand branch of the industry and the elimination of two skilled trades and one semiskilled, and also the partial elimination of the skilled flatteners. The contest for supremacy now lies between the cylinder and the sheet machine processes.” Both of these patents are in the top 10% of the Kelly et al. (2020) measure. In terms of our methodology, we identify “glaziers” and “molders, shapers, and casters, except metal and plastic” as being among the most related occupations to these two patents. Specifically, the latter occupation, which corresponds SOC code 519195, has a sub-occupation called “glass blowers, molders, benders, and finishers”.

These two examples illustrate that the impact of a new technology on a given worker is not ex-ante obvious. Some technologies may replace un-skilled workers, while others may displace highly specialized and skilled workers. Indeed as Jerome (1934) notes, glass workers displaced by the sheet and cylinder machines in their time were considered to be members of skilled trade. Goldin and Katz (2008, Chapter 3) provide a number of historical examples where technological advances standardized tasks formerly performed by skilled artisans so that they could be performed by unskilled workers (see also Acemoglu, Gancia, and Zilibotti, 2012, for a related theoretical treatment of such a process). Further, new technologies may also generate demand for new skills—for example, the operators of the Barber-Colman warp-tying machine—hence their long-run effects may be different than their short-run impact. Hence, a significant part of the paper focuses on examining the correlation between our technology exposures and subsequent labor market outcomes.

2.2.3 Identifying Variation in Technology Exposures over Time

Our analysis so far delivers a measure of similarity between a given patent and a given occupation. The next step is to construct time-series indices of technological exposure at the occupation level. The key challenge in constructing a time-varying index lies in choosing how to appropriately quantify the 'degree' of innovation that occurs at a given point in time. One possibility would be to count the number of patents; however, this approach is unlikely to be fruitful, since not all patents are equally important. Various approaches have been proposed, which essentially weight patents by their forward citations (Hall, Jaffe, and Trajtenberg, 2005); estimates of their market value (Kogan, Papanikolaou, Seru, and Stoffman, 2017); or their textual similarity to prior and subsequent patents (Kelly et al., 2020).

For our purposes, we choose the Kelly et al. (2020) approach for two reasons: first, unlike forward citations, their measure is available for the entirety of our sample; second, it is available for all patents, and not just patents issued to publicly traded firms; and third, we are primarily interested in the contribution of a patent on the technology frontier rather than their private value to their firm.

We define our time series index of exposure of occupation i to technology at time t as

$$\eta_{i,t} = \frac{1}{\kappa_t} \sum_{j \in \Gamma_t} \tilde{\rho}_{i,j} \times \mathbf{1}(\tilde{q}_{j,t} \geq \tilde{q}_{p90}). \tag{2.12}$$

Our time-series index (2.12) aggregates our patent-occupation similarity scores across all breakthrough patents issued in year t. Specifically, we sum over the occupation-patent similarity score $\tilde{\rho}_{i,j}$ across the set of patents $j \in \Gamma_t$ that are issued in year t. We restrict attention to breakthrough patents, that is, patents whose Kelly et al. (2020) ratio of importance $\tilde{q}_{j,t}$ exceeds the (unconditional) 90th percentile $\tilde{q}_{p90}$.

When computing (2.12), we use an adjusted occupation-patent exposure metric $\tilde{\rho}_{i,j}$. Specifically, we perform the following adjustments to our raw occupation-patent exposure $\rho_{i,j}$ from (2.6). First, we remove yearly fixed effects. We do so in order to account for language and structural differences in patent documents over time and technology areas.[9] Second, we impose sparsity: after removing the fixed effects we set all patent × occupation pairs to zero that are below the 80th percentile in fixed-effect adjusted similarity. Last, we scale the

[9] Patents have become much longer and use much more technical language over the sample period, and the OCR text recognition of very early patents is far from perfect. We also slightly modify the Kelly et al. (2020) procedure by adjusting for the interaction between year and technology fixed effects, since some tech classes tend to have a naturally higher ratio of forward to backward patent similarity.

remaining non-zero pairs such that a patent/occupation pair at the 80th percentile of yearly adjusted similarities has a score equal to zero and the maximum adjusted score equals one.

2.2.4 Which occupations are more exposed?

Overall, we find that over the span of our sample, service-type occupations that specialize in person-to-person interaction scoring especially low on average exposure. In particular, Table 2.12 lists the top and bottom five occupations by average exposure over the time period spanning since 1850. The most exposed occupation is titled "Inspectors, Testers, Sorters, Samplers, and Weighers". The top occupations tend to be those working in production and manufacturing type jobs, which are commonly posited to be among the type of occupations most affected by new technologies. The least exposed occupations are mental health counselors, dancers, funeral attendants, judges, and clergy, all representing service job types that are unlikely to have the nature of their work substantially changed by new technologies.

Related to this point, Autor and Dorn (2013) argue that middle skill occupations have been significantly more exposed to technological innovation that low-skill workers. Using our direct measure of technology exposure, we can verify this is indeed the case. Specifically, we can examine how the technology exposure of occupations varies by 'skill levels' as proxied by wages. We obtain information on average wages by occupation from the Current Population Survey Merged Outgoing Rotation Groups (MORG, see Appendix 2.6.4 for more details). Given the short time dimension of the data (MORG starts in 1980), we focus on cross-sectional comparisons.

Figure 2-2 we plot exposures against average wage percentile ranks for the post-1980 period. Consistent with Autor and Dorn (2013), we see that the most exposed occupations tend to be found in the middle of the income distribution.

2.2.5 Long-run trends

We begin by documenting the types of occupation that are exposed to technological innovation, and how these exposures have shifted over time. To this end, we group each occupation into eight broad categories: service; sales and office; production, transportation, and material moving; natural resources, construction, and maintenance; management, business, and financial; healthcare practitioners; education, legal, community service, arts and media; and, computer, engineering, and science. Within each of these groups we take the average of $\eta_{i,t}$ and then scale across the eight groups each year so that the total sums to one. Figure 2-3 plots these shares over the entire sample (1850 to the present).

Examining Figure 2-3, there are two points worth noting. First, 'blue-collar' occupations, that is, those related to production and construction, have been consistently more exposed to technological progress than the others. Second, this trend has materially shifted over the recent decades, possibly due to the Information Technology (IT) revolution. Starting around the 1950s, there has been a secular increase in the relative technology exposure of 'white-collar' occupations. This rise is most visible in the increased exposure of computer, engineering, and science occupations. Sales and office occupations have also seen an increased relationship with innovation, as well as management/business occupations, though these two groups remain small in their overall exposure.

A useful way of summarizing these trends is examining the characteristics of occupations most exposed to technology at a given point in time. We first examine what kinds of tasks are performed by these occupations. Basing our analysis on Acemoglu and Autor (2011), we focus on four task categories: manual tasks (routine and physical); non-routine manual (interpersonal); routine cognitive; and non-routine cognitive.[10] Let $T_w(i)$ be an indicator function that equals 1 if occupation i has a score in the top quintile across occupations for task w; also denote by ω_i the Acemoglu and Autor (2011) employment shares for occupation i. We then construct an index $\lambda_{w,t}$ of the technological exposure of task category w as follows:

$$\lambda_{w,t} = \sum_i \eta_{i,t} \times T_w(i) \times \omega_i \tag{2.13}$$

[10] Because the routine manual and non-routine manual (physical) task scores are highly correlated and also move similarly with technological exposure, we group these two task types into one category by taking the average of the two scores. For similar reasons we take the average of non-routine cognitive (analytical) and non-routine cognitive (interpersonal) to get a non-routine cognitive score.

Figure 2-4 plots our measure of technology exposure $\lambda_{w,t}$, now separated for each of these task categories. The top panel (Panel A) plots these series in levels; the bottom panel (Panel B) plots their composition. The overall time-series behavior of our measures largely mimics the series of Kelly et al. (2020)—which is not surprising, given that we use their definition of breakthrough patents. We note three major innovation waves, lasting from 1870 to 1890; 1910 to 1930; and from 1970 to the present. The first peak corresponds to the beginning of the second industrial revolution, which saw technological advances such as the telephone and electric lighting and improvements in railroads. The second peak corresponds to advances in manufacturing, particularly in plastics and chemicals, consistent with the evidence of Field (2003). The latest wave of technological progress includes revolutions in information technology.

Importantly, we see that the first two major innovation waves were primarily related to occupations performing non-interpersonal manual tasks. By contrast, cognitive tasks are significantly less exposed. However, starting from the 1970s, there is a shift in the relative exposure of occupations emphasizing cognitive tasks, especially routine cognitive tasks. As a result, in the last few decades, these occupations are almost as exposed to innovation as occupations emphasizing manual tasks. This pattern is driven by information technology revolution that has led to the modern digitalization of the workplace. Occupations that relate to these type of innovations have a distinctly different task profile than the most prevalent technologies of past innovation waves. That said, even in the recent period, occupations emphasizing interpersonal tasks remain the least exposed to technological change. This pattern is consistent with the findings of Autor and Dorn (2013), who show that service occupations have increased in importance at the expense of occupations heavily exposed to automation, and also Deming (2017), who documents an increased importance of social skills in the labor market.

We next separate occupations by their education requirements. Specifically, we compute the share of workers in that occupation who have either completed a 4-year college degree or have attained a high-school diploma or lower in in a given year. For this analysis we crosswalk occupations to David Dorn's revised Census occ1990 level. We impute college grad and above/high school or below occupation shares for years between Census decades by linearly interpolating between the nearest available Census years and similarly interpolate occupational employment shares $\omega_{i,t}$ between Census years. We then let $S_{s,t}(i)$ be an indicator for whether occupation i is in the top quintile of the share of workers in education category s in year t. Due to data availability, we begin our analysis in 1950. We define the education exposure index $\zeta_{s,t}$ similarly to $\lambda_{w,t}$:

$$\zeta_{s,t} = \sum_i \eta_{i,t} \times S_{s,t}(i) \times \omega_{i,t} \tag{2.14}$$

Figure 2-5 presents our results. For most of the sample, we see that occupations requiring a college degree are significantly less exposed to innovation than occupations requiring a lower level of education. However, and consistent with the discussion above, we see that this pattern is shifting in the recent decades: towards the end of the sample, the difference in technology exposure between occupations requiring a college degree with those that do not has shrunk dramatically. This is especially evident in panel B of Figure 2-5 where we plot the composition rather than the levels of technological exposure. It's also important to note that this pattern is not driven by the increase in the share of workers with a college degree, since we assign occupations to the high education group based on their ranking in the cross-sectional distribution of occupational college grad shares. Rather, this pattern is driven by compositional shifts in the types of technologies being introduced, with an increasing share of technologies being targeted towards the tasks performed by relatively more educated occupations.

2.3 Technology Shocks and Labor Market Outcomes

In Section 2.1 we documented that breakthrough innovations are on average associated with increases in measured productivity and declines in labor share. Now, armed with an additional source of cross-sectional variation—differences in occupation exposure to specific technologies—we can examine a more direct link between technological progress and outcomes for specific workers.

We begin by focusing on group (i.e. occupation-level) outcomes in Section 2.3.1. The advantage of doing so is that we can examine the relation between our measure and labor market outcomes over a long period of time. The disadvantage of doing so is that patterns in some occupation-level outcomes, specifically wages, can mask important worker-level heterogeneity within the occupation. Thus, in Section 2.3.2 we focus on outcomes of individual workers using administrative data on worker earnings. Focusing on individual workers allows us to more closely trace the correlation between innovation and individual outcomes, while also enabling us to condition on certain observable characteristics. The disadvantage of the Census administrative data is that it is available only since the mid-1990s.

2.3.1 Occupation-level Evidence from repeated cross-sections

We begin by examining the relation between innovation and subsequent growth in the employment shares and average wage earnings of exposed occupations.

Data

The availability of public Census data allows us to examine employment outcomes over a long period of time (1850 to today). The Census surveys consist of repeated cross-sectional observations. Important for our purposes they contain information on occupations, which we can link to our innovation measure $\eta_{i,t}$ constructed in equation (2.12). Specifically, we use the 1950 Census occupation definition for pre-1950 Census years since the more updated 1990 Census classification scheme is only available in post-1950 Census years. We make use of the 1990 Census occupation classifications for the years they are available. We then crosswalk Census occupations to the David Dorn occ1990dd classification scheme using the crosswalk files provided on his website and aggregate our measure $\eta_{i,t}$ to the occ1990dd-level by averaging across 6-digit SOC codes within an occ1990dd code. This results in a Census-year by occ1990dd panel of occupation employment shares. Census records for the year 1890 were destroyed in a fire, and so the employment growth observations for the 20-year horizon in 1870 or for the 10-year horizon in 1880 are not available. The final dataset consists of an unbalanced panel of occupation–Census year employment shares and spans the Census years from 1850 to 2010. Appendix 2.6.4 provides additional details.

In addition, we use more recent data from the Current Population Survey Merged Outgoing Rotation Groups (MORG) which provides data on both wages and employment outcomes for the post-1980 period. We use the data to create a balanced panel of wage earnings and employment growth at the level of occupation and calendar year. We obtain the cleaned MORG extracts provided by the Center for Economic Policy Research (CEPR). In particular, we use the "wage3" variable that combines the hourly earnings for hourly workers and non-hourly workers, adjusts for top-coding using a lognormal imputation, and is constructed to match the NBER's recommendation for the most consistent hourly wage series from 1979 to the present.

Technology exposure and employment (1850–present)

We examine employment outcomes using the following specification,

$$\frac{1}{k}\Big(\log Y_{i,t+k} - \log Y_{i,t}\Big) = \alpha_0 + \alpha_t + \beta(k)\,\eta_{i,t} + \lambda Y_{i,t} + \varepsilon_{i,t}, \qquad k = 10, 20 \text{ years.} \tag{2.15}$$

The main dependent variable $Y_{i,t}$ is the employment share in total non-farm employment. Observations are weighted by the employment share of the given occupation and standard errors are clustered by occupation. As before, $\eta_{i,t}$ is normalized to unit standard deviation. All specifications include time fixed effects; depending on the specification, we include controls for the lagged 10-year employment growth rate.

Panel A of Table 2.1 presents our findings. We note that there is a strong and statistically significant negative correlation between our innovation measure η and subsequent changes in employment at the occupation level. The magnitudes are significant: a one-standard deviation increase in $\eta_{i,t}$ is associated with a 0.41% annualized decline in employment over the next 10 years and a 0.70% percent decrease in employment over the next 20 years.

We next allow the slope coefficients β to vary across Census years, focusing on horizons of $k = 20$ years. Figure 2-7 plots the point estimates of β for each Census year along with the 90% confidence intervals based on standard errors clustered by occupation. Examining the figure, we see that the point estimates are negative for all but the 1860 and 1940 Censuses, and are significant in 1880, 1910, 1920, 1950, 1970, 1980, and 1990. The magnitude of this correlation is also fairly stable over time, implying that occupations that are exposed to innovation have had consistently experienced employment declines over the entire 150 year period.[11]

One potential concern with these findings is that they reflect industry trends. To separate our findings from industry-level sources of variation, we next aggregate the Census data at the occupation by industry level over time. We use the 1950 Census industry designations, which are available the furthest back in time. Because Census industry codes are unreliable before 1910 we start our analysis using the data from the 1910 Census. We therefore estimate a slightly modified version of equation (2.15),

$$\frac{1}{k}\Big(\log Y_{i,j,t+k}) - \log Y_{i,j,t}\Big) = \alpha_0 + \beta(k)\eta_{i,t} + \delta Z_{t,j} + \varepsilon_{i,j,t}, \qquad k = 10, 20 \text{ years.} \tag{2.16}$$

The dependent variable $Y_{i,j,t}$ now represent the share of total non-farm employment for occupation i in industry j. The vector of controls $Z_{t,j}$ now contains time, or industry–time dummies depending on the specification, as well as lagged values of the dependent variable at the industry–occupation cell. The inclusion of industry–time allows us to isolate our findings from industry-specific trends in the sample.

Examining Panel B of Table 2.1, we see that the estimated slope coefficient on our innovation measure is consistently negative and economically and statistically significant. Overall, a one-standard deviation increase in $\eta_{i,t}$ is associated with a 0.60% to 0.89% decline in employment over the next 20-year horizon. The fact that this negative relation is essentially

[11] In 1930, John Maynard Keynes wrote: "We are being afflicted with a new disease of technological unemployment...due to our discovery of means of economising the use of labor outrunning the pace at which we can find new uses for labor." (Keynes, 1930). Indeed, the 1920–40 period corresponded with a large innovation wave that was associated significant declines in employment for occupations whose tasks were related to those innovations.

unaffected by the inclusion of industry-time fixed effects illustrates that our findings are not merely driven by the decline of certain industries which happen to employ workers with high technology exposure. Rather, much of the negative employment effects exist within, rather than between, industries. That said, the fact that the coefficients do attenuate slightly when including industry fixed effects indicates that high $\eta_{i,t}$ occupations tend to be employed in industries that have experienced overall employment declines.

Technology exposure, employment, and wages (1980–present)

We next turn our attention to the post-1980 period. We estimate the following specification

$$\frac{1}{k}\Big(\log Y_{i,t+k}) - \log Y_{i,t}\Big) = \alpha + \beta(k)\eta_{i,t} + \delta Z_{i,t} + \varepsilon_{i,t}, \qquad k = 5 \ldots 20 \text{ years.} \tag{2.17}$$

Here, $Y_{i,t}$ represents wage earnings or employment for a given occupation$i$ in calendar year t. The vector of controls $Z_{i,t}$ includes three lagged one-year growth rates of the dependent variable and time fixed effects. As before, $\eta_{i,t}$ is normalized to a unit standard deviation.

Figure 2-8 plots the estimated coefficients β along with 90% confidence intervals. We note that the responses for both wages and employment are strongly negative for all horizons. The point estimates are both economically and statistically significant and are comparable across horizons, suggesting these are permanent effects. Focusing on employment changes, a one-standard deviation increase in our technology exposure is associated with approximately a 1.1% annualized decline in occupation employment over the next five to twenty years. Similarly, our innovation measure also predicts a significant decline in wage earnings: a one-standard deviation increase in $\eta_{i,t}$ is followed by a decline in average wage earnings of approximately 0.2% to 0.3% per year over the same period.

Comparing the magnitude of employment declines in the MORG sample to the long-run (Census) results in Table 2.1 above, we note that the coefficient magnitudes are largely comparable. This is noteworthy in lieu of the fact the nature of breakthrough innovations in the post-1980 period is somewhat distinct than in the pre-1980 sample. As we saw in Figure 2-4, recent innovations are significantly more related to occupations performing routine cognitive tasks.

Recent work has argued that recessions are periods of technological transformation and thus accelerated automation of routine jobs (Jaimovich and Siu, 2018; Kopytov, Roussanov, and Taschereau-Dumouchel, 2018; Zhang, 2019). Consistent with this view, we next provide some direct evidence using our measure of technology exposure. In Figure 2-9 we see that occupations which were in the top quintile of $\eta_{i,t}$ in 1985 experienced stark declines in employment around the 1991, 2001, and 2007-2008 recessions, with a flatter but slightly declining profile in between recessions. Meanwhile, assigning occupations into the top quintile of routine-task intensity in 1985, we do see a persistent decline over the time period, but a much less pronounced pattern around recessions. This pattern is consistent with models of innovation-related job displacement where the opportunity to replace labor with an automation technology is a real option for firms. For example, Zhang (2019) shows that in a production based asset pricing model where firms choose to invest in labor-saving technologies, they choose to exercise this option when expected cash flows are temporarily low. Therefore the pattern exhibited in Figure 2-9 is consistent with employers replacing high $\eta_{i,t}$ workers with capital when the exercise value for doing so is high.

2.3.2 Individual-level evidence from panel administrative earnings data

Next, we turn our attention to individual worker outcomes. Doing so allows us to not only more directly link innovation to specific worker outcomes, but it also allows us to examine how this relation varies with observable worker characteristics. In particular, the detailed nature of administrative data allows us to examine how the relation between innovation and worker outcomes varies with proxies for worker skill, such as education or past earnings.

Data

We use a random sample of individual workers tracked by the Current Population Survey (CPS) and their associated Detailed Earnings Records from the Census—which contains their W2 tax income. The CPS includes information on occupation as well as demographic information such as age and gender. We limit the sample to individuals who are older than 25 and younger than 55 years old and to periods where the CPS interview date is less than 5 years old so that the occupation information is relatively recent.

We construct a measure of forward looking wage earnings growth following Autor, Dorn, Hanson, and Song (2014); Guvenen, Ozkan, and Song (2014); Kogan et al. (2020)

$$g_{i,t:t+h} \equiv w^i_{t+1,t+h} - w^i_{t-2,t}. \tag{2.18}$$

where $w^i_{t+1,t+h}$ refers to average *age-adjusted earnings* over the period, defined as

$$w^i_{t,t+h} \equiv \log\left(\frac{\sum_{j=0}^{h} \text{W-2 earnings}_{i,t+j}}{\sum_{j=0}^{k} D(\text{age}_{i,t+j})}\right). \tag{2.19}$$

refers to worker earnings net of life cycle effects. We focus on horizons of $h = 3, 5$, and 10 years. Appendix 2.6.5 provides more details on the construction of the data and the patch to patent information.

Given that the administrative data sample has a shorter time dimension and a larger cross-section than our other data, we make some modifications to our construction of our worker technology exposure measure η. First, we identify important patents based on their 5-year, as opposed to 10-year forward similarity. This allows us to extend the sample by five additional years, which helps with the short length of the sample. Further, to fully take advantage of the larger cross-section, we allow our baseline innovation exposure measure η to also vary by industry, by restricting attention to patents issued to firms in the same 4-digit NAICS industry as the worker. Letting j index patents as before; $\Gamma_{k,t}$ denote the set of patents issued in industry k in year t; $o(i)$ the occupation of individual i; and $k(i,t)$ the industry of individual i in year t, we redefine our time series index of exposure of worker i to technology at time t as

$$\eta_{i,t} = \frac{1}{\kappa_t} \sum_{j \in \Gamma_{k(i,t),t}} \tilde{\rho}_{o(i),j} \times \mathbf{1}(\tilde{q}_{j,t} \geq \tilde{q}_{p90}). \tag{2.20}$$

In brief, our technology exposure metric (2.20) is largely the same as before, except that we now focus only on breakthrough patents in the industry in which worker i is currently employed—that is, $\eta_{i,t}$ defined in (2.20) now varies by occupation, industry, and year instead of just occupation and year as (2.12).

Given these restrictions imposed by the Census-CPS merged file and the nature of our innovation data, we are left with approximately 1.2 million person-year observations spanning the period from 1988 to 2016. In terms of demographics, approximately 58% of the sample is male and 46% of the observations correspond to workers with a college degree. Table 2.2 provides more details on the distribution of age and worker earnings: the median worker in the sample is 41 years old and earns approximately \$58k per year. The distribution of earnings is rather skewed however: the average is equal to \$75k while the 5th and 95th percentiles are equal to \$14k and \$172k, respectively. The last set of rows of Table 2.2 summarize the distribution of our (age-adjusted) cumulative earnings growth (2.18). At a horizon of $h = 5$ years, the median is equal to 0.011 while the mean is -0.063; given that (2.18) corresponds to a log difference, the large dispersion in earnings induces the mean growth rate to be negative due to Jensen's inequality. That said, the distribution is also highly negatively skewed: the 5th percentile is equal to -0.968 log points while the 95th percentile is equal to 0.574.

An illustrative example

Before examining the correlation between our technology measure and subsequent wage growth, it is informative to look at particular examples. We choose the rise of e-commerce—and more specifically the automatic fulfillment of retail purchase orders. Advances in information technology and telecommunications have obviated the need for manual processing of customer orders. Using our patent-based indicators we can identify the 1996 to 2002 period as a period of significant innovation related to the tasks performed by order-fulfillment clerks. Examples of such breakthrough innovations early on include U.S. Patent 5,696,906 for "Telecommunication user account management system and method"; Patent 5,592,560 for "Method and system for building a database and performing marketing based upon prior shopping history"; or Patent

5,628,004 for “System for managing database of communication of recipients”. Appendix Table 2.18 contains a longer list—the top 10 most related breakthrough patents to order fulfillment clerks issued in the 1997 to 2000 period.

To study the impact of these breakthroughs on worker earnings, Figure 2-10 contrasts labor market outcomes for order fulfillment clerks versus a set of workers in two related occupations unlikely to be affected by these innovations: personnel and library clerks.[12] The top panel of Figure 2-10 plots the average real wage (in 2015 US dollars) differential across the two sets of workers—normalized to be zero in 1997. The bottom panel plots the difference in average technology exposure from (2.20) for workers employed as order clerks across industries relative to workers employed as personnel or library clerks.

Examining the top panel, we note that relative wage trends for the two occupations were fairly flat prior to 1997. However, since then they begin to systematically diverge. The bottom panel shows that this divergence coincided with significant breakthrough innovations that were related more to order than library clerks. Beginning in 1996, there is a sustained increase in (relative) innovation that persists for several years. By 2007, order clerks’ average annual wages had declined by nearly $30,000 relative to personnel/library clerks.

Our preferred interpretation of the patterns in Figure 2-10 is that improvements in the automatic processing of orders displaced workers whose primary task was to fulfill orders. Naturally, we cannot exclude the possibility that these patterns reflect industry- or occupation-specific trends. Fortunately, our empirical design outlined below allows us to control for both industry- as well as occupation-specific year fixed effects.

Innovation and worker earnings growth

We estimate the following specification,

$$g_{i,t:t+h} = \alpha + \beta(h)\,\eta_{i,t} + \delta Z_{i,t} + \varepsilon_{i,t}. \tag{2.21}$$

[12] The DOT indicates that an order clerk ‘Processes orders for material or merchandise received by mail, telephone, or personally from customer or company employee, manually or using computer or calculating machine...informs customer of any information needed...using mail or telephone. Writes or types order form, or enters data into computer, to determine total cost for customer. Records or files copy of orders."

Here, i now refers to a particular worker; as such, the left hand side represents the growth in her average earnings over the next h years compares to the last three. The main variable of interest $\eta_{i,t}$ now refers to the workers' technology exposure, specifically, the level of breakthrough innovations that are closely related to her occupation that are filed by firms in the same industry (based on NAICS 4-digit codes) as the worker. We examine earnings responses over the next $h = 3, 5$ and 10 years. The vector Z includes a rich set of controls that aim to soak up ex-ante worker heterogeneity. Specifically, we include various combinations of year, occupation and industry fixed effects—our most conservative specification interacts the latter two with calendar year to account for occupation- or industry-specific time trends. In addition, we include flexible non-parametric controls for worker age and past worker earnings as well as recent earnings growth rates.[13] Standard errors are clustered at the industry (NAICS 4-digit) level.

Table 2.3 summarizes our findings for the average worker in our sample. Overall, we find that workers' technology exposure is negatively related to their subsequent earnings growth. Panel A reports the estimated slope coefficients $\beta(h)$ for horizons of $h = 3, 5, 10$ years; different columns correspond to different fixed effect combinations. The magnitudes are both economically and statistically significant. Focusing on the 5-year horizon and the most conservative specification that includes both industry-year and occupation-year fixed effects, we see that a one standard deviation increase in innovation is associated with a 0.013 log point decline in average worker earnings over the next five years. These magnitudes increase with the horizon h, ranging from 0.011 to a 0.014 log point decline in average earnings at horizons of three and ten years, respectively.[14]

In brief, Panel A shows that the average effect of technology exposure on worker earnings

[13] We construct controls for worker age and lagged earnings $w^i_{t-4,t}$ by linearly interpolating between 3rd degree Chebyshev polynomials in workers' lagged income quantiles within an industry-age bin at 10-year age intervals. In addition, to soak up some potential variation related to potential mean-reversion in earnings (which could be the case following large transitory shocks), we also include 3rd degree Chebyshev polynomials in workers' lagged income growth rate percentiles, and we allow these coefficients to differ by gender as well as past income levels based on five gender-specific bins formed based upon a worker's rank relative to her peers in the same industry and occupation.

[14] We find no meaningful differences across genders: over the next five years, on average men experience a 0.013 log point decline and women experience a 0.011 log point decline.

is negative. This negative average effect, however, likely masks considerable heterogeneity in ex-post worker outcomes. To that end, in Panel B we next examine whether technology exposure is associated with increases in the second-moment of earnings growth: we estimate a modified version of (2.21), in which now the dependent variable is the *absolute value* of earnings growth. We see that technology exposure is associated with an increase in the second moment: the absolute value of earnings growth (2.21) increases by approximately 0.5 percentage points in response to a one-standard deviation increase in $\eta_{i,t}$, which corresponds to approximately a 3% increase relative to the sample median value of the dependent variable (0.157).

Similarly, Panel C examines the increase in the skewness of the earnings distribution, or equivalently the extent to which the negative average effects documented in Panel A are concentrated in a subset of workers. Specifically, we construct indicators for whether a given worker's income growth over a given horizon falls in the bottom 10th percentile of all our observations within a given year. Focusing on the same specification, we see that an increase in a worker's technology exposure is associated an economically significant increase in the likelihood of large earnings losses for affected workers: a one-standard deviation increase in $\eta_{i,t}$ is associated with a 0.4 percentage point increase in the likelihood that a worker's subsequent earnings growth is in the bottom 10th percentile.

Our findings in this section reinforce our findings in Section (2.3.1) that technology is associated with occupation-level declines in employment and wages. By tracking the earnings growth of individual workers, we can ensure that our findings on worker earnings are not driven by selection across occupations. In addition, individual-level data allows us to paint a more complete picture of how earnings losses are distributed across workers, even at a particular industry–occupation cell at a point in time. Building on this idea, the next section examines how these findings vary with worker characteristics.

Innovation, worker earnings and ex-ante heterogeneity

We next allow the effects to vary by observable worker characteristics. In terms of the heterogeneous effects of technological progress across workers, existing work has emphasized that (a) much of technological change is skill biased (see, e.g. Goldin and Katz, 2008, for a textbook reference); (b) an important component of human capital is likely specific to a technological vintage (Chari and Hopenhayn, 1991; Jovanovic and Nyarko, 1996; Violante, 2002). Accordingly, we focus on education and prior income as common proxies for worker skill; allowing the impact of technology to vary by worker age helps tease out the effect of vintage-specific human capital. In what follows, we re-estimate equation (2.21) and now allow the slope coefficient $\beta(h)$ to vary across worker sub-groups. For brevity we focus on the most conservative specification that includes both industry-year and occupation-year fixed effects. Appendix Tables 2.20 to 2.21 illustrate that results are similar across alternate specifications.

Table 2.4 examines how the effect of technology on worker earnings vary with education, specifically on whether the worker has a college degree. Focusing on mean growth rates across horizons—columns (1) to (3)—we see that an increase in technology exposure $\eta_{i,t}$ is associated with an economically and statistically significant decline in average earnings growth for both college and non-college educated workers. If we interpret education as a proxy for skill, the fact that earnings decline for both groups is somewhat at odds with the canonical model of skill-based technical change. That said, we do find that non-college workers experience a somewhat larger decline in average earnings growth than workers with a college degree, with the difference being marginally statistically significant (p-values range from 0.043 to 0.11 across horizons). Columns (4) and (5) show that both groups experience a similar increase in their second moment of earnings growth, while non-college educated workers experience a somewhat larger increase in the probability of large earnings declines than college educated workers (0.48 vs 0.35 percentage points).

Table 2.5 examines how the response of worker earnings growth to $\eta_{i,t}$ varies by prior income. In terms of point estimates, we find a U-shaped pattern: a given increase in a worker's technology exposure $\eta_{i,t}$ has the largest impact on the earnings growth of not only the least- but also the highest-paid workers (relative to their peers in the same occupation and industry). These estimates are noisy: we cannot reject the null that the response of earnings growth of workers at the bottom quartile is equal to the responses of workers in the 25th to 95th percentile. However, the response of the most-highly paid workers is both statistically as well as economically distinct from the workers in the middle group: a one-standard deviation increase in technology exposure is associated with a 0.025 log point decline in their average earnings growth—approximately twice as large as the effect on the average worker.

At face value, these facts seem at odds with the canonical model of skill-biased technical change: if technology is complementary to the labor input of skilled workers, to the extent that prior income is related to worker skill, one would expect to see that top workers should experience an increase in their average earnings (or at the very least a smaller decline). By contrast, the opposite pattern obtains. Comparing across columns (1) to (3) we see that these patterns are quantitatively similar across horizons, suggesting that these effects are highly persistent.

One potential reconciliation of the patterns in Table 2.5 with the standard view of technology-skill complementarity is allowing for vintage-specific human capital. That is, skill is not an immutable characteristic of the worker; it is the result of experience and learning by operating a particular technology. When new technologies are introduced, some of that accumulated knowledge becomes obsolete: skilled workers in the old technology need not remain skilled in the new. If that is the case, we would expect skilled workers to face greater earnings risk in response to increased rate of technological innovation due to the possibility of skill displacement. Consistent with this view columns (4) and (5) of Table 2.5 show that top earners face significantly greater labor income risk than the average worker in response to an increase in their technology exposure. Focusing on the last column, a one-standard deviation increase in $\eta_{i,t}$ is associated with a 1.26 percentage point increase that these workers experience a large earnings decline (earnings growth in the bottom 10-the percentile) which is approximately three times higher than the average worker.

Table 2.6 provides additional support to the idea of vintage-specific human capital by examining how these effects vary with worker age. Older workers are both more likely to have accumulated skills in existing technology and also less likely to be able to become familiar with new production methods. Accordingly, we see that older workers (those in the 45 to 55 range) experience significantly greater declines in earnings growth (0.02 to 0.025 log points across horizons) relative to workers aged 35–45 (0.9 to 1.4 log point decline) or 25–35 (0.4 to 0.9 log point decline). Columns (4) and (5) similarly show that earnings risk in response to an increase in technology exposure is increasing in age: a one-standard deviation increase in $\eta_{i,t}$ is associated with a 0.9 percentage point increase in the likelihood of a large earnings decline for workers aged 45–55, compared to a 0.2 percentage point increase for younger workers.

Discussion

In brief, we find that a given improvement in technology leads to lower earnings growth across all workers. These patterns, together with the positive correlation to industry productivity and the decline in the labor share documented in Section 2.1 is consistent with the view that most of the breakthrough innovations in the sample are labor-saving, that is, they partly replace tasks performed by workers. The standard view is that increases in automation are more likely to affect low-skill workers, since low-productivity workers are likely to be replaced first. Some of our findings are consistent with this view: we find some evidence that workers without a college degree and the lowest-paid workers experience larger than average earnings declines in response to technological innovation (though the latter difference is not statistically significant).

However, some of our findings are harder to reconcile with a view that skill is an immutable characteristic of the worker. Workers that are more highly paid relative to their peers in the same occupation and industry experience on average significantly larger earnings declines than the average worker; further, our findings suggest that much of these average decrease reflects an increase in the probability of large earnings losses as opposed to a small consistent decline in earnings. Put differently, the distribution of earnings losses is heterogenous: a subset of skilled workers experience large earnings declines rather than the entire group experiencing small declines.

This pattern, together with the increased earnings response of older workers suggests a role of vintage-specific human capital, or equivalently for technology making certain worker skills obsolete. The model in the next Section 3.1 formalizes and quantifies this idea more fully. That said, one caveat in interpreting these patterns is that the period covered by the Census-CPS administrative data coincides with the rise of very specific technologies, namely ICT. Thus, we should be careful when extrapolate these findings to other periods of rapid technological progress.

2.3.3 Additional Results and Robustness Checks

Here we discuss a number of additional results that frame our work to the existing literature and explore the extent to which our measure underestimates the degree of labor-displacive innovations.

Comparison to Existing Measures of Exposure to Technical Change

Our work is not the first to construct occupation-level measures of exposure to technological change (Autor and Dorn, 2013; Webb, 2019). A key advantage of our measure relative to existing work is that it also incorporates time-series variation. That said, it is instructive to explore the extent to which it contains additional information regarding cross-sectional differences in technology exposures. Here, we explore this question, and compare the performance of our technology measure in predicting wage and employment declines relative to the routine-task intensity and the measure of occupation-level offshorability from Autor and Dorn (2013), and the Webb (2019) measures of exposure to robotics or software.

To compare our different approaches, we estimate a long-difference cross-sectional specification similar to Webb (2019) as follows

$$\frac{1}{k}\Big(\log Y_{i,t+T}) - \log Y_{i,t}\Big) = \alpha + \alpha_j + \beta\eta_{i,1980} + \delta X_i + \epsilon_{i,j} \tag{2.22}$$

Here i indexes occupations and j indexes industries. In estimating (2.22), we combine information on wages and employment in the 1980 Census and the 2012 ACS. In particular, we use the 1980 Census and 2012 ACS data from Deming (2017), which are reported at the occupation by industry by education level, and aggregate the data to industry by occupation. Thus, the dependent variable denotes either the log change in employment or the change in log wages over the 1980-2012 time period. We include industry fixed effects α_j to account for industry specific shocks that may be correlated with occupational outcomes. Controls X_i include occupation employment share in 1980, occupation log wage in 1980, three indicators for the occupations education level in 1980, the routine-task intensity and the measure of occupation-level offshorability from Autor and Dorn (2013), and the Webb (2019) measures of exposure to robotics or software patents, depending on the specification. We weight observations by the employment share in 1980 and cluster standard errors by industry.

Tables 2.9 reports our findings for employment (Panel A) and wages (Panel B). Examining the first row of each panel, we see that the point estimates (and statistical significance) of β are essentially unaffected by including the Autor and Dorn (2013) or Webb (2019) measures. We conclude that cross-sectional differences in $\eta_{i,t}$ contain independent information relative to these alternate cross-sectional metrics.

Alternative approaches to measuring labor displacement

Our technology exposure measure is constructed based on the similarity between patents and tasks performed by a specific occupation. Ex-ante, it is not entirely obvious whether a high level of similarity is likely to capture complementarity or substitution between the technology and the tasks performed by labor. Even though we are finding a consistently negative relation between our technology exposure measure and subsequent labor market outcomes, it is possible these effects are muted because our measure mixing labor-saving and labor-enhancing innovations.

To explore this possibility, we next compare the performance of our measure to a purely statistical predictor that is calibrated to predict employment declines in-sample. To do so, we leverage recent advances in topic modeling to construct a composite predictor from patent text whose purpose is to maximize the in-sample predictability of employment declines. This measure is akin to a principal component; it has no straightforward economic interpretation, but it rather provides a statistical upper bound on how large the labor-displacive effects could be on a factor that is constructed from the text of breakthrough patents.[15] The correlation between our baseline measure $\eta_{i,t}$ and the statistical predictor constructed to represent exposure to labor-saving technologies is approximately 73 percent.

We then compare the performance of our baseline measure based on patent-task similarity to the in-sample performance of this statistical factor in the wage and employment regressions in Figure 2-8—which can be viewed as an upper bound in the ability of patent text to predict employment and wages. We find that the performance of our baseline measure is close to this upper bound: the annualized employment and wage declines predicted by this statistical displacement factor are 1.25% and 0.25%, respectively—compared to 1.12% and 0.20% for our technology measure based on patent-task similarity. Appendix **??** provides more details. We conclude that our technology measure based on patent-task similarity primarily captures labor-saving innovations.

Robustness

We perform several checks to explore the extent to which our findings are specific to a particular period. Appendix Table **??** shows that our findings on the long-run negative relation between technology exposure and employment at the occupation level are robust across sub-samples. That said, Appendix Table 2.15 shows that this negative relation is primarily driven by innovation waves: we separate the sample into two sub-samples, where we

[15] We build on the approach proposed by Cong, Liang, and Zhang (2019), which is well-suited to prediction exercises using large-scale textual data. In brief, this approach can be summarized as follows. We first extract the 500 most important common factors (topics) from the text of breakthrough patents using the approach of Cong et al. (2019) and the vector representations of word embedings discussed in Section 2.2. We then use these 500 textual factors to form a single predictor that is optimized to predict occupation declines in-sample. To do so, we examine the univariate performance of each factor in predicting employment declines, and then form a linear combination (the first principal component) of the predictors that are statistically significant negative predictors at the 5%. We also construct a labor-enhancing factor using the converse exercise. Appendix **??** describes the procedure in detail.

define as innovation waves the 20 year periods beginning in years 1880, 1910, 1920, and 1980, 1990—with the remaining years representing non-innovation wave periods (for a discussion of the 1920s, see e.g. Field, 2003).

Further, we verify the robustness of our findings to alternative specifications. Appendix Tables 2.20 through 2.22 show that our results on heterogenous responses by education, age, and prior income are largely robust to different combinations of fixed effects and horizons over which we measure worker earnings.

That said, one particular caveat in interpreting the heterogenous patterns we uncover using the Census-CPS administrative data is that they are estimated during a very specific time period which coincides with the rise of very specific technologies (i.e. ICT). The extent to which similar patterns obtain more broadly during earlier periods is an open question. As a step towards this direction, we re-estimate our long-run employment results and allow the coefficient to vary by worker age (the only variable consistently reported since the 1850s in the population sensus). In particular, we re-estimate 2.15 but now add a second panel dimension, worker age. That is, the unit of analysis is not worker occupation-age groups and when we compute employment growth rates we track a particular cohort. For instance, to compute the 20-year growth rate in employment in 1900 for workers aged 20–29 in occupation i, we compare it to the employment of 40–49 year old workers in occupation i in 1920.

Table 2.23 reports our findings. Panel A focuses on the full sample. We see that the employment growth of older workers in a specific occupation is significantly more exposed than the employment growth of younger workers (difference has a p-value of 0.015). In terms of magnitudes, a one-standard deviation increase in technology exposure is followed by a 1.1% annual decline in employment for older workers over the next twenty years, compared to a 0.7% annual decline for younger workers. However, Panel B shows that this pattern is largely driven by the latter part of the sample. Specifically, there are no differences in employment outcomes across age groups during the 1850–1920 period. In the 1930 to 1960 period, there is some evidence that older workers are significantly more exposed, but the results are too noisy to infer meaningful differences (p-value of 0.13). By contrast, focusing on the post-1970 period, the difference is between younger and older workers becomes larger and statistically significant. We conclude that we cannot rule out the possibility that our worker heterogeneity results are somewhat specific to the ICT revolution: it is entirely possible that older workers were significantly more displaced than younger workers by ICT.

2.4 Model

Here we provide a model that features skill-biased technological change and allows for skill displacement. The model contains a continuum of workers who supply high- and low-skill labor inputs. Consistent with the literature, the output of the high-skill labor input is more complementary to technology than the output of the low-skill input. As a result, improvements in technology lead to an increase in the wages of high-skill workers relative to the wages of low-skill workers. This is a feature of the standard model.

We extend the model to allow for skill displacement. In particular, improvements in technology may render some of workers' skills obsolete—a specific worker may lose a part of her skill when the technology frontier improves. As a result, even though wages of skilled workers rise in response to technology, an incumbent skilled worker may experience a decline in wages.

2.4.1 Setup

We model the output of a given industry. Output is produced by three factors of production: low-skilled labor L, high-skilled labor H, and intangible capital (technology) ξ. For simplicity, we will abstract from labor growth and model output per capita Y as

$$Y_t = \left[\mu \Big(H_t \Big)^{\sigma} + (1-\mu) \left(\lambda \Big(\xi_t \Big)^{\rho} + (1-\lambda) \Big(L_t \Big)^{\rho} \right)^{\sigma/\rho} \right]^{1/\sigma} \tag{2.23}$$

Here, ρ denotes the elasticity of substitution between technology and unskilled labor and σ denotes the elasticity of substitution between skilled labor and the composite output of technology and unskilled labor.[16] Since the total mass of workers is normalized to one, equation (2.23) also refers to labor productivity (output per worker).

The factor ξ is the stock of intangible capital/knowledge embodying the technology used for producing output Y, similar in spirit to Acemoglu and Restrepo (2018). When we map our empirical analysis to the model, we will interpret our technology exposure metric $\eta_{i,t}$ as a shock to ξ, which however affects only a subset of workers involved in the production of Y—the model equivalent to 'occupations', described below. Keeping with the literature, we expect technology to be more complementary to skilled labor relative to unskilled labor, so we will impose the condition that, in relative terms, shifts in technology are more complementary to skilled than unskilled labor, or equivalently that

$$\sigma < \rho < 1. \tag{2.24}$$

Put differently, technology ξ is a better substitute for unskilled rather than skilled labor.

Technology ξ evolves exogenously according to

$$d\xi_t = -g\,\xi_t\,dt + \kappa\,d\,N_t. \tag{2.25}$$

Technology improves according to the process N_t whose increments are Poisson with arrival rate $\omega\,dt$. Recall that we have set up output in per capita terms. As such, the negative drift term in equation (2.25) reflects the fact that ξ also a per-capita quantity and population grows at rate g. Given (2.25), the level of ξ is stationary with a long-run mean equal to $\kappa\,\omega/g$.

Workers are heterogenous along two dimensions. In particular, there is a unit mass of workers differentiated by their type $\theta \in [0,1]$, which determines their endowment of high- and low-skill labor inputs; workers also vary in their ability to acquire new skills $s = \{l, h\}$. Specifically, each worker can provide θ units of skilled labor H and $1-\theta$ units of unskilled

[16] In setting up (2.23), we have included technology and unskilled labor in the inner nest, and skilled labor in the outer nest. This is in contrast to the formulation in Krusell et al. (2000); Eisfeldt, Falato, and Xiaolan (2021), but note that the comparison with these two papers is imperfect, since ξ denotes intangible capital (technology) rather than physical equipment (machines).

labor L. As a result, the total supply of skilled labor as a share of population is equal to

$$H_t = \int_0^1 \theta \, p_t(\theta) \, d\theta, \tag{2.26}$$

where $p_t(\theta)$ is the measure of workers of skill level θ at time t. Since we normalize the total supply of labor to one,

$$L_t = 1 - H_t. \tag{2.27}$$

In addition, workers vary in their ability to acquire new skills—that is, increase their skill level θ. The share of workers who cannot acquire new skills s_l can produce only in the low-skill task so $\theta = 0$. The remaining share of workers $s_h = 1 - s_l$, have skill $\theta \in (\underline{\theta}, 1)$ that evolves over time due to learning by doing and technological displacement according to

$$d\theta_{i,t} = m \, \theta_{i,t} \, dM_{i,t} - h \, \theta_{i,t} \, d_{i,t} \, d \, N_t, \tag{2.28}$$

Here, $dM_{i,t}$ is a Poisson jump with arrival rate $\phi \, dt$ that reflects the stochastic acquisition of new expertise. Since we limit $\theta \in (\underline{\theta}, 1)$ for these workers, we impose reflecting boundaries at $\underline{\theta}$ and 1.

Importantly, the last term in equation (2.28) captures the displacive effect of the arrival of new technologies ($d \, N_t = 1$). There is a stochastic element in how technology improvements affect workers: this uncertainty is captured by $d_{i,t}$, which is a random variable with support on the unit interval and is independent of θ_{it}. For now, we assume that $d_{i,t}$ i.i.d. distributed across agents and follows a binomial distribution $d \in \{0, 1\}$ with $Prob(d = 1) = \alpha$. More generally, we could allow the distribution of d_i to vary with certain worker characteristics such as age or education. Affected workers experience a proportional loss in their human capital (skill) by a factor h. Last, workers of each type die at Poisson rate $\delta \, dt$ and are replaced by newborn skilled workers with either zero skill ($\theta = 0$) or the minimum level of skill ($\theta = \underline{\theta}$) for skilled workers with probabilities s_l and $1 - s_l$ respectively.[17]

[17] Our formulation for θ is related to Jones and Kim (2018) in that the skill of an individual worker grows on average over time but occasionally resets to a lower level.

Given our assumptions (2.26)–(2.28), the aggregate supply of skilled labor H_t increases with learning, decreases as skilled older workers are replaced with unskilled young workers, and decreases temporarily following periods of rapid technological progress. The latter effect captures the idea that technological improvements may be associated with lower output in the short run as agents in their economy need to upgrade their skills to fully take advantage of new innovations—similar in spirit to Brynjolfsson, Rock, and Syverson (2018).

The current wage of an individual worker with skill level $\theta_{i,t}$ is equal to

$$w_{i,t} = W_{L,t} + \theta_{i,t} \left(W_{H,t} - W_{L,t} \right). \tag{2.29}$$

In equilibrium $W_{H,t}$ and $W_{L,t}$ are equal to the marginal product of skilled and unskilled labor, respectively

$$W_{H,t} = \frac{\partial Y_t}{\partial H_t}, \quad \text{and} \quad W_{L,t} = \frac{\partial Y_t}{\partial L_t}. \tag{2.30}$$

In sum, we provide a model in which the skill premium increases with the level of technology, yet the wage earnings of individual skilled workers can fall as they potentially are displaced. We discuss the model calibration next.

2.4.2 Model Calibration

Here we discuss how we fit the model to the data.

Methodology

The model has a total of 14 parameters. We choose these parameters via a mixture of calibration and indirect inference. Specifically, we choose $s_l = 0.375$ so that workers with only low-skill labor inputs constitute the lowest income bin (25% of the sample), and half of the second-lowest. Since m and ϕ are not separately identified, we set the learning rate $m = 0.03$; when choosing the grid for θ, we assume that skilled workers human capital $\theta \in (0.03, 1)$. Last, we set the worker exit rate, at $\delta = 2.5\%$ which corresponds to a 40 year average working life.

To estimate the remaining 10 parameters $\Theta = \{\mu, \lambda, \rho, \sigma, \phi, \alpha, \kappa, \omega, h, g\}$, we target the mean level of the skill premium, the response of labor productivity and the labor share to changes in technology estimated in Section 2.1, and the response of worker earnings growth and likelihood of large wage declines conditional on levels of prior income, estimated in Section 2.3.2. Since the model has no mechanism for delayed responses, whereas in the data the diffusion of technology likely takes some time, we match the model responses on impact to the empirical responses over five years. Table 2.8 summarizes the 14 statistics that we target.

To obtain some intuition for how the model parameters are identified, we next discuss how these quantities help identify model parameters. In the model, the skill premium defined as the ratio of wages for the high-skill versus the low-skill labor input equals

$$\frac{W_{H,t}}{W_{L,t}} = \frac{\mu}{(1-\mu)(1-\lambda)} \left(\frac{H_t}{L_t}\right)^{\sigma-1} \left(\lambda \left(\frac{\xi_t}{L_t}\right)^{\rho} + (1-\lambda)\right)^{\frac{\rho-\sigma}{\rho}}. \tag{2.31}$$

Examining (2.31), we see that as long as technology is more complementary to high-skill than low-skill labor inputs ($\rho > \sigma$) and we hold fixed the supply of skilled and unskilled labor H and L then the skill premium is increasing with the level of technology ξ. Thus, just like the standard model, our model generates an increase in the skill premium over the long run during periods when the rate of technology rises faster than average. However, in our model the supply of high- and low-skill inputs H and L varies in the short run, due to skill displacement (2.28). This process leads to drop in H/L and thus a further increase in the skill premium in the short-run.

When mapping the model to the data, we define the skill premium as the mean ratio of earnings of workers in the 75th vs the 25th percentile. This ratio combines information on the ratio W_H/W_L and the ergodic distribution of θ. In terms of identifying model parameters, the mean level of the skill premium thus helps identify the factor share parameters μ and λ and the elasticities ρ and σ. Further, it affects the parameters driving the ergodic distribution of θ, namely ω, ϕ, h and α.

The labor share of output in the model can be written as

$$\frac{W_{H,t}\,L_t + W_{L,t}\,H_t}{Y_t} = \frac{(1-\lambda)\,(1-\mu)\left(\lambda\left(\frac{\xi_t}{L_t}\right)^{\rho} + 1-\lambda\right)^{\frac{\sigma}{\rho}-1} + \mu\,\left(\frac{H_t}{L_t}\right)^{\sigma}}{(1-\mu)\left(\lambda\left(\frac{\xi_t}{L_t}\right)^{\rho} + 1-\lambda\right)^{\frac{\sigma}{\rho}} + \mu\,\left(\frac{H_t}{L_t}\right)^{\sigma}} \tag{2.32}$$

The response of the labor share (2.32)— and output (2.23)—to increases in technology ξ is ambiguous in the model. These both depend on the extent to which different tasks contribute to output (μ and λ); technology-labor complementarity (ρ and σ) and the response of H and L to a technology shock (which depends on h and α). What helps with identification in our case is our finding that technology improvements are associated with declines in the labor share of output and an increase in output per worker (see Section 2.1). The fact that output/productivity and the labor share respond with opposite signs helps narrow down the set of admissible parameters quite significantly.

To identify the parameters involved in the dynamics of worker skill acquisition and displacement in (2.28), we also target the heterogeneity in earnings responses to changes in technology (see Section 2.3.2). Specifically, we target the mean earnings growth responses (column (2) in Table 2.5), and changes in the probability of large declines in wage earnings (column (5) in Table 2.5). To construct the analogue. Recall that, in the data, there is a U-shape relation in mean responses, with the highest-paid workers and lowest-paid workers experiencing the largest earnings declines in response to technology shocks. In terms of higher moments though, we find that the highest-paid workers are far more likely than other groups to experience large earnings declines. In the model, whether higher-paid workers are more exposed to technology is largely ambiguous.

To see this, we can derive the following decomposition for wage earnings growth in the model over any horizon h:

$$\frac{w_{i,t+h} - w_{i,t}}{w_{i,t}} = \underbrace{\frac{w_{l,t}}{w_{l,t} + \theta_{i,t} s_{p,t}}}_{\substack{\text{low skill}\\ \text{income share}}} \underbrace{\frac{\Delta_h w_{l,t+h}}{w_{l,t}}}_{\substack{\text{low skill}\\ \text{wage chg}}} + \underbrace{\frac{\theta_{i,t} s_{p,t}}{w_{l,t} + \theta_{i,t} s_{p,t}}}_{\substack{\text{high skill}\\ \text{income share}}} \left[\frac{s_{p,t+h}}{s_{p,t}} \cdot \underbrace{\frac{\Delta_h \theta_{i,t+h}}{\theta_{i,t}}}_{\substack{\text{skill}\\ \text{displacement}}} + \underbrace{\frac{\Delta_h s_{p,t+h}}{s_{p,t}}}_{\substack{\text{high skill}\\ \text{wage chg}}}\right] \tag{2.33}$$

where $s_{p,t} = W_{h,t} - W_{l,t}$.

As we see from the last term in brackets in (2.33), whether the highest-earning workers experience larger declines depends on whether the increase in the skill premium in is sufficient to offset the loss of worker skill $\theta_{i,t}$ due to skill displacement—see equation (2.28). For the high-income (i.e. θ) workers, the primary income risk in the model comes from having human capital displaced, while the lowest-income workers (those in the s_l group with $\theta = 0$) face income losses from changes in wages. Improvements in technology lead to an increase in the skill price of H and a drop in the skill price of L because of both differences in complementarity and skill displacement—since workers fall down the ladder following a shock, H is scarcer and L is more abundant. These effects depend on the size of human capital losses and increases, as well as the associated skill prices following displacement, including h, ϕ, ω, λ, μ, σ, and ρ.

Mapping the empirical regressions in Table 2.5 to the model entails two challenges: first, our technology measure varies at the industry and occupation level whereas the model refers to a single industry; second, our empirical specifications include occupation, industry, and time fixed effects so the main coefficients are also identified by comparing to workers in other occupations or industries. To narrow the gap between the model and the data, we construct the closest equivalent to a regression coefficient in the model as follows. We first calculate a set of wage responses that vary by income bins that match the empirical equivalents. Within each income bin, we compute wage growth for exposed ($d_{i,t} = 1$) and unexposed ($d_{i,t} = 0$) workers in the case of a technology shock occurring ($dN_t = 1$) or not ($dN_t = 0$). The equivalent of the regression coefficient in the model is the coefficient of wage growth on the interaction between a shock occurring and the worker being exposed, while separately controlling for exposure and shock dummies and everything interacted with income bins.[18]

We calibrate the remaining 10 model parameters by minimizing the distance between the

[18] When constructing these regression coefficients in the model, we use the ergodic distribution of wage growth, so we take into account the share of exposed workers α, the frequency of technology shocks ω and the likelihood each worker falls in a given income bin.

output of the model $\hat{X}(\Theta)$ and the data X,

$$\hat{\Theta} = \arg\min_{\Theta} \left(X - \hat{X}(\Theta)\right)' W \left(X - \hat{X}(\Theta)\right). \tag{2.34}$$

Our choice of weighting matrix W emphasizes percent deviations of the model vs the empirical values and places relatively more weight in the aggregate moments.

Table 2.7 summarizes our parameter choices. Similar to Krusell et al. (2000); Eisfeldt et al. (2021), we find that technology is a good substitute for the low-skill labor input ($\rho = 0.74$) whereas the high-skill labor input is complementary to technology ($\sigma = -0.12$). Technology shocks are relatively frequent ($\omega = 1.56$) and sizeable ($\kappa = 0.32$). Importantly, however, the model features a modest degree of skill displacement: workers who fall down the ladder only lose $h = 6\%$ of their existing level of θ. That said, these losses are pervasive: the probability of skill loss conditional on a shock is $\alpha = 32\%$. Further, these skill losses are transient: workers are able to acquire skills (increase θ) at an average rate of $m\,\phi = 7.2\%$ per year.

Model Fit and Discussion of the Mechanism

Examining Table 2.8, we see that the model does a good job matching the target statistics, including the labor share and responses of aggregate quantities to technology shocks. Specifically, the model is able to capture the fact that output and labor productivity rise following a technology shock whereas the labor share falls. In addition, the model is able to largely replicate both the marginal effect of a shock on exposed workers as well as the U-shape pattern of coefficients by income rank.

Figure 2-11 plots the impulse responses generated by the model in response to a one-standard deviation shock to the level of technology ξ (panel A). Panel B shows that this improvement in technology leads to a 2.5% rise in output/productivity on impact. By contrast, Panel C shows that the labor share declines by approximately 1.5%. This decline in the labor share is driven by a combination of two factors. First, as we see in Panel D, the quantity of the high-skill labor input declines by approximately 2.5% as workers' skills are displaced. This fall is temporary, as H gradually increases to skill acquisition. Since the wages for the high-skill task exceed the wages of the low-skill task, the total wage bill in the economy falls. Second, Panel E shows that improvements in technology are associated with decline in the price of the low-skill labor input (W_L) which further depresses the labor share; by contrast, even though the price of the high-skill labor input rises in Panel F, the rise is not sufficient to cause the labor share to rise because H falls. These movements in skill prices are driven by a combination of two forces: first, the high-skill input is complementary to ξ whereas the low-skill input is a substitute; second, skill prices change in response to the reduction in the effective supply of H due to skill displacement.

Figure 2-12 summarizes the distributional impact of technology shocks in the cross-section of workers. Panel A focuses on differences in growth rates in response to a technology shock relative to the no-shock counterfactual. The blue bars correspond to unexposed workers ($i.e.d_i = 0$). For these workers, the only effect in play is changes in skill prices. Low-income workers supply only the low-skill labor input L. Since the price W_L of the low-skill input falls, these workers experience a decline in wages. By contrast, the high-income workers supply mostly the high-skill input H; since the price of the high-skill input W_H rises, these workers experience an increase in wages. Absent skill displacement would be the only impact on wage growth in the model—and would be similar to the models in Krusell et al. (2000) and Eisfeldt et al. (2021). Yet, such a model would be unable to produce the empirical patterns in Table 2.5.

The orange bars in Panel A of Figure 2-12 correspond to the wage growth of exposed (i.e. $d_i = 1$) workers following a shock relative to the no-shock counterfactual. These workers experience the same change in skill prices as the unexposed workers, but they are also subject to skill displacement (loss of human capital θ). As a result, the wage growth of the high-income exposed workers is markedly different than the wage growth of the unexposed high-income workers: despite the fact that skill prices W_H rise, these workers experience a fall in wages due to loss of human capital θ. Further, just like the data, their wages fall significantly more than the low-income workers, implying that this loss in skill is significant.

Panel B Figure 2-12 plots the equivalent of the regression coefficient in the model, that is, the OLS coefficient of a regression of wage growth on a shock and exposure dummy, controlling for income. Since these slope coefficients are estimated using the ergodic distribution of wages at the model steady state, which factor in the relative size of the different worker groups and the frequency of technology shocks they cannot be expressed as simple functions of the coefficients in Panel A. However, they display a similar pattern as the orange bars: improvements in technology have an asymmetric effect on the wages of exposed workers. The workers most affected are the high-income workers—with some mild evidence of a U-shaped response.

In brief, Figure 2-12 summarizes the impact of technology of wages, which is a combination of shifts in skill prices and changes in the quantity of human capital. The combination of these effects generate the U-shape in earnings losses we see in column (2) of Table 2.5. The lowest-income workers have $\theta = 0$, and as a consequence have wages which fall dramatically relative to a non-shock period. Workers in the middle part of the income distribution experience some loss of human capital and suffer from the decline in the price of low-skill labor input W_L, but these losses are partly offset from the increases in the high-skill price W_L. Workers at highest income group has the farthest to fall: these workers who are exposed to technology experience the largest wage declines of anyone in the model due to skill displacement. By contrast, unexposed workers who stay at the top of the ladder following a technological innovation see large wage increases due to higher W_H—which results from scarcer H and the complementarity of H and ξ.

2.5 Conclusion

We develop a new method for identifying the arrival of labor-displacive innovations. Our time series indicators of worker technology exposure date are available since the mid 19th century and are available at a high level of granularity—industry and worker occupation. Examining the type of worker tasks most exposed to innovation, we find that while non-routine manual (physical) and routine-manual tasks have been highly exposed throughout the last 150+ years, the innovations of the information technology revolution in the post-1980 period saw an increased relationship with cognitive tasks.

More importantly, we find that our technology exposure measures are consistently negatively related to workers' future labor market outcomes, both at the group (occupation) but also at the individual level. Using a panel of administrative data on worker earnings, we show that the earnings of older and less educated workers are more responsive to our technology exposure measure, which is in line with the existing view of technology-skill complementarity. By contrast, our finding that the earnings of more highly paid workers (relative to their peers in the same industry and same occupation) respond more to our technology measure is somewhat at odds.

We reconcile these patterns with the standard view by allowing for skill displacement in the standard model of technology-skill complementarity. Our calibrated model is able to quantitatively replicate our main findings: improvements in technology are associated with increases in productivity but a decline in the labor share; lower earnings across workers of all income groups. In the model, the earnings of high-income workers respond more to technology improvements because these workers have further to fall: the loss of skills following technological progress is sufficient to offset any wage gains associated with higher skill prices.

Overall, we provide long-run evidence that the technological displacement of labor has been a persistent phenomenon over the past century and a half. Our findings illustrate the utility of our technology exposure indicators that can be used to study an array of questions in economics. That said, we should emphasize that our indicators are constructed largely from the perspective of incumbent workers and are primarily intended to capture technological substitution of existing tasks. A likely feature of technological progress that we are missing is that it facilitates the creation of new tasks and occupations. Building on our work, Autor, Salomons, and Seegmiller (2021) represents a promising step along that direction.

References

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics, Volume 4. Amsterdam: Elsevier-North*, pp. 1043–1171.

Acemoglu, D., G. Gancia, and F. Zilibotti (2012). Competing engines of growth: Innovation and standardization. *Journal of Economic Theory 147*(2), 570–601.

Acemoglu, D. and P. Restrepo (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review 108*(6), 1488–1542.

Acemoglu, D. and P. Restrepo (2020). Robots and jobs: Evidence from us labor markets. *Journal of Political Economy forthcoming*.

Acemoglu, D. and P. Restrepo (2021). Tasks, automation, and the rise in us wage inequality. Working Paper 28920, National Bureau of Economic Research.

Akerman, A., I. Gaarder, and M. Mogstad (2015). The skill complementarity of broadband internet. *The Quarterly Journal of Economics 130*(4), 1781–1824.

Arora, S., Y. Liang, and T. Ma (2017). A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*.

Atack, J., R. A. Margo, and P. W. Rhode (2019). Automation of manufacturing in the late nineteenth century: The hand and machine labor study. *Journal of Economic Perspectives 33*(2), 51–70.

Autor, D., D. Dorn, G. H. Hanson, and J. Song (2014). Trade adjustment: Worker-level evidence. *The Quarterly Journal of Economics 129*(4), 1799–1860.

Autor, D., A. Salomons, and B. Seegmiller (2021). New frontiers: The origins and content of new work, 1940–2018. Working paper, MIT.

Autor, D. H. and D. Dorn (2013). The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review 103*(5), 1553–97.

Autor, D. H., L. F. Katz, and M. S. Kearney (2006). The Polarization of the U.S. Labor Market. *American Economic Review 96*(2), 189–194.

Autor, D. H., F. Levy, and R. J. Murnane (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics 118*(4), 1279–1333.

Bianchi, F. (2016). Methods for measuring expectations and uncertainty in markov-switching models. *Journal of Econometrics 190*(1), 79–99.

Braxton, J. C. and B. Taska (2020). Technological change and the consequences of job loss.

Brynjolfsson, E., D. Rock, and C. Syverson (2018). The productivity j-curve: How intangibles complement general purpose technologies. Working Paper 25148, National Bureau of Economic Research.

Chari, V. V. and H. Hopenhayn (1991). Vintage human capital, growth, and the diffusion of new technology. *Journal of Political Economy 99*(6), 1142–1165.

Cong, W., T. Liang, and X. Zhang (2019). Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information.

Dechezleprêtre, A., D. Hémous, M. Olsen, and C. Zanella (2021). Induced automation: evidence from firm-level patent data. *University of Zurich, Department of Economics, Working Paper* (384).

Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Deming, D. J. and K. Noray (2020). Earnings dynamics, changing job skills, and stem careers. *The Quarterly Journal of Economics 135*(4), 1965–2005.

Eisfeldt, A. L., A. Falato, and M. Z. Xiaolan (2021). Human capitalists. Working Paper 28815, National Bureau of Economic Research.

Feigenbaum, J. and D. P. Gross (2020). Automation and the fate of young workers: Evidence from telephone operation in the early 20th century. Working Paper 28061, National Bureau of Economic Research.

Field, A. J. (2003). The most technologically progressive decade of the century. *American Economic Review 93*(4), 1399–1413.

Goldin, C. and L. Katz (2008). *The Race Between Education and Technology.* Harvard University Press.

Goldin, C. and L. F. Katz (1998). The origins of technology-skill complementarity. *The Quarterly Journal of Economics 113*(3), 693–732.

Goldschlag, N., T. J. Lybbert, and N. J. Zolas (2020). Tracking the technological composition of industries with algorithmic patent concordances. *Economics of Innovation and New Technology 29*(6), 582–602.

Goos, M. and A. Manning (2007). Lousy and lovely jobs: The rising polarization of work in britain. *The Review of Economics and Statistics 89*(1), 118–133.

Guvenen, F., S. Ozkan, and J. Song (2014). The Nature of Countercyclical Income Risk. *Journal of Political Economy 122*(3), 621–660.

Hall, B. H., A. Jaffe, and M. Trajtenberg (2005). Market value and patent citations. *The RAND Journal of Economics 36*(1), 16–38.

Hemous, D. and M. Olsen (2021). The rise of the machines: Automation, horizontal innovation, and income inequality. *American Economic Journal: Macroeconomics*.

Hornstein, A., P. Krusell, and G. L. Violante (2005). The Effects of Technical Change on Labor Market Inequalities. In P. Aghion and S. Durlauf (Eds.), *Handbook of Economic Growth*, Volume 1 of *Handbook of Economic Growth*, Chapter 20, pp. 1275–1370. Elsevier.

Hornstein, A., P. Krusell, and G. L. Violante (2007). Technology—Policy Interaction in Frictional Labour-Markets. *Review of Economic Studies 74*(4), 1089–1124.

Huckfeldt, C. (2021). Understanding the scarring effect of recessions. *American Economic Review forthcoming.*

Humlum, A. (2019). Robot adoption and labor market dynamics. *Princeton University.*

Jaimovich, N. and H. Siu (2018). The trend is the cycle: Job polarization and jobless recoveries. *Review of Economics and Statistics, forthcoming.*

Jerome, H. (1934). *Mechanization in Industry.* NBER.

Jones, C. I. and J. Kim (2018). A schumpeterian model of top income inequality. *Journal of Political Economy 126*(5), 1785–1826.

Jovanovic, B. and Y. Nyarko (1996). Learning by doing and the choice of technology. *Econometrica 64*(6), 1299–1310.

Kambourov, G. and I. Manovskii (2009). Occupational specificity of human capital. *International Economic Review 50*(1), 63–115.

Karabarbounis, L. and B. Neiman (2013). The Global Decline of the Labor Share*. *The Quarterly Journal of Economics 129*(1), 61–103.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2020). Measuring technological innovation over the long run. *American Economic Review: Insights forthcoming.*

Kerr, W. and S. Fu (2008). The survey of industrial r&d—patent database link project. *The Journal of Technology Transfer 33*, 173–186.

Keynes, J. M. (1930). *Economic Possibilities For Our Grandchildren*, Volume 9 of *The Collected Writings of John Maynard Keynes*, pp. 321–332. Royal Economic Society.

Kogan, L., D. Papanikolaou, L. D. W. Schmidt, and J. Song (2020). Technological innovation and labor income risk. Working Paper 26964, National Bureau of Economic Research.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017). Technological Innovation, Resource Allocation, and Growth. *The Quarterly Journal of Economics 132*(2), 665–712.

Kopytov, A., N. Roussanov, and M. Taschereau-Dumouchel (2018). Short-run pain, long-run gain? recessions and technological transformation. *Journal of Monetary Economics 97*, 29–44.

Krusell, P., L. E. Ohanian, J.-V. Ríos-Rull, and G. L. Violante (2000). Capital-skill complementarity and inequality: A macroeconomic analysis. *Econometrica 68*(5), 1029–1053.

Mann, K. and L. Püttmann (2018). Benign effects of automation: New evidence from patent texts. *Working Paper*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. *CoRR abs/1310.4546*.

Neal, D. (1995). Industry-specific human capital: Evidence from displaced workers. *Journal of labor Economics 13*(4), 653–677.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *EMNLP*.

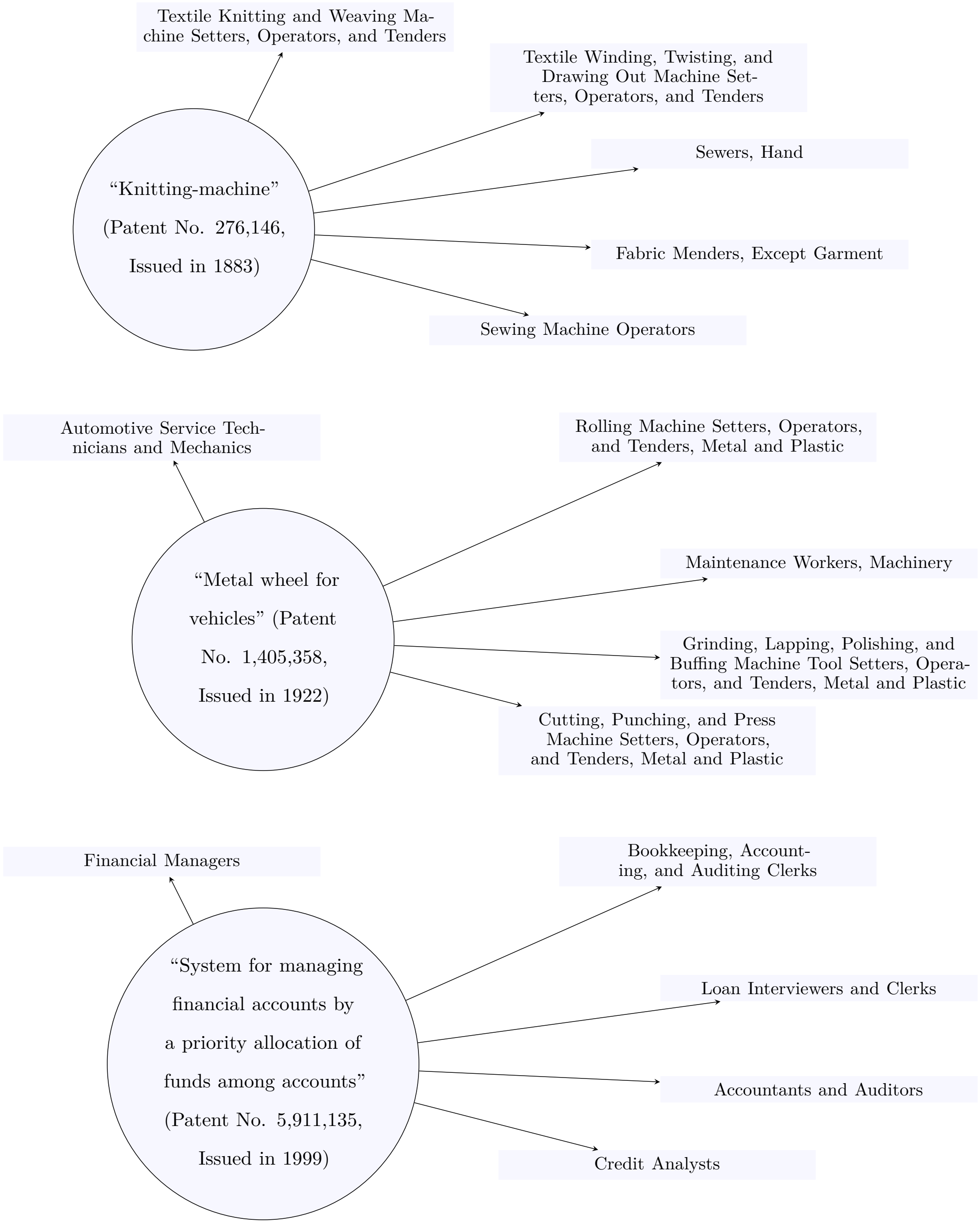
Violante, G. L. (2002). Technological acceleration, skill transferability, and the rise in residual inequality. *The Quarterly Journal of Economics 117*(1), 297.

Webb, M. (2019). The impact of artificial intelligence on the labor market.

Zhang, M. B. (2019). Labor-technology substitution: Implications for asset pricing. *The Journal of Finance 74*(4), 1793–1839.

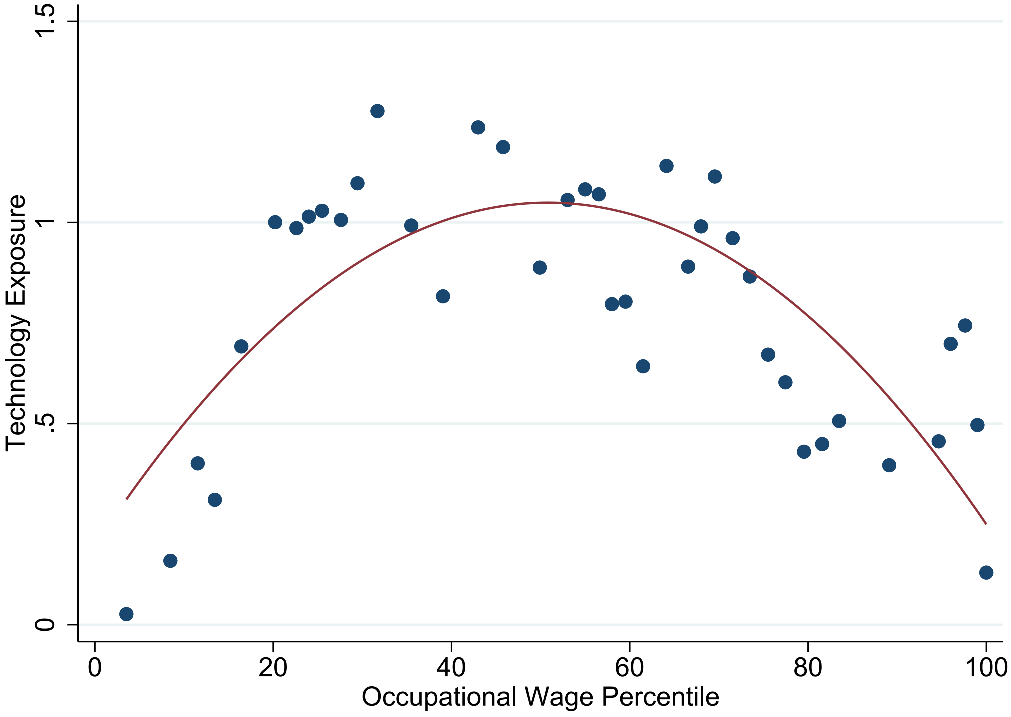
Figures and Tables

Figure 2-1: Examples of Technology Exposure: By Innovations



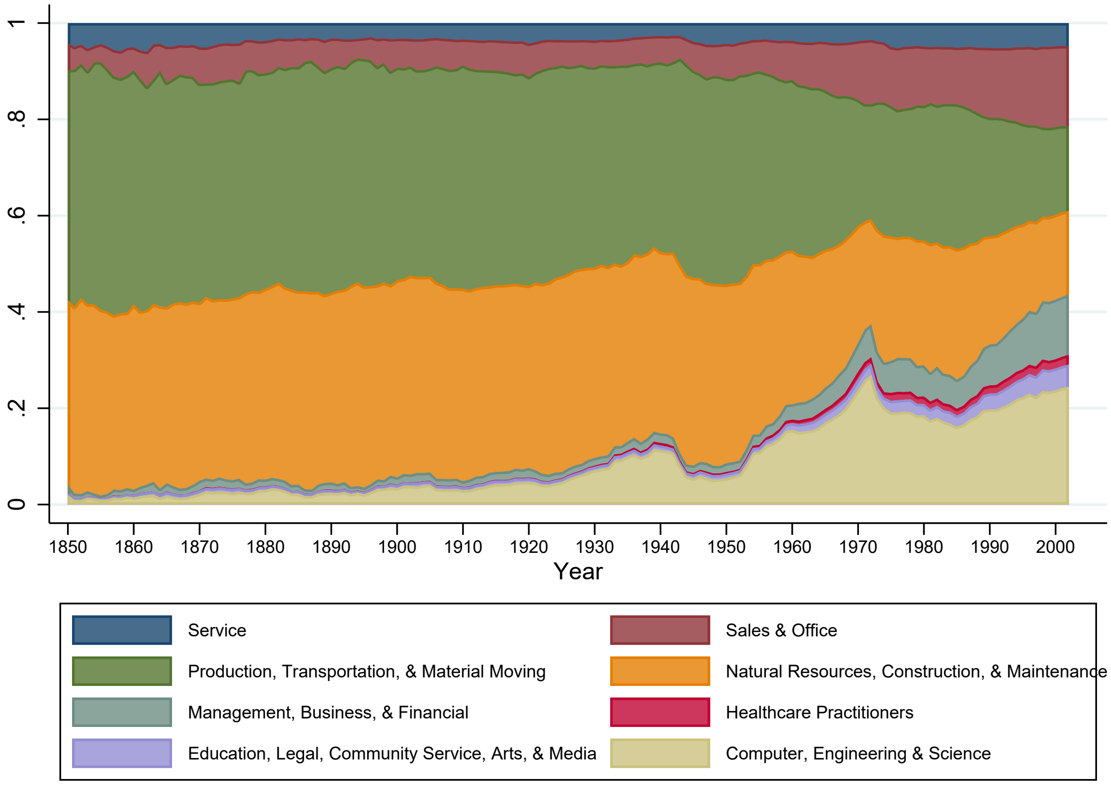
Textile Knitting and Weaving Machine Setters, Operators, and Tenders
Textile Winding, Twisting, and Drawing Out Machine Setters, Operators, and Tenders
Sewers, Hand
“Knitting-machine” (Patent No. 276,146, Issued in 1883)
Fabric Menders, Except Garment
Sewing Machine Operators
Automotive Service Technicians and Mechanics
Rolling Machine Setters, Operators, and Tenders, Metal and Plastic
Maintenance Workers, Machinery
“Metal wheel for vehicles” (Patent No. 1,405,358, Issued in 1922)
Grinding, Lapping, Polishing, and Buffing Machine Tool Setters, Operators, and Tenders, Metal and Plastic
Cutting, Punching, and Press Machine Setters, Operators, and Tenders, Metal and Plastic
Financial Managers
Bookkeeping, Accounting, and Auditing Clerks
“System for managing financial accounts by a priority allocation of funds among accounts” (Patent No. 5,911,135, Issued in 1999)
Loan Interviewers and Clerks
Accountants and Auditors
Credit Analysts

Figure 2-2: Technological Exposure, by occupation income level



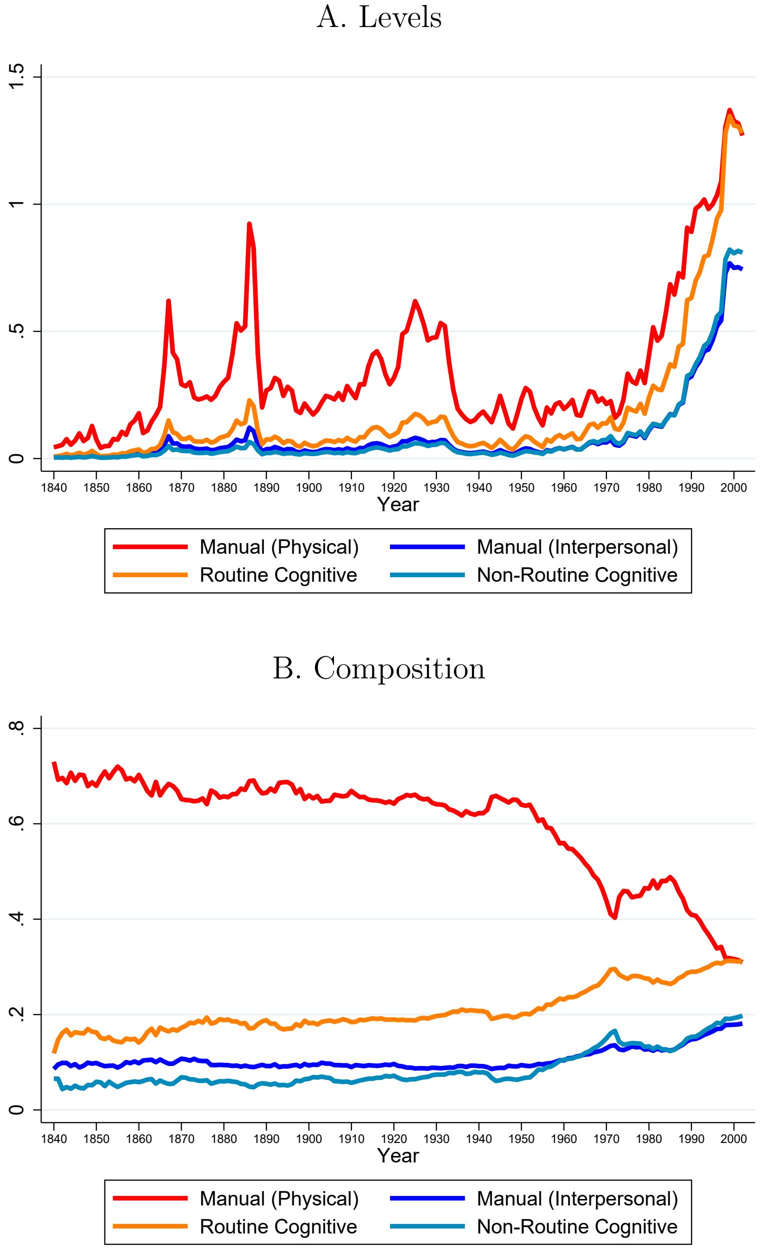
Note: This figure plots average $\eta_{i,t}$ for occupations over the 1980 to 2002 period by wage percentile rank. The wage data come from the Current Population Survey Merged Outgoing Rotation Groups.

Figure 2-3: Technological Exposure, composition by major occupation group



Note: This figure plots the average of our occupation-level innovation exposure index, $\eta_{i,t}$, where $\eta_{i,t}$ has been averaged separately within eight broad occupation groups. The occupation group averages are re-scaled each year so that the total across all groups sums to one in the given year.

Figure 2-4: Technology Exposure, by task type


A. Levels

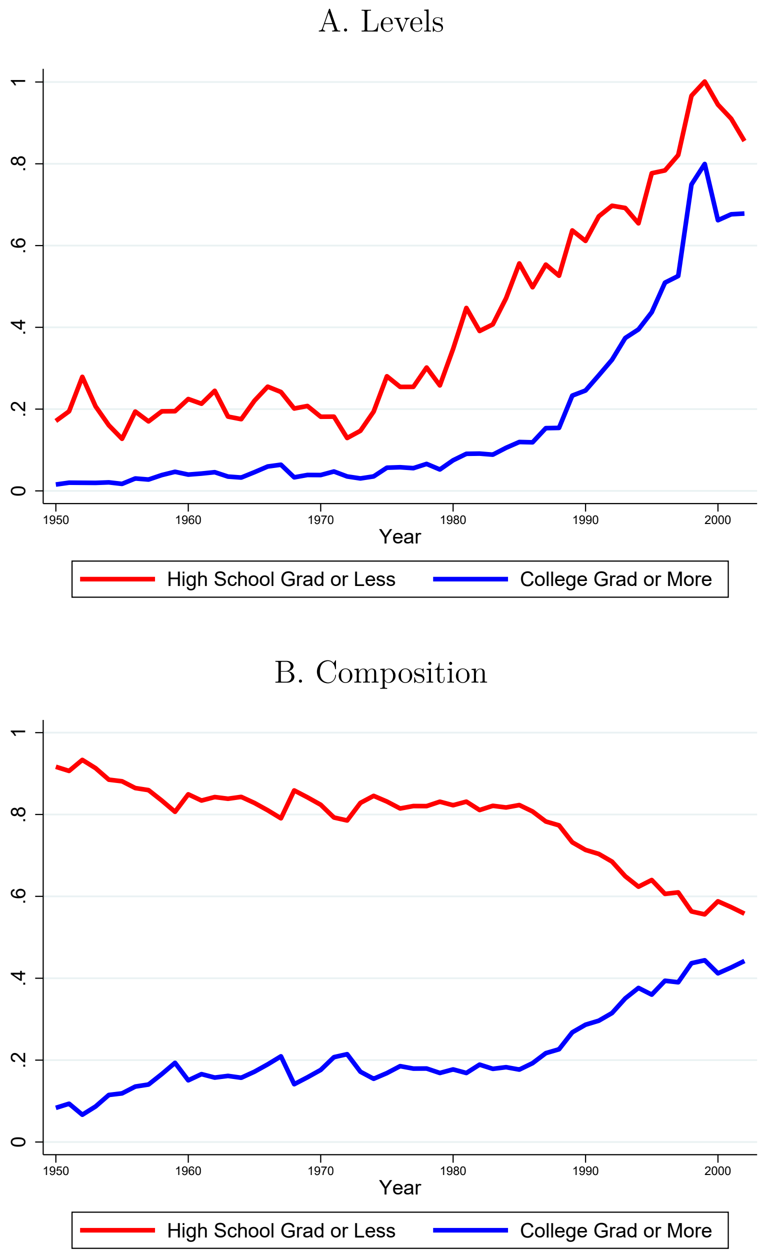
B. Composition

Note: This figure plots the level and composition of our index of technological exposure by task category:

$$\lambda_{w,t} = \sum_i \eta_{i,t} \times T_{w,t}(i) \times \omega_i \tag{2.35}$$

Panel A plots the raw index $\lambda_{w,t}$ and panel B plots the relative shares $\lambda_{w,t}/\sum_{w'} \lambda_{w',t}$ Here w represents one of the four given task categories. $T_{w,t}$ is an indicator that takes a value of 1 if occupation i is in the top quintile of the cross-sectional distribution of task scores for task category w. $\eta_{i,t}$ is our index of technological exposure and ω_i gives the Acemoglu and Autor (2011) occupational employment shares. See main text for more details.

Figure 2-5: Technology Exposure, by education requirements


A. Levels

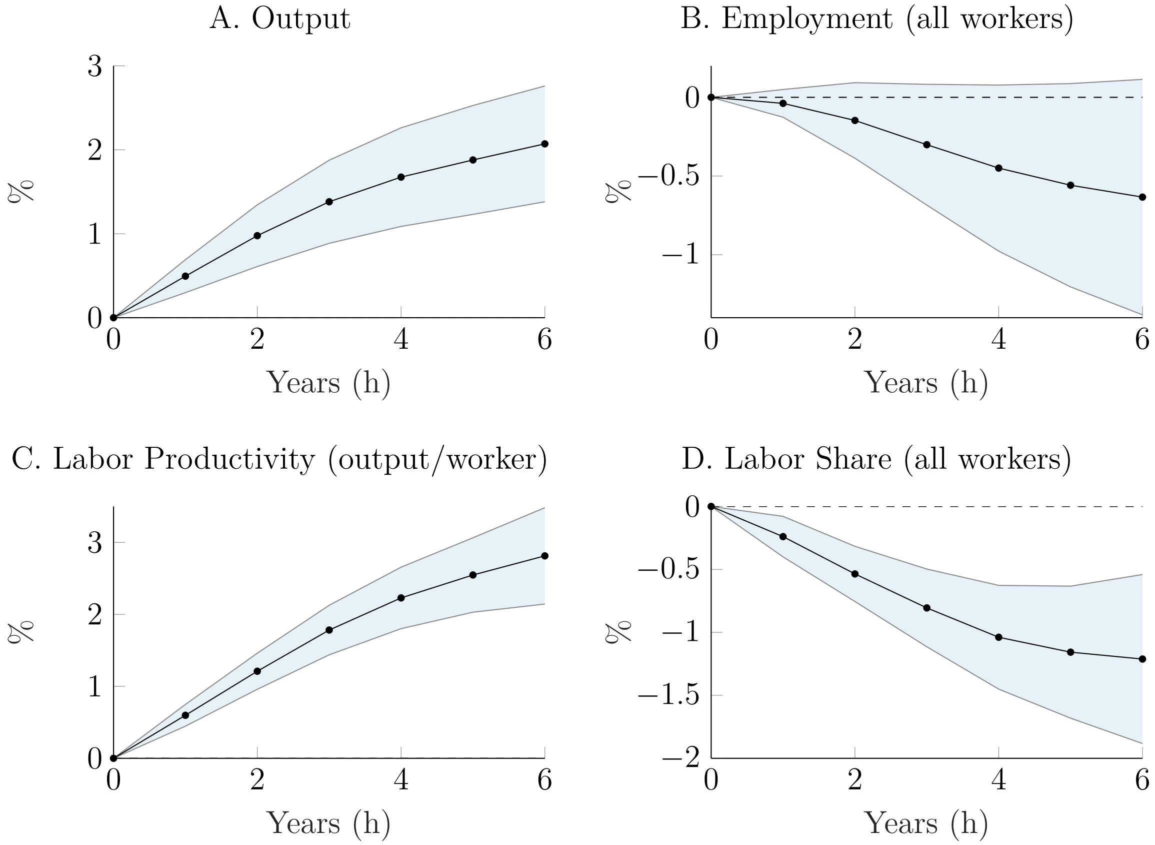
B. Composition

Note: This figure plots the level and composition of our index of technological exposure by education category:

$$\zeta_{s,t} = \sum_i \eta_{i,t} \times S_{s,t}(i) \times \omega_{i,t} \tag{2.36}$$

Panel A plots the raw index $\zeta_{s,t}$ and panel B plots the relative shares $\zeta_{s,t}/\sum_{s'} \zeta_{s',t}$ Here s represents either the educational category "high school or less" or "college grad or more". $S_{s,t}$ is and indicator that takes a value of 1 if occupation i is in the top quintile of the time t cross-sectional distribution of shares of workers falling in category s. $\eta_{i,t}$ is our index of technological exposure and $\omega_{i,t}$ gives occupational employment shares. See main text for more details.

Figure 2-6: Innovation: Productivity vs Labor Share

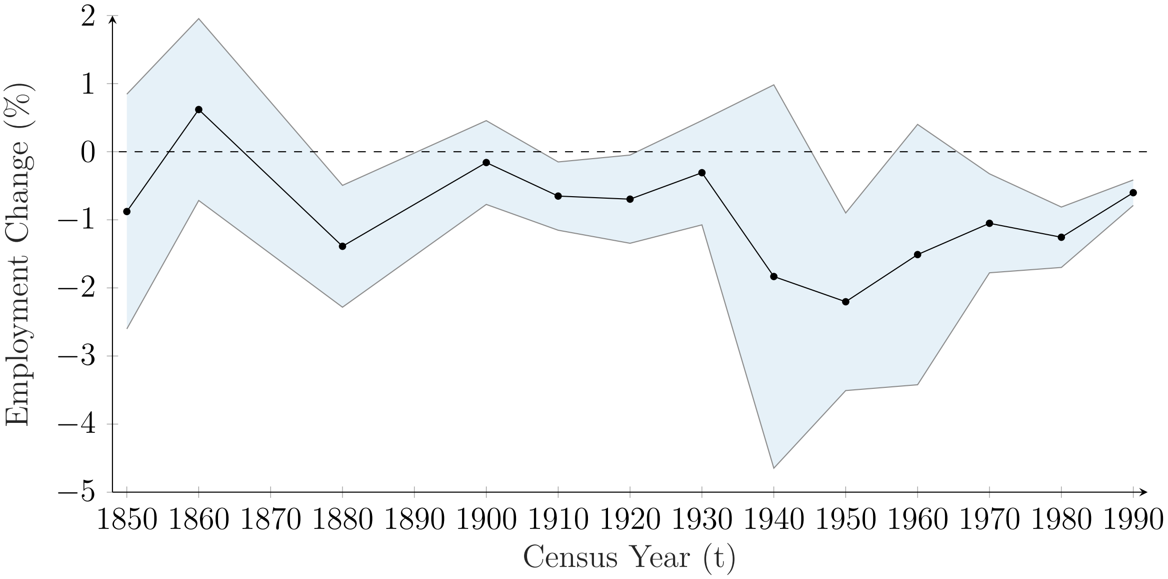


Note: The figure plots the estimated coefficients $\beta(k)$ from regressions of the form

$$\log X_{j,t+k} - \log X_{j,t} = \alpha(k) + \beta(k)\,\psi_{j,t} + \delta(k)Z_{j,t} + \epsilon_{j,t} \qquad \text{for } k = 1\ldots T \text{ years}$$

The main independent variable $\psi_{j,t}$ is an index of innovation in industry j in year t, constructed as follows. First, we assign breakthrough patents to industries using the patent CPC tech class to industry crosswalk from Goldschlag et al. (2020). Second, we only include breakthrough patents whose average similarity to the industry's occupations (using occupation-by-industry employment weights) are above the (unconditional) median. We scale $\psi_{j,t}$ by US population and normalize to unit standard deviation. Controls $Z_{j,t}$ include industry employment shares, year fixed effects and lagged 5-year growth rate of the dependent variable. Standard errors are clustered by industry, and corresponding t-stats are shown in parentheses.

Figure 2-7: Employment and Technology Exposure (long-run: 1850–present)

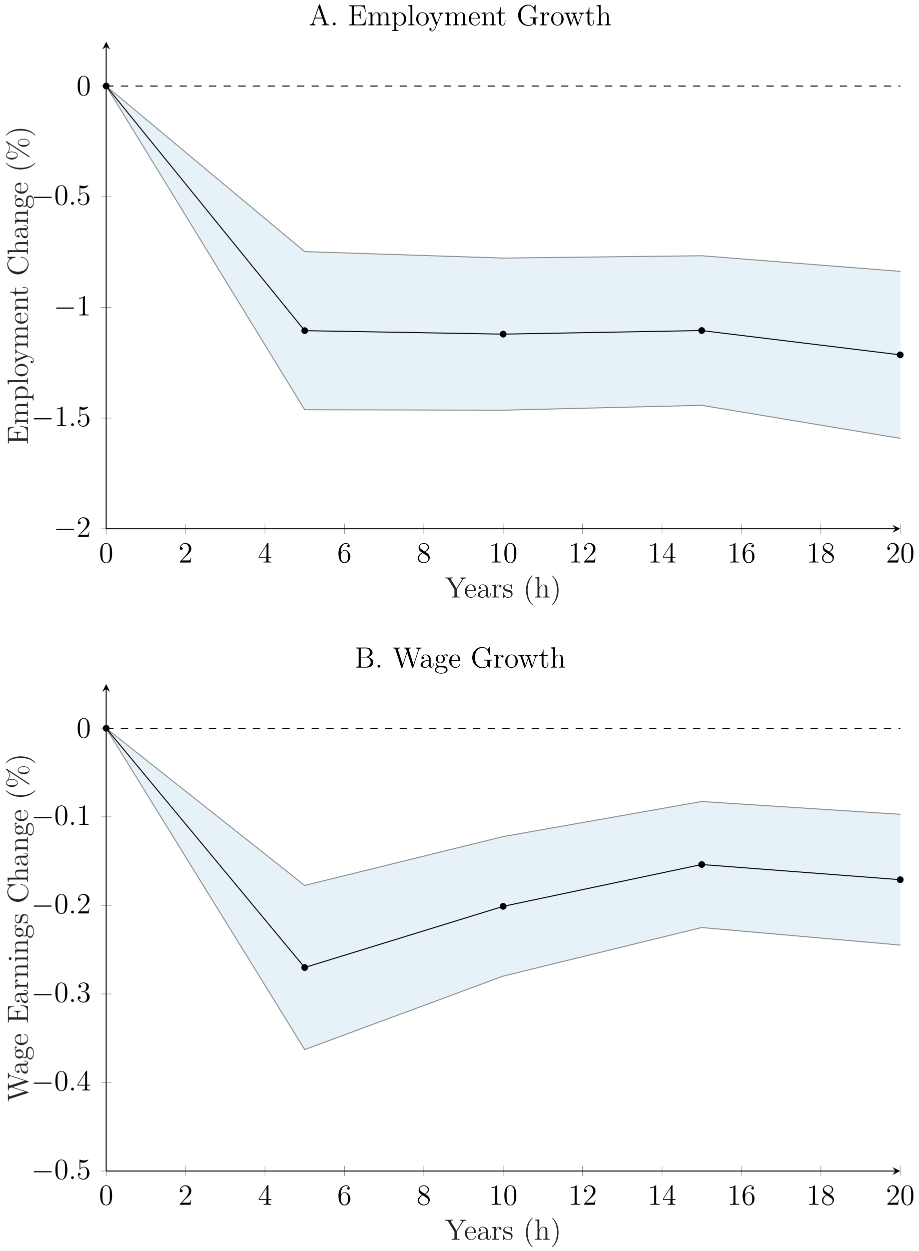


Note: Figure shows the slope coefficients on annual regressions of 20-Year employment share growth on our technology exposure $\eta_{i,t}$, using Census Years from 1850 to 1990. Specifically, we plot the β coefficients from

$$\frac{1}{k}\Big(\log Y_{j,t+k} - \log Y_{j,t}\Big) = \alpha_t + \beta_t \eta_{i,t} + \lambda_t + \epsilon_{i,t}$$

Here $Y_{i,t}$ is the occupation's share in total non-farm employment. Standard errors are clustered by occupation and shaded area represents the corresponding 90% confidence intervals for β_τ. Growth rates are expressed in annualized percentage terms and $\eta_{i,t}$ is standardized.

Figure 2-8: Employment, wage earnings and technology exposure (recent period: 1980–present)



A. Employment Growth

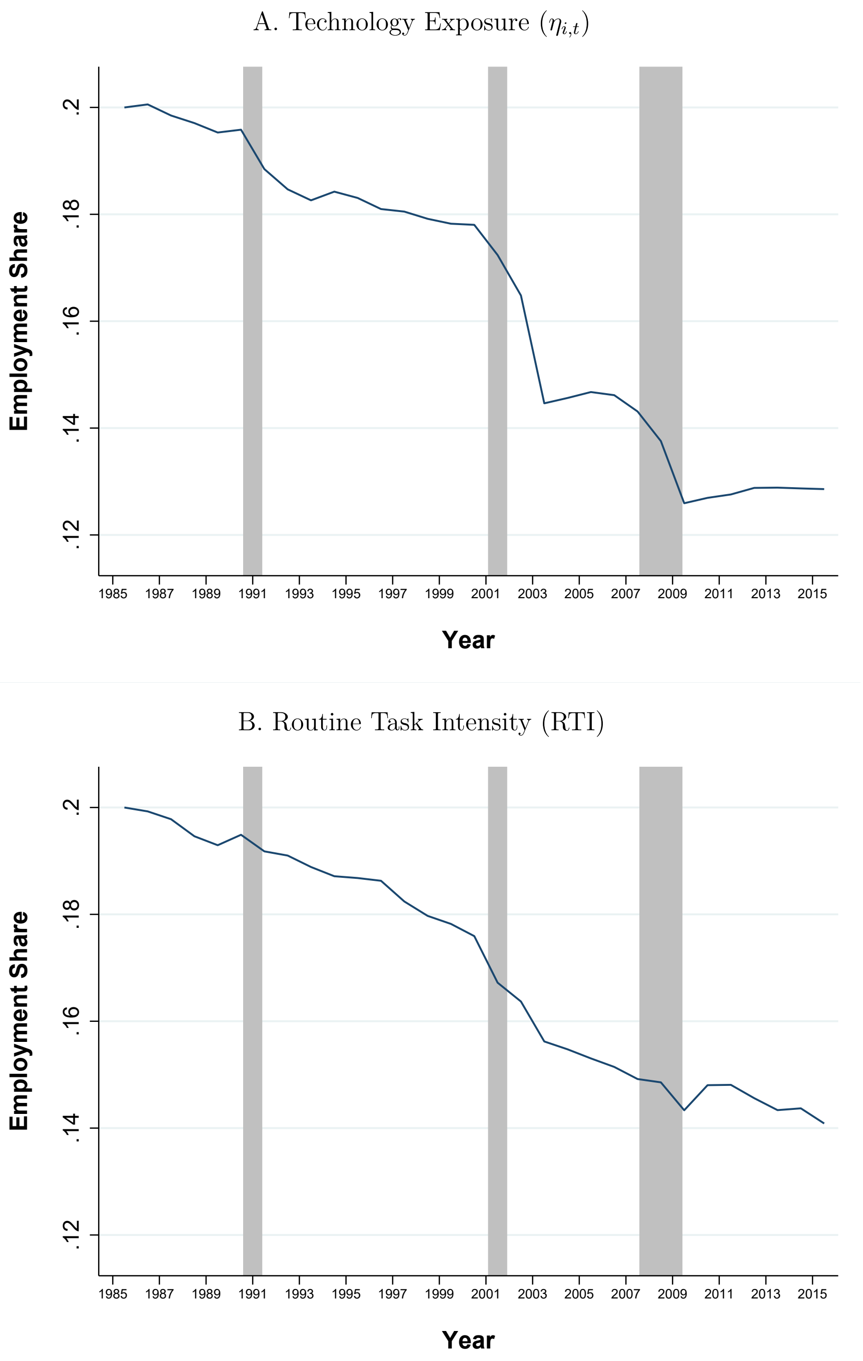
B. Wage Growth

Note: The Figures above plot coefficients from panel regressions of annualized wage and income growth rates over different time horizons on occupation innovation exposures:

$$y_{i,t+k} - y_{i,t} = \alpha + \beta \eta_{i,t} + \delta X_{i,t} + \varepsilon_{i,t}$$

Controls $X_{i,t}$–includes three one-year lags of dependent variable, and time fixed effects. Dependent variable is expressed in annualized percentage terms and $\eta_{i,t}$ is standardized. Figures plot 90% confidence interval for each time horizon. Data come from the CPS Merged Outgoing Rotation Groups (MORG) and cover the 1985–2018 period.

Figure 2-9: Employment share over the business cycle (Technology Exposure vs RTI)



A. Technology Exposure ($\eta_{i,t}$)

B. Routine Task Intensity (RTI)

Note: The above figure plots aggregate employment shares over time for occupations that were in the top quintiles of innovation exposure ($\eta_{i,t}$) and routine-task intensity in the year 1985. Vertical shaded bars represent NBER recession dates. Data source: CPS Merged Outgoing Rotation Groups extracts obtained from the Center For Economic Policy Research website.

Figure 2-10: Example: Order Clerks versus Personnel and Library Clerks

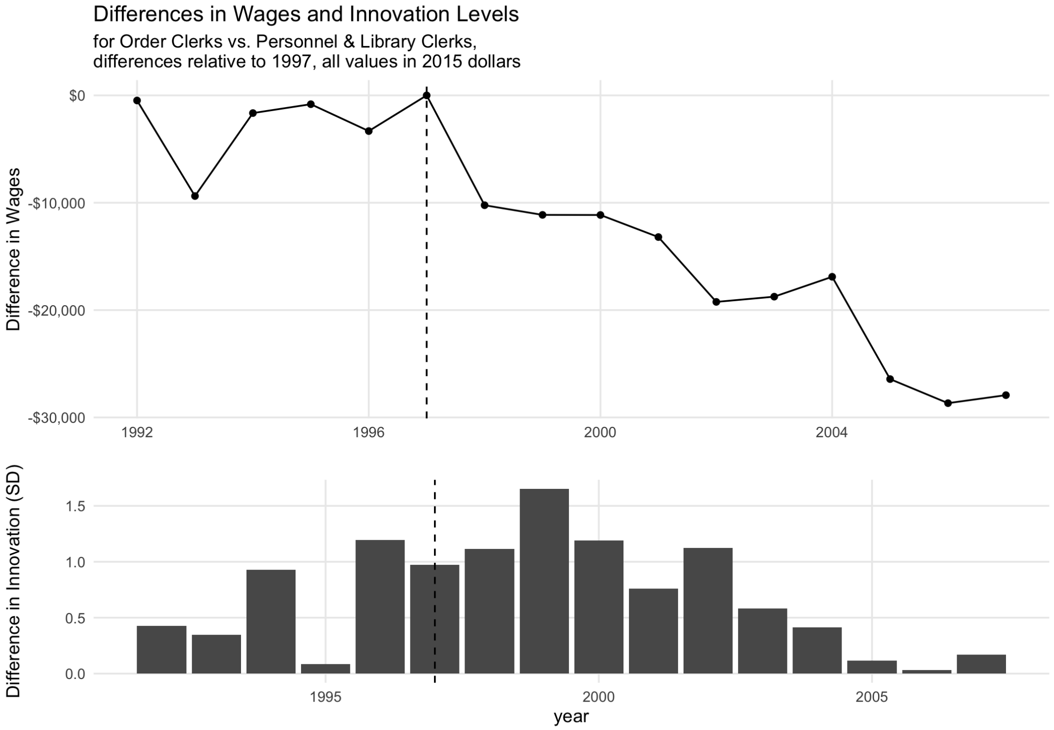
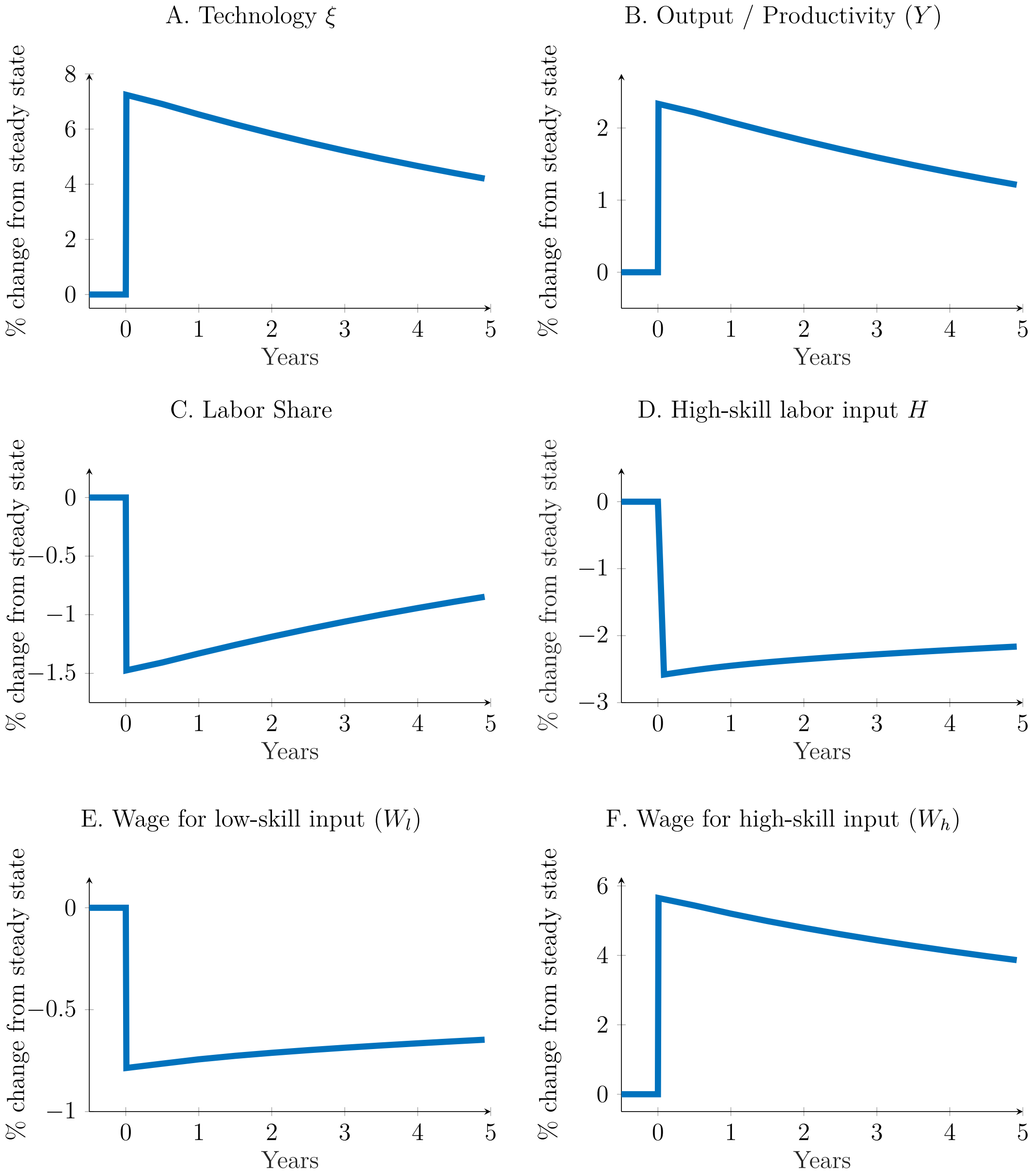
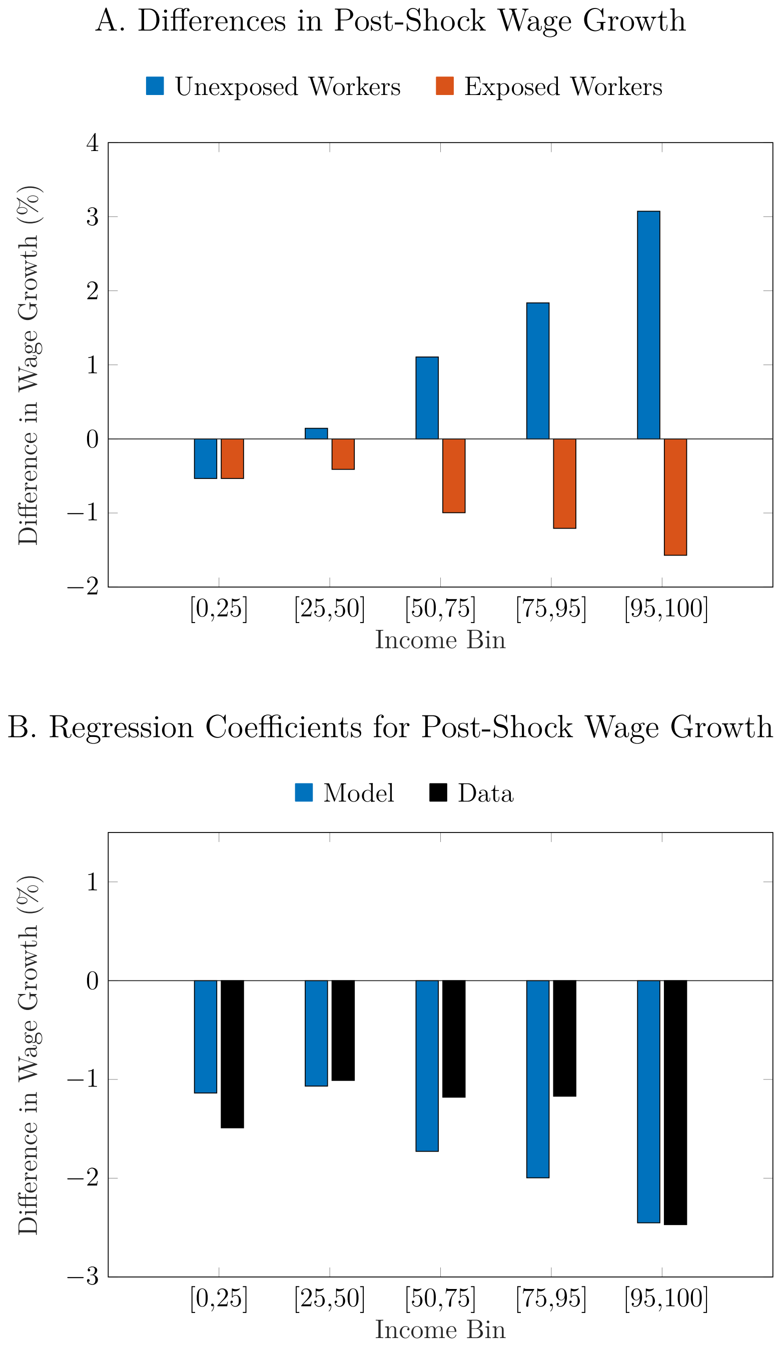


Figure 2-11: Model: Impulse Responses



Note: This figure shows the impulse responses of key model quantities following a one-standard deviation technology shock evaluated at the steady state of the model.

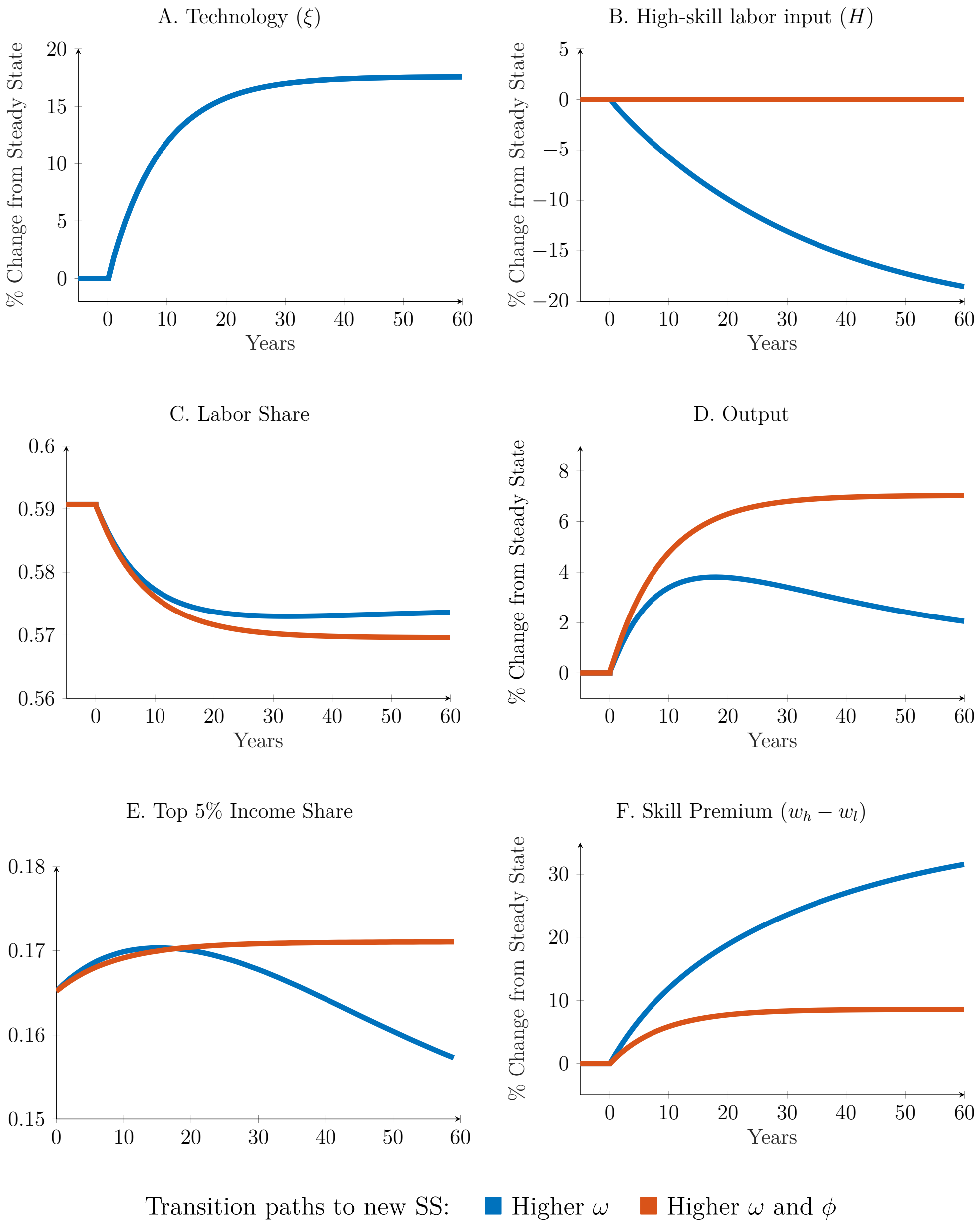
Figure 2-12: Model: Innovation and Worker Earnings


A. Differences in Post-Shock Wage Growth

B. Regression Coefficients for Post-Shock Wage Growth

Note: Panel A shows raw differences between wage growth during a shock period and wage growth if there had not been a shock for workers who are exposed to the shock, workers who are not exposed to the shock. Panel B shows the associated regression coefficients, which represent the marginal effects of a shock on wage growth given exposure. The left part of the figure shows the results in our baseline calibration. The right part of the figure compares to the case where there is no displacement of human capital.

Figure 2-13: The race between education and technology



Note: Figure computes the transition paths from the old to the new steady state for two permanent parameter shifts: 1) the blue line plots a permanent increase in the frequency of technological innovation ω, calibrated so that the level of technology ξ is permanently higher by one standard deviation relative to the old steady state (panel A); and 2) the orange line plots the transition paths associated with the same shift in ω but also with an increase in the rate of new skill acquisition ϕ such that the total supply of the high-skill labor input remains the same as the old steady state (panel B). Panel C plots the labor share of output; panel D plots total output/productivity; panel E plots income inequality, defined as the top 5% income share in the model; and panel F plots the skill premium.

Table 2.1: Technology And Employment Over the Long Run (1850–present)

	A. Occupation-level Employment				B. Industry X Occupation level employment			
	10 Years	20 Years	10 Years	20 Years	10 Years	20 Years	10 Years	20 Years
Technology Exposure, $\eta_{i,t}$	-0.48***	-0.77***	-0.37***	-0.67***	-0.56***	-0.97***	-0.60***	-1.10***
(past decade average)	(-5.19)	(-6.84)	(-4.39)	(-6.93)	(-3.02)	(-3.92)	(-3.18)	(-4.33)
Observations	2,865	2,574	2,492	2,208	102,400	81,009	72,451	54,662
Controls								
Time FE	Y	Y	Y	Y				
Industry X Time FE					Y	Y	Y	Y
Lagged Dependent Variable			Y	Y			Y	Y

Note: The table above reports results from regressions of the form

$$\frac{1}{k}\Big(\log Y_{i,t+k} - \log Y_{i,t}\Big) = \alpha_0 + \alpha_t + \beta(k)\eta_{i,t} + \rho\left(\log Y_{i,t} - \log Y_{i,t-k}\right) + \epsilon_{i,t}$$

for $k = 10, 20$ years for Census years spanning from 1850-2010. Here $Y_{i,t}$ is the occupation's share in total non-farm employment. $\eta_{i,t}$ is standardized and growth rates are in annualized percentage terms. Standard errors are clustered by occupation and corresponding t-stats are shown in parentheses. Observations are weighted by occupation employment share at time t. Census year 1870 does not show up in the first column of the 20-year subsample regressions because the 1890 Census records no longer exist.

Table 2.2: Summary Statistics: Worker-Level Data

Variable	Mean	SD	5%	25%	Median	75%	95%
W2-Earnings	100,100	282,400	18,060	47,100	76,010	119,600	233,900
Age	41	7	29	34	41	47	53
Age: Lowest 25% Earners	39	7	29	33	38	44	51
Age: Top 5% Earners	44	7	31	39	46	50	54
Earnings growth, 3-years	-0.053	0.573	-0.890	-0.124	0.015	0.143	0.540
Earnings growth, 5-years	-0.063	0.592	-0.968	-0.158	0.011	0.156	0.574
Earnings growth, 10 -years	-0.073	0.639	-1.098	-0.215	0.005	0.190	0.666

Note: The table reports summary statistics for our wage earnings data from the Census-CPS sample, which covers the 1988 to 2016 period. W2-Earnings are reported in terms of 2015 dollars.

Table 2.3: Worker Earnings and Technology Exposure

	(1)	(2)
A. Cond. Mean: E[g], by Horizon		
3 years	-1.21	-1.12
	(-6.32)	(-5.65)
5 years	-1.55	-1.30
	(-6.84)	(-4.77)
10 years	-1.43	-1.24
	(-4.96)	(-3.14)
B. Risk: Absolute Income Growth E[|g|]		
3 years	0.73	0.25
	(3.03)	(1.59)
5 years	0.76	0.37
	(2.40)	(1.78)
10 years	0.38	0.47
	(0.95)	(1.68)
C. Skewness: Prob. Large Income Decline $p(g < p^{10})$		
3 years	0.56	0.38
	(6.31)	(4.18)
5 years	0.58	0.41
	(4.76)	(3.31)
10 years	0.37	0.27
	(2.09)	(1.48)
Controls:		
Industry FE	Y	
Occupation FE	Y	
Year FE	Y	
Industry × Year FE		Y
Occupation × Year FE		Y

Note: Panel A shows the estimated slope coefficients $\beta(h)$ (times 100) from equation (2.21) in the main text for horizons h of 3,5, and 10 years. Panels B and C focus on the 5-year horizon. Panel B shows the slope coefficients of a variant of the above specification where we replace the dependent variable $g_{i,t:t+h}$ with its absolute value $|g_{i,t:t+h}|$ to capture the response of second moments to changes in technology exposure $\eta_{i,t}$. Similarly, Panel C replaces the dependent variable with a dummy that takes the value of one if $g_{i,t:t+h}$ lies in the bottom 10-th percentile; this specification allows us to capture increases in negative skewness in response to an increase in $\eta_{i,t}$. We report t-statistics (in parentheses) using standard errors clustered at the industry (NAICS 4-digit) level. All specifications include industry times year and occupation times year fixed effects. We normalize $\eta_{i,t}$ to unit standard deviation. The bottom panel shows the p-values associated with the hypotheses that the coefficients are equal across the reported subgroups.

Table 2.4: Worker Earnings and Technology Exposure, by Education

	(1)	(2)	(3)	(4)	(5)
Education		Cond. Mean		St. Dev	Skew
		$E[g]$		$E[\|g\|]$	$p(g < p^{10})$
	3-year	5-year	10-year	5-year	5-year
College	-0.91	-1.14	-1.21	0.49	0.35
	(-3.33)	(-3.34)	(-2.49)	(2.65)	(2.57)
No College	-1.30	-1.50	-1.76	0.39	0.48
	(-5.58)	(-5.18)	(-4.47)	(1.51)	(4.33)
Coeff. Differences			p-values		
College = No College	0.043	0.110	0.052	0.537	0.071

Note: Columns (1) to (3) show the estimated slope coefficients (times 100) from equation (2.21) in the main text: the dependent variable is worker earnings growth over horizons of 3,5, and 10 years; the main independent variable of interest is a worker's technology exposure $\eta_{i,t}$. The slope coefficient $\beta(h)$ is allowed to vary with the worker's education. Columns (1) to (3) correspond to horizons of 3,5, and 10 years. Columns (4) and (5) focus on the 5-year horizon. Column (4) shows the slope coefficients of a variant of the above specification where we replace the dependent variable $g_{i,t:t+h}$ with its absolute value $|g_{i,t:t+h}|$ to capture the response of second moments to changes in technology exposure $\eta_{i,t}$. Similarly, Column (5) replaces the dependent variable with a dummy that takes the value of one if $g_{i,t:t+h}$ lies in the bottom 10-th percentile; this specification allows us to capture increases in negative skewness in response to an increase in $\eta_{i,t}$. We report t-statistics (in parentheses) using standard errors clustered at the industry (NAICS 4-digit) level. All specifications include industry times year and occupation times year fixed effects.We normalize $\eta_{i,t}$ to unit standard deviation. The bottom panel shows the p-values associated with the hypotheses that the coefficients are equal across the reported subgroups.

Table 2.5: Worker Earnings and Technology Exposure, by Prior Income

	(1)	(2)	(3)	(4)	(5)
Income Percentile	Cond. Mean			St. Dev	Skew
	$E[g]$			$E[\|g\|]$	$p(g < p^{10})$
	3-year	5-year	10-year	5-year	5-year
0 to 25-th	-1.24	-1.49	-1.85	-0.26	0.29
	(-5.76)	(-5.28)	(-4.70)	(-1.07)	(2.23)
25 to 50-th	-0.85	-1.01	-0.96	0.10	0.26
	(-4.14)	(-3.57)	(-2.04)	(0.40)	(1.45)
50 to 75-th	-1.01	-1.18	-1.01	0.48	0.39
	(-3.51)	(-3.07)	(-1.87)	(1.65)	(2.84)
75 to 95-th	-1.05	-1.17	-0.81	0.76	0.39
	(-4.12)	(-3.45)	(-1.64)	(3.25)	(2.76)
95 to 100-th	-2.24	-2.47	-2.28	2.01	1.26
	(-6.11)	(-4.75)	(-3.83)	(5.78)	(5.50)
Coeff. Differences			p-values		
[95-100] = [25-95]	0.000	0.000	0.002	0.000	0.000
[0-25] = [25-95]	0.610	0.630	0.331	0.175	0.853
[95-100] = [0-25]	0.019	0.092	0.498	0.000	0.000

Note: Columns (1) to (3) show the estimated slope coefficients (times 100) from equation (2.21) in the main text: the dependent variable is worker earnings growth over horizons of 3,5, and 10 years; the main independent variable of interest is a worker's technology exposure $\eta_{i,t}$. The slope coefficient $\beta(h)$ is allowed to vary with the worker's prior income rank. Columns (1) to (3) correspond to horizons of 3,5, and 10 years. Columns (4) and (5) focus on the 5-year horizon. Column (4) shows the slope coefficients of a variant of the above specification where we replace the dependent variable $g_{i,t:t+h}$ with its absolute value $|g_{i,t:t+h}|$ to capture the response of second moments to changes in technology exposure $\eta_{i,t}$. Similarly, Column (5) replaces the dependent variable with a dummy that takes the value of one if $g_{i,t:t+h}$ lies in the bottom 10-th percentile; this specification allows us to capture increases in negative skewness in response to an increase in $\eta_{i,t}$. We report t-statistics (in parentheses) using standard errors clustered at the industry (NAICS 4-digit) level. All specifications include industry times year and occupation times year fixed effects. We normalize $\eta_{i,t}$ to unit standard deviation. The bottom panel shows the p-values associated with the hypotheses that the coefficients are equal across the reported subgroups.

Table 2.6: Worker Earnings and Technology Exposure, by Age

	(1)	(2)	(3)	(4)	(5)
Worker Age	Cond. Mean			St. Dev	Skew
	$E[g]$			$E[\|g\|]$	$p(g < p^{10})$
	3-year	5-year	10-year	5-year	5-year
25–35 years	-0.39	-0.64	-0.92	0.38	0.18
	(-1.86)	(-2.55)	(-2.42)	(1.84)	(1.50)
35–45 years	-0.86	-1.04	-1.37	0.05	0.23
	(-5.24)	(-5.08)	(-3.63)	(0.37)	(2.55)
45–55 years	-1.95	-2.25	-2.51	1.13	0.88
	(-3.72)	(-3.55)	(-3.02)	(2.5)	(3.36)
Coeff. Differences			p-values		
45–55 = 25–35	0.001	0.002	0.006	0.051	0.001
45–55 = 35–45	0.015	0.020	0.044	0.010	0.011

Note: Columns (1) to (3) show the estimated slope coefficients (times 100) from equation (2.21) in the main text: the dependent variable is worker earnings growth over horizons of 3,5, and 10 years; the main independent variable of interest is a worker's technology exposure $\eta_{i,t}$. The slope coefficient $\beta(h)$ is allowed to vary with the worker's age. Columns (1) to (3) correspond to horizons of 3,5, and 10 years. Columns (4) and (5) focus on the 5-year horizon. Column (4) shows the slope coefficients of a variant of the above specification where we replace the dependent variable $g_{i,t:t+h}$ with its absolute value $|g_{i,t:t+h}|$ to capture the response of second moments to changes in technology exposure $\eta_{i,t}$. Similarly, Column (5) replaces the dependent variable with a dummy that takes the value of one if $g_{i,t:t+h}$ lies in the bottom 10-th percentile; this specification allows us to capture increases in negative skewness in response to an increase in $\eta_{i,t}$. We report t-statistics (in parentheses) using standard errors clustered at the industry (NAICS 4-digit) level. All specifications include industry times year and occupation times year fixed effects.We normalize $\eta_{i,t}$ to unit standard deviation. The bottom panel shows the p-values associated with the hypotheses that the coefficients are equal across the reported subgroups.

Table 2.7: Model Parameters

Description	Parameter	Value
Share of workers who do not move up the ladder	s_l	0.375
Minimum level of skill	$\underline{\theta}$	0.03
Probability of worker exit	δ	0.025
Amount of skills acquired	m	0.03
CES parameter in inner nest (technology ξ and low-skill labor L)	ρ	0.74
Share of technology in inner nest	λ	0.27
CES parameter in outer nest (high-skill labor H and ξ/L composite)	σ	-0.12
Share of high-skill labor in outer nest	μ	0.17
Size of technology improvement	κ	0.31
Arrival rate of technology shocks	ω	1.56
Share of exposed workers	α	0.32
Human capital loss percentage conditional on fall	h	0.06
Rate of depreciation of technology	g	0.12
Likelihood of worker skill acquisition	ϕ	2.40

Note: Table reports the parameter used to calibrate the model. The first four parameters are calibrated a priori; the latter 10 parameters are chosen to fit the statistics reported in Table 2.8.

Table 2.8: Model Fit

Statistic	Data	Model
Labor share, average	0.66	0.59
Labor share, response to ξ	-1.29	-1.48
Skill premium (p75 / p25 ratio), average	2.45	1.68
Labor productivity, response to ξ	2.81	2.31
Worker earnings growth response to ξ		
0 to 25-th percentile	-1.49	-1.13
25 to 50-th percentile	-1.01	-1.06
50 to 75-th percentile	-1.18	-1.72
75 to 95-th percentile	-1.17	-2.00
95 to 100-th percentile	-2.47	-2.45
Likelihood of large wage declines in response to ξ		
0 to 25-th percentile	0.29	0.51
25 to 50-th percentile	0.26	0.51
50 to 75-th percentile	0.39	0.51
75 to 95-th percentile	0.39	0.51
95 to 100-th percentile	1.26	1.40

Note: Table reports the fit of the model to the statistics that we target. The parameters used in our calibration are listed in Table 2.8.

Table 2.9: Technology and Labor Market Outcomes: Comparison to Other Measures

	A. Employment				
	(1)	(2)	(3)	(4)	(5)
Technology Exposure $\eta_{i,1980}$	-0.79***	-0.87***	-0.74***	-0.74***	-0.82***
	(-5.68)	(-6.25)	(-5.26)	(-5.12)	(-5.99)
Routine Task Intensity (RTI)		-0.058			0.011
		(-0.50)			(0.08)
Offshorability		-0.15**			-0.20**
		(-2.04)			(-2.50)
Robot Exposure			-0.66**		-0.93**
			(-2.21)		(-2.21)
Software Exposure				-0.36	0.069
				(-1.34)	(0.21)
	B. Wages				
	(1)	(2)	(3)	(4)	(5)
Technology Exposure $\eta_{i,j,1980}$	-0.082***	-0.064***	-0.071***	-0.100***	-0.078***
	(-4.11)	(-3.11)	(-3.87)	(-5.37)	(-4.21)
Routine Task Intensity (RTI)		-0.066*			-0.059*
		(-1.97)			(-1.70)
Offshorability		0.0042			-0.0026
		(0.17)			(-0.09)
Robot Exposure			-0.13*		-0.27***
			(-1.86)		(-2.76)
Software Exposure				0.13**	0.24***
				(2.10)	(4.73)
Observations	17536	16959	17448	17448	16959

Note: This table shows results from estimating

$$\frac{1}{k}\Big(\log Y_{i,j,t+k} - \log Y_{i,j,t}\Big) = \alpha + \alpha_j + \beta\eta_{i,1980} + \delta X_i + \epsilon_{i,j}$$

Here i indexes occupations and j indexes industries; we report results for $k = 32$ years using Deming (2017) data from the 1980 Census and the 2012 ACS. The dependent variables are employment (Panel A) or average wages (Panel B). All specifications include industry fixed effects and controls for occupation employment share in 1980, occupation log wage in 1980, three categorical indicators for the occupation's average education level in 1980. We additionally include the routine-task intensity and the measure of occupation-level offshorability from Autor and Dorn (2013) and the measure of exposure to robots or software from Webb (2019) depending on the specification. Observations are weighted by employment share in 1980.

2.6 Appendix

2.6.1 Converting Patent Text for Numerical Analysis

Here, we briefly overview our conversion of unstructured patent text data into a numerical format suitable for statistical analysis. We obtain text data for measuring patent/job task similarity from two sources. Job task descriptions come from the revised 4th edition of the Dictionary of Occupation Titles (DOT) database. We use the patent text data parsed from the USPTO patent search website in Kelly et al. (2020), which includes all US patents beginning in 1976, comprising patent numbers 3,930,271 through 9,113,586, as well as patent text data obtained from Google patents for pre-1976 patents. Our analysis of the patent text combines the claims, abstract, and description section into one patent-level corpus for each patent. Since the DOT has a very wide range of occupations (with over 13,000 specific occupation descriptions) we first crosswalk the DOT occupations to the considerably coarser and yet still detailed set of 6-digit occupations in the 2010 edition of O*NET. We then combine all tasks for a given occupation at the 2010 O*NET 6-digit level into one occupation-level corpus. The process for cleaning and preparing the text files for numerical representation follows the steps outlined below.

We first clean out all non-alphabetic characters from the patent and task text, including removing all punctuation and numerical characters. We then convert all text to lowercase. At this stage each patent and occupation-level task text are represented by a single string of words separated by spaces. To convert each patent/occupation into a list of associated words we apply a word tokenizer that separates the text into lists of word tokens which are identified by whitespace in between alphabetic characters. Since most words carry little semantic information, we filter the set of tokens by first removing all "stop words"– which include prepositions, pronouns, and other common words carrying little content–from the union of several frequently used stop words lists.

Stop words come from the following sources:

- `https://pypi.python.org/pypi/stop-words`

- `https://dev.mysql.com/doc/refman/5.1/en/fulltext-stopwords.html`
- `http://www.lextek.com/manuals/onix/stopwords1.html`
- `http://www.lextek.com/manuals/onix/stopwords2.html`
- `https://msdn.microsoft.com/zh-cn/library/bb164590`
- `http://www.ranks.nl/stopwords`
- `http://www.webconfs.com/stop-words.php`
- `http://www.nltk.org/book/ch02.html` (NLTK stop words list)

We also add to the list of stop words the following terms that are ubiquitous in the patent text but don't provide information regarding the content and purpose of the patent: abstract, claim, claims, claimed, claiming, present, invention, united, states, patent, description, and background. The final stop word list contains 1337 unique terms that are filtered out.

Even after removing stop words, we expect much of the remaining text to offer little information regarding the purpose and use of a given patent or the core job functions expected to be performed by workers in a given occupation. In order to focus on the parts of the document most likely to contain relevant information, we retain descriptive and action words–i.e. nouns and verbs–and remove all other tokens. We do this using the part-of-speech tagger from the NLTK Python library. Finally, we lemmatize all remaining nouns and verbs, which is to convert them to a common root form. This converts all nouns to their singular form and verbs to their present tense. We use the NLTK WordNet Lemmatizer to accomplish this task. After these steps are completed, we have a set of cleaned lists of tokens for each patent and each occupation's tasks that we can then use to compute pairwise similarity scores.

2.6.2 Description of Word Embedding Vectors

To appreciate how our metric differs from the standard bag-of words approach it is useful to briefly examine how word embeddings are computed in Pennington et al. (2014). Denote the matrix X as a $V \times V$ matrix of word co-occurence counts obtained over a set of training documents, where V is the number of words in the vocabulary. Then $X_{i,j}$ tabulates the number of times word j appears in the context of the word i.[19] Denote $X_i = \sum_k X_{i,k}$ as the number of times any word appears in the context of word i, and the probability of word j occuring in the context of word i is $P_{i,j} \equiv X_{i,j}/X_i$. The goal of the word embedding approach is to construct a mapping $F(\cdot)$ from some d-dimensional vectors x_i, x_j, and $\tilde{x}_k$ such that

$$F(x_i, x_j, \tilde{x}_k) = \frac{P_{i,k}}{P_{j,k}} \tag{2.37}$$

Imposing some conditions on the mapping $F(\cdot)$, they show that a natural choice for modeling $P_{i,k}$ in (2.37) is

$$x_i^T \tilde{x}_k = \log(X_{i,k}) - \log(X_i) \tag{2.38}$$

Since the mapping should be symmetric for i and k they add "bias terms" (essentially i and k fixed effects) which gives

$$x_i^T \tilde{x}_k + b_i + b_k = \log(X_{i,k}) \tag{2.39}$$

Summing over squared errors for all pairwise combinations of terms yields the weighted least squares objective

$$\text{Min}_{x_i, \tilde{x}_k, b_i, b_k} \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{i,j}) \left(x_i^T \tilde{x}_k + b_i + b_k - \log(X_{i,j}) \right)^2 \tag{2.40}$$

Here the observation-specific weighting function $f(X_{i,j})$ equals zero for $X_{i,j} = 0$ so that the log is well defined, and is constructed to avoid overweighting rare occurences or extremely frequent occurences. The objective (2.40) is a highly-overidentified least squares minimization problem. Since the solution is not unique, the model is trained by randomly instantiating x_i and $\tilde{x}_k$ and performing gradient descent for a pre-specified number of iterations, yielding d-dimensional vector representations of a given word. Here d is a hyper-parameter; Pennington

[19] Pennington et al. (2014) use a symmetric 10 word window to determine "context" and weight down occurences that occur further away from the word (one word away receives weight 1, two words away receives weight 1/2, etc.).

et al. (2014) find that $d = 300$ works well on word analogy tasks.

Since (2.40) is symmetric it yields two vectors for word i, x_i and $\tilde{x}_i$, so the final word vector is taken as the average of the two. The ultimate output is a dense 300-dimensional vector for each word i that has been estimated from co-occurence probabilities and occupies a position in a word vector space such that the pairwise distances between words (i.e. using a metric like the cosine similarity) are related to the probability that the words occur within the context of one another and within the context of other similar words. Note that the basis for this word vector space is arbitrary and has no meaning; distances between word embeddings are only well-defined in relation to one another and a different training instance of the same data would yield different word vectors but very similar pairwise distances between word vectors.

Our method for backing out a geometric representation of the "meaning" of a document in (2.7) is to construct a weighted average of the meaning of all words in the document. Thus our vector representation of documents retains the 300-dimensional structure of the individual word constituents; these vectors are much denser and smaller than the very large and sparse document vectors in the standard bag of words methodology. In brief, there are two key characteristics that differentiate our approach relative to bag of words techniques. First, X_i is no longer a sparse vector like V_i. Moreover, because of the way word vectors are estimated, our method allows vectors containing similar words to be "close" to one another. Thus, relative to the bag of words approach our method: (1) constitutes a large dimensionality reduction; and, (2) can incorporate a notion of synonyms/distances between word meanings.

Armed with a vector representation of the document that accounts for synonyms, we next use the cosine similarity to measure the similarity between patent i and occupation j:

$$\text{Sim}_{i,j} = \frac{X_i}{||X_i||} \cdot \frac{X_j}{||X_j||} \tag{2.41}$$

This is the same distance metric as the bag of words approach, except now X_i and X_j are dense vectors carrying a geometric interpretation akin to a weighted average of the semantic meaning of all nouns and verbs in the respective documents.

To illustrate the difference between our approach and the standard bag of words, consider the following example of two documents, with the first document containing the words 'dog' and 'cat' and the other containing the words 'puppy' and 'kitten'. Even though the two documents carry essentially the same meaning, the bag of words approach will conclude that they are distinct: the representation of the two documents is

$$V_1 = [1, 1, 0, 0], \quad \text{and} \quad V_2 = [0, 0, 1, 1] \tag{2.42}$$

which implies that the two documents are orthogonal, $\rho_{1,2} = 0$. Here, the TF-IDF weights in our simple example satisfy $TF_{1,dog} = 1/2$ and $IDF_{dog} = \log(2)$, with similar logic applying to "cat"; this proceeds analogously for document 2 containing "puppy" and "kitten".

By contrast, in the word embeddings approach, these two documents are now represented as

$$X_1 = (1/2) \times \log(2) x_{dog} + (1/2) \times \log(2) x_{cat} \tag{2.43}$$

and similarly for X_2. Here x_{dog}, x_{cat} would have been trained using the Pennington et al. (2014) method described above on a very large outside set of documents. Hence, in this case since word vectors are estimated such that $x_{dog} \approx x_{puppy}$ and $x_{cat} \approx x_{kitten}$, we now have $\text{Sim}_{1,2} \approx 0.81$ using the word vectors estimated by Pennington et al. (2014). A weighted average word embedding approach has been shown in the natural language processing literature to achieve good performance on standard benchmark tests for evaluating document similarity metrics relative to alternative methods that are much more costly to compute (see, e.g. Arora, Liang, and Ma, 2017). A relative disadvantage is that it ignores word ordering—which also applies to the more standard 'bag of words' approach for representing documents as vectors. However, since we have dropped all stop words and words that are not either a noun or a verb, retaining word ordering in our setting is far less relevant.

Last, our methodology bears some similarities to recent work by Webb (2019), who also analyzes the similarity between a patent and O*NET job tasks. Webb (2019) focuses on similarity in verb-object pairs in the title and the abstract of patents with verb-object pairs in the job task descriptions and restricts his attention to patents identified as being related to robots, AI, or software. He uses word hierarchies obtained from WordNet to determine similarity in verb-object pairings. By contrast, we infer document similarity by using geometric representations of word meanings (GloVe) that have been estimated directly from word co-occurence counts. In addition to employing a different methodology, we also have a broader focus: we compute occupation-patent distance measures for all occupations and the entire set of USPTO patents since 1836. Furthermore, we use not only the abstract but the entirety of the patent document—which includes the abstract, claims, and the detailed description of the patented invention.

2.6.3 Constructing the Industry Innovation Measure

For each breakthrough patent p we assign it to an industry using probabilistic patent CPC tech class to NAICS crosswalks constructed by Goldschlag et al. (2020). The Goldschlag et al. (2020) crosswalk assigns probabilities that patents from a given technology class originated from a particular NAICS industry for different levels of NAICS aggregation. The NBER manufacturing database reports data the 6-digit NAICS code level, and so we use the Goldschlag et al. (2020) 6-digit NAICS to 3-digit CPC probabilistic crosswalk. We then aggregate the data to the 4-digit NAICS level to parallel the level of industry classification we use in our analysis in 2.3.2 of the main text.

Label the set of breakthrough patents issued in year t by Γ_t; $\alpha_{j,p}$ the probability of breakthrough patent p being issued to industry j, and κ_t the US population in year t. We then define the industry-level breakthrough patent index (including only patents with high average textual similarity to the industry workforce) by

$$\psi_{j,t} = \frac{1}{\kappa_t} \sum_{p \in \Gamma_t} \alpha_{j,p}) \tag{2.44}$$

2.6.4 Census public-use data

We gather Census data from IPUMS and compute aggregate employment shares for occupations in Census years spanning 1850-2010. We use the 1950 Census occupation definition for pre-1950 Census years since the more updated 1990 Census classification scheme is only available in post-1950 Census years. We make use of the 1990 Census occupation classifications for the years they are available. We then crosswalk Census occupations to the David Dorn occ1990dd classification scheme using the crosswalk files provided on his website and aggregate our measure $\eta_{i,t}$ to the occ1990dd-level by averaging across 6-digit SOC codes within an occ1990dd code. This results in a Census-year by occ1990dd panel of occupation employment shares. Census records for the year 1890 were destroyed in a fire, and so the employment growth observations for the 20-year horizon in 1870 or for the 10-year horizon in 1880 are not available.

For the post-1980 results, we use the Current Population Survey Merged Outgoing Rotation Groups (MORG). We obtain the cleaned versions of MORG extracts provided by the Center for Economic Policy Research (CEPR). We use the "wage3" variable that combines the usual hourly earnings for hourly workers and non-hourly workers, which adjusts for top-coding using a lognormal imputation and is constructed to match the NBER's recommendation for the most consistent hourly wage series from 1979 to the present. Using these data we construct a time series of wage and employment growth for occupations at the occ1990dd level. Because occ1990dd cannot be crosswalked to a balanced panel of occupations using the Census 1970 occupation codes, we start our analysis in the post-1982 time period when these extracts began using the 1980 Census occupation classification scheme.

2.6.5 Census-CPS administrative data

We use a random sample of individual workers tracked by the Current Population Survey (CPS) and their associated Detailed Earnings Records from the Census—which contains their W2 tax income. We limit the sample to individuals who are older than 25 and younger than 55 years old.

The CPS includes information on demographic information such as age and gender, but more importantly occupation at the time of the interview. We assign workers to occupations based on their response to the CPS survey (CPS “occ” variable). We construct a crosswalk between the yearly CPS occupations codes and the occ1990dd classification scheme and assign all CPS occupations their corresponding occ1990dd code. We assign this occupation to the worker for the next 5 years, thus effectively dropping observations where the CPS interview date is older than 5 years—so that the occupation information is relatively recent.

We merge the individual worker records from the Census-CPS matched sample to patent data at the industry (NAICS 4) level. Specifically, we identify the industry of where the patent origination by relying on the Census SSEL patent–assignee database, which provides a corresponding SSEL firm identifier (“firmid"), which we then use to obtain the firms' 4-digit NAICS code. In particular, we use two SSEL patent–assignee crosswalks: the newer Business Dynamics Statistics of Patenting Firms database (BDS-PF) and an older patent-SSL crosswalk created by Kerr and Fu (2008). The BDS-PF links are available starting with the 2000 SSL. We use the BDS-PF firmid-patent links for any patents for which it is available. Otherwise we rely on the links from Kerr and Fu (2008) created from the 1999 SSL. In cases where a firmid matches to multiple NAICS codes we apply the 4-digit NAICS code of highest employment. We drop any industry-year observations with no patents filed in a given year.

To allow the effects to vary with prior income, we assign workers into five groups based on their average income over the last three years (the last term in (2.18)) compared to workers in the same occupation and NAICS4 industry. These groups are defined based on the following percentiles of prior income [0%, 25%), [25%, 50%), [50%, 75%), [75%, 95%), [95%, 100%] calculated within industry–occupation cells. In the (uncommon) case when NAICS4 industries have cells which are too small to rank, we broaden the industry definition from 4 digit NAICS to 2 digit NAICS. Subsequently, any Industry–Occupation cells with fewer than 10 individuals are dropped.

2.6.6 Constructing a Statistical Displacement Factor

To construct our predictor we use a method proposed by Cong et al. (2019), which is well-suited to prediction exercises using large-scale textual data. Our adaptation of their method for the task of predicting occupation outcomes can be summarized in the following steps. Let the number of patent documents be N_p (where we restrict just to the set of breakthrough patents from Kelly et al. (2020) as described in section 2.2), the number of occupation task descriptions be N_o, and the number of words in the vocabulary formed from the union of all patent and occupation documents be N_w:

1. Perform approximate nearest neighbor search using a locality-sensitive hashing routine (LSH) on vector representations of word meanings to form K clusters ("topics") of related words. Label the kth cluster of words C_k.

2. Create a $N_p \times N_w$ matrix of breakthrough patent documents by word counts weighted by term-frequency inverse document frequency (TF-IDF), computed over all patents (i.e. TF-IDF is computed also including non-breakthrough patents). Call this matrix A. Loop over each word cluster C_k from step 1 for $k = 1, \dots, K$, and extract the submatrix of A formed by taking the columns in A corresponding to the words contained in cluster C_k. Call this submatrix A_k. Perform a singular-value decomposition of A_k and take its top singular value v_k (in absolute value) and corresponding top right singular vector V_k. Then take the $N_p \times 1$ vector $P_k = \frac{|A_k v_k|}{v_k' v_k}$ to be the loadings of each patent document on topic/word cluster k. Retain only the clusters C_k which rank in the top 500 based on their top absolute singular values.

3. Perform step 2 for all occupations, except only for the top 500 clusters that were retained. Call the resulting $N_o \times 1$ vector of occupation loadings O_k. Denote the set breakthrough of patents issued in year t by $\widehat{\Gamma}_t$. Let $O_{k,i}$ represent the ith element of O_k and $P_{k,j}$ the jth element of P_k, the vector of patent loadings on cluster k. Then

occupation i's exposure to the kth topic in year t is given by

$$\psi_{i,k,t} = \frac{O_{k,i}}{\kappa_t} \sum_{j \in \widehat{\Gamma}_t} P_{k,j} \tag{2.45}$$

As before we only sum over breakthrough patents and normalize by U.S. population in year t (denoted by κ_t). This yields an occupation's exposure in each year to the 500 topics which are found to be the most important among the breakthrough patents. Though equation 2.45 looks a bit like our construction of $\eta_{i,t}$ in equation 2.12, it differs in that we no longer directly use word vectors to compute similarities. Instead, the Cong et al. (2019) technique only uses the word vectors to give an educated guess on the topics contained in the set of documents. Thus occcupations are similar to a given topic when they contain words that are also found in that topic.

We focus on the period of time covered by our CPS merged outgoing rotation group sample (1985-2018) used in the employment regressions in Figure 2-8. This is for two reasons: first, this is the period where our employment and wage data coverage is most comprehensive, with a yearly time series and relatively stable occupation classifications. Second, the task composition of innovations has begun to change in this period of time relative to all previous innovation waves. In particular, cognitive skills have started to become more related to innovations, and this has been driven by the rising importance of information technology and electronics patents, which was not the case prior to the late 20th century. If skill-biased technological change has complemented the skillset of cognitive occupations, then innovations related to these occupations may be complementary to rather than a substitute for their skills. Thus if our measure mixes these two channels it is particularly likely to occur during this period of time.

Steps 1 and 2 above simply group documents into topics of related terms, compute how related a given topic is to each individual document, and provide an estimate of how important each topic is to the overall set of documents. Justification for the use of LSH clustering of word vectors to obtain topics and the singular value decomposition to infer topic importance/document topic loadings are discussed at length in Cong et al. (2019), to which we refer the interested reader for further details. For our purposes it suffices that by performing steps 1 through 3 we are able to obtain a panel of 500 predictors at the occupation-by-year level and which represent exposures to topics of words which are particularly relevant to patents.

In brief, this approach can be summarized as follows. We first extract the 500 most important common factors (topics) from the text of breakthrough patents using the approach of Cong et al. (2019) and the vector representations of word embedings discussed in Section 2.2. We then use these 500 textual factors to form a single predictor that is optimized to predict occupation declines in-sample. To do so, we examine the univariate performance of each factor in predicting employment declines, and then form a linear combination (the first principal component) of the predictors that are statistically significant negative predictors at the 5%. We also construct a labor-enhancing factor using the converse exercise.

Appendix Table 2.16 summarizes our findings. By design, both factors predict employment with the correct sign in-sample. More importantly, both of these factors predict wage growth with the same sign, despite the fact that they were not designed to do so and wage growth is not highly correlated with employment growth. That said, the displacement factor (the factor calibrated for employment declines) is a much stronger predictor of both employment and wage growth than its counterpart designed to predict positive employment growth.

In terms of magnitudes, the employment and wage declines predicted by this statistical displacement factor are comparable to our baseline measure—that is, 1.25% vs 1.12% employment declines at the 10-year horizon and 0.25% vs 0.20% decline in wage earnings). The correlation between our baseline measure $\eta_{i,t}$ and the statistical predictor constructed to represent exposure to labor-saving technologies is approximately 73 percent. By contrast, the correlation with the factor calibrated to predict employment increases is negative at -11 percent.

2.6.7 Model Appendix

The model we use consider a continuum of workers, with a state parameter θ on the [0, 1] interval, corresponding to their ability to produce H. Share s_h of workers have the ability to accumulate H over time, and have H reset by a certain amount when new technology enters the playing field. Technology is given by ξ and has shocks of size κ, which (in expectation) displaces the human capital of α share of workers, reducing their θ by m.

Production is given by a nested CES production function, where composite good X is produced via a combination of L and ξ

$$X = \left(\xi^{\rho}\lambda + L^{\rho}(1-\lambda)\right)^{(1/\rho)}.$$

Output, Y, is produced as a combination of X and H:

$$Y = \left(\mu H^{\sigma} + (1-\mu)X^{\sigma}\right)^{(1/\sigma)}.$$

Technology evolves with process

$$\xi_t = (1-g)\xi_{t-1} + \kappa\, d\, N_t$$

where dN is a random variable with expectation ω.

Worker $i \in s_h$ has evolving human capital such that

$$\theta_{i,t} = m\,\theta_{i,t-1}\,dM_{i,t} - d_{i,t}\,h\,\theta_{i,t-1}.$$

In this case, dM is a random variable representing human capital acquisition, m is the size of the jump relative to initial human capital, dN is the same shock variable as in the equation for technology, and h_{-} is the scale of the loss of human capital (H) if a shock occurs. $d_{i,t}$ is an i.i.d. binomial random variable with expectation $E(d) = \alpha$, indicating whether someone is "exposed" to a technology shock or not. If you are exposed to a technology shock when one occurs, you experience the human capital loss, otherwise you do not.

We solve the model in discrete time, with a monthly time-step δt, and we approximate the continuum of workers with an exponentially increasing, finite grid of points on the [0, 1] interval. Since we are approximating a continuum of workers, each gridpoint has an infinite number of observations, and we can work directly with expectations when solving the model. This means for a given starting grid point on the θ interval, we have

$$E(\theta_{i,t}) = \theta_{i,t-1} + m\,\phi\,\theta_{i,t-1} - \alpha\,h\,\theta_{i,t-1}$$

Note that in discrete time, the technology and human capital processes admit a two-state Markov-Switching VAR representation with a shock state ($s = 0$) and a no-shock state $s = 1$. Let i index the starting gridpoint of a worker, with $i + 1$ and $i - 1$ being the adjacent gridpoints.

With some abuse of notation, instead of thinking of $\theta_{i,t}$ as a single worker on the grid, we can think of it as the probability mass (share of workers) on a gridpoint i.

$$E(\theta_{i,t}|s = 0) = \theta_{i,t-1} - \phi\theta_{i+1,t-1} + \phi\theta_{i-1,t-1}$$

This works because we set the distance between gridpoints is m. If we have a shock, we have

$$E(\theta_{i,t}|s = 1) = \theta_{i,t-1} - \phi\theta_{i+1,t-1} + \phi\theta_{i-1,t-1} - \alpha\theta_{i,t-1}) + \alpha\,h\,\theta_{i+m,t-1}).$$

In other words, if we experience a shock, we get some mass from the gridpoint which is mh above us, lose share α of our previous density as exposed workers, lose share ϕ to a higher gridpoint as workers acquire new skills, and gain share ϕ from the gridpoint below.

We can represent this as a VAR process, with transition probability α to a gridpoint which is mh_{-} points below the current, ϕ to a gridpoint above us, and so on. In practice, in order to make h_{-} a continuous parameter, we split the fall probability across the two relevant gridpoints, with density allocated between them to make the fall have expectation h_{-}. For example, if m was 3%, and we needed a 5% fall, conditional on the fall a worker would have (roughly) a 2/3 chance of falling two gridpoints and a 1/3 chance of falling 1 gridpoint.

The transition process for the workers in s_l is very simple, as it is an absorbing state with no entry or exit. So

$$\theta_{s_l,t} = \theta_{s_l,t-1}$$

in both shock periods and no-shock periods.

The transition process for ξ is given by

$$\xi_t | s_t = 0 = (1-g)\xi_{t-1}$$

and

$$\xi_t | s_t = 1 = (1-g)\xi_{t-1} + \kappa.$$

Suppose we set up the VAR coefficient matrices, A, accordingly. Each period has probability ω of experiencing a shock, and probability $1-\omega$ of not experiencing a shock. This gives us transition process

$$E(A_t) = (1-\omega)A_{t-1,0} + \omega A_{t-1,1}.$$

Bianchi (2016) demonstrates how to find the steady state of the Markov-Sitching VAR model. For this exposition, we rely on his notation. He considers the MS-VAR process

$$Z_t = c_{\xi_t} + A_{\xi_t} Z_{t-1} + V_{\xi_t}\epsilon_t,$$

and

$$V_{\xi_t} = R_{\xi_t}\Sigma_{\xi_t},$$

where z_t is a vector of variables, c_t is a vector of constants,

In practice, our process for the mass of θ has a reflecting barrier at 1, and in order to enforce the sum-to-1 constraint for the total mass on the θ grid we represent the VAR through a VECM process.

We then solve for the steady state by finding the largest real eigenvalue of the Bianchi representation of the MS-VAR system. Because we have an absorbing state at the bottom of the θ grid whose density is a known value (fixed before calibration), we exclude that point from the solution, scaling down the intercept term by $1 - s_l$. Finally, to compute the value at the top of the θ grid (call it gridpoint j), we simply compute $1 - \sum_{i\ j} \theta_{ss,i} - s_l$.

Once we've solved for the ergodic steady state, we can begin calculating the model moments which correspond to our empirical calibration targets. Our process sets time step δt to one month. Our theoretical moments are calculated as one-month impact responses, scaled by an annualization factor $\sqrt{12\omega(1-\omega)}$ for aggregate moments (labor share and output), and $\sqrt{12\omega\alpha(1-\omega\alpha)}$.

Impact responses for labor share and output are calculated relative to the ergodic steady state for all state variables. The steady state values are iterated forward one period using the transition matrices constructed above, but where a shock happens with certainty (effectively setting $\omega = 1$ for a single period). To compare output, labor share, wages, and other desired targets, between the values at the ergodic steady state relative to the shock period, we follow this procedure in each period. First, we calculate the level of H at the steady state as

$$H = \sum_{i=1}^{N} \theta_i m(\theta_i)$$

where $m(theta_i)$ is the mass of theta at gridpoint i. The workers in s_l produce no H, so this is sum of the value for θ at each gridpoints times the mass of workers at that rung of the ladder. L is calculated as $1 - H$. Output Y and the composite good X are calculated with the equations provided above. σ, ρ, μ, and λ are free parameters. If we call ξ^* the value of the technology state variable at the ergodic steady state, ξ post-shock is $\xi^* + \kappa$, where κ is a free variable. Wages associated with H (w_h) and L (w_l) are calculated as the marginal product of each task. Given output, these are calculated as

$$w_h = \frac{(1-\mu)\mu H^{\sigma-1}(\lambda\xi^\rho - L^\rho(\lambda-1))^{(\sigma/\rho)} + \mu H^\sigma)^{(1/\sigma)}}{(1-\mu)(\lambda\xi^\rho - L^\rho(\lambda-1))^{(\sigma/\rho)} + \mu H^\sigma},$$

and

$$w_l = \frac{(1-\mu)(\lambda\xi^{\rho} - L^{\rho}(\lambda-1))^{(\sigma/\rho)} + \mu H^{\sigma})^{(1/\sigma)}(\lambda-1)(\lambda\xi^{\rho} - L^{\rho}(\lambda-1))^{(\sigma/\rho)}(\mu-1)L^{\rho-1}}{(\lambda\xi^{\rho} + (1-\lambda)L^{\rho})((1-\mu)(\lambda\xi^{\rho} - L^{\rho}(\lambda-1))^{(\sigma/\rho)} + \mu H^{\sigma})}$$

For each period in question for the impact responses, wages are calculated by plugging in the relevant state variables. Impact responses are calculated in log differences to align with our calibration targets (e.g. $\log Y_{shock} - \log Y_{ss}$), and subsequently scaled by the annualization factor.

2.7 Appendix Figures and Tables

Table 2.10: Most Similar Patents For Select Occupations

Cashiers (SOC Code 412011)	
5055657	Vending type machine dispensing a redeemable credit voucher upon payment interrupt
5987439	Automated banking system for making change on a card or user account
5897625	Automated document cashing system
6012048	Automated banking system for dispensing money orders, wire transfer and bill payment
5598332	Cash register capable of temporary-closing operation
Loan Interviewers and Clerks (SOC Code 434131)	
6289319	Automatic business and financial transaction processing system
5611052	Lender direct credit evaluation and loan processing system
6233566	System, method and computer program product for online financial products trading
5940811	Closed loop financial transaction method and apparatus
5966700	Management system for risk sharing of mortgage pools
Railroad Conductors (SOC Code 534031)	
5828979	Automatic train control system and method
6250590	Mobile train steering

3944986	Vehicle movement control system for railroad terminals
6135396	System and method for automatic train operation
5797330	Mass transit system

Table 2.11: Most Similar Occupations For Select Patents

“Knitting-machine” (Patent No. 276146, Issued in 1883)
Textile Knitting and Weaving Machine Setters, Operators, and Tenders
Sewing Machine Operators
Sewers, Hand
Fabric Menders, Except Garment
Textile Winding, Twisting, and Drawing Out Machine Setters, Operators, and Tenders
“Metal wheel for vehicles” (Patent No. 1405358, Issued in 1922)
Automotive Service Technicians and Mechanics
Cutting, Punching, and Press Machine Setters, Operators, and Tenders, Metal and Plastic
Maintenance Workers, Machinery
Grinding, Lapping, Polishing, and Buffing Machine Tool Setters, Operators, and Tenders, Metal and Plastic
Rolling Machine Setters, Operators, and Tenders, Metal and Plastic
“System for managing financial accounts by a priority allocation of funds among accounts”

(Patent No. 5911135, Issued in 1999)
Financial Managers
Credit Analysts
Loan Interviewers and Clerks
Accountants and Auditors
Bookkeeping, Accounting, and Auditing Clerks

Table 2.12: Occupations Most and Least Exposed to Innovation

Top 5 Occupations by Average $\eta_{i,t}$	Bottom 5 Occupations by Average $\eta_{i,t}$
Inspectors, Testers, Sorters, Samplers, and Weighers	Mental Health Counselors
Metal Workers and Plastic Workers, All Other	Dancers
Cutting, Punching, and Press Machine Setters, Operators, and Tenders, Metal and Plastic	Funeral Attendants
Production Workers, All Other	Judges, Magistrate Judges, and Magistrates
Electromechanical Equipment Assemblers	Clergy

Table 2.13: Unconditional Correlations of $\eta_{i,t}$ With Task Categories

NR Cog (Analytical)	-0.12**					
	(-2.53)					
NR Cog (Interpersonal)		-0.16***				
		(-4.65)				
NR Man (Physical)			0.24***			
			(5.65)			
NR Man (Interpersonal)				-0.33***		
				(-8.43)		
Routine Cognitive					0.033	
					(0.95)	
Routine Manual						0.24***
						(5.74)

This figure plots the correlations of $\eta_{i,t}$ with the occupation task types computed from O*NET in Acemoglu and Autor (2011). Correlations are weighted by the Acemoglu and Autor (2011) occupation employment weights used to normalize the distribution of tasks to mean zero and standard deviation one.

Table 2.15: Technology And Employment During and Outside of Innovation Waves

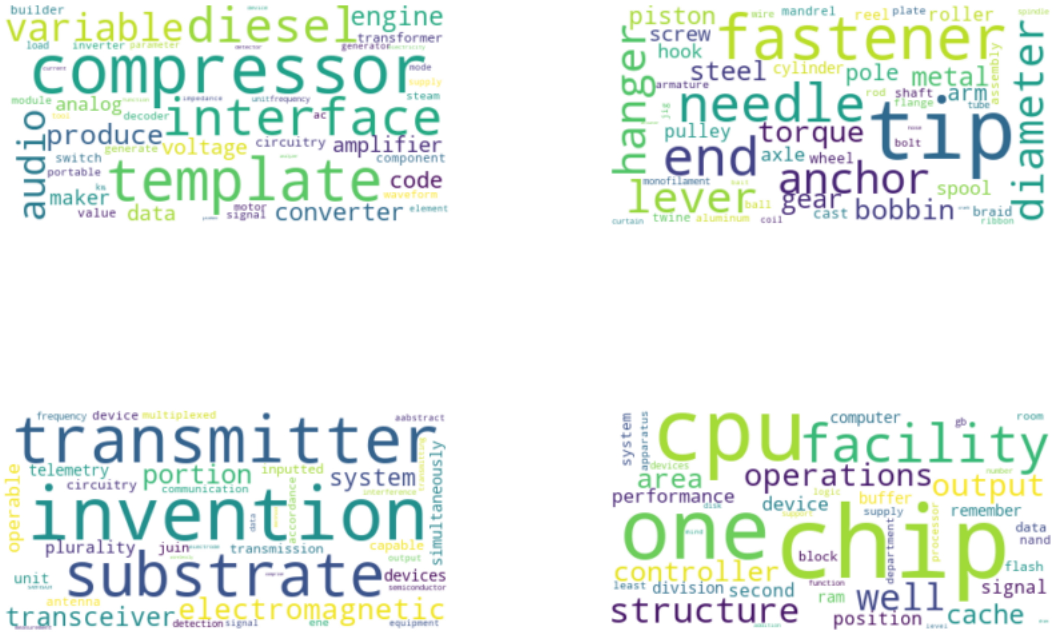
	Innovation Wave	Other Years
Technology Exposure, $\eta_{i,t}$	-0.82***	-0.53
	(-5.92)	(-1.54)
Time FE	X	X
N	1106	1468
R^2 (Within)	0.091	0.018

The table above plots results from regressions of the form

$$\log(Y_{i,t+k}) - \log(Y_{i,t}) = \alpha_0 + \alpha_t + \beta\eta_{i,t} + \epsilon_{i,t}$$

for $k = 20$ years for Census years spanning from 1850-2000. Here $Y_{i,t}$ is occupation's share in total non-farm employment. $\eta_{i,t}$ is standardized and growth rates are in annualized percentage terms. The sample is split into periods of innovation waves as identified by the breakthrough patent index of Kelly et al. (2020). The 20 year periods beginning in years 1880, 1910, 1920, and 1980, 1990 are labelled innovation waves with the remaining years representing non-innovation wave periods. Standard errors are clustered by occupation and corresponding t-stats are shown in parentheses. Observations are weighted by occupation employment share at time t.

Figure 2-14: Sample Patent Topic Word Clusters



The above are four of the topics resulting from the LSH approximate nearest neighbors routine used to separate words into clusters as described in section 2.3.3. The relative size of the word corresponds to the importance of that word within the topic.

Table 2.16: Predictive Performance of 10-Year Employment and Wage Growth on Predictors Constructed From Patent Topics

Panel A: Negative Constructed Predictor

	Employment Growth		Wage Growth	
ξ_{Mean}	-1.25***		-0.21***	
	(-6.79)		(-6.10)	
ξ_{PC1}		-1.09***		-0.20***
		(-6.32)		(-6.11)
Year FEs	X	X	X	X
Controls	X	X	X	X

Panel B: Positive Constructed Predictor

	Employment Growth		Wage Growth	
γ_{Mean}	0.59***		0.019	
	(3.84)		(0.70)	
γ_{PC1}		0.50***		0.0062
		(3.67)		(0.24)
Year FEs	X	X	X	X
Controls	X	X	X	X

Note: The tables above show coefficients from panel regressions of annualized wage and income growth rates over the 10-year horizon on textual factors constructed to predict employment as described in section 2.3.3. Regressions are of the form

$$y_{i,t+k} - y_{i,t} = \alpha + \beta Z_{i,t} + \delta X_{i,t} + \varepsilon_{i,t}$$

For $Z_{i,t} = \xi_{i,t}$ ("labor-saving") or $\gamma_{i,t}$ ("productivity enhancing"). Controls $X_{i,t}$ include three one-year lags of dependent variable, time fixed effects, wage, and occupation employment share. Subscripts $PC1$ and $Mean$ denote versions computed using either the first principal component or cross-sectional mean across individual textual predictors derived from the patent topics identified by the Cong et al. (2019) method. Dependent variable is expressed in annualized percentage terms and $\eta_{i,t}$ is standardized. Standard errors are clustered by occupation and independent variables are standardized. Observations are weighted by occupation's employment share at time t. The sample uses CPS merged outgoing rotation group data starting in 1982.

Table 2.17: Correlations Between Predictors Constructed From Patent Topics and Different Versions of Occupation Technology Exposure $\eta_{i,t}$

	All Patents		Drop ICT Patents		Just ICT Patents	
ξ_{Mean}	0.73^{***}		0.88^{***}		0.41^{***}	
	(20.53)		(31.71)		(10.21)	
γ_{Mean}		-0.11^{***}		-0.17^{***}		-0.030
		(-5.26)		(-6.39)		(-1.13)

Note: This table reports correlations between versions of technology exposure $\eta_{i,t}$ formed using different sets of patents and the composite predictors constructed from textual factors using the Cong et al. (2019) method to predict employment outcomes either negatively (ξ_{Mean}) or positively (γ_{Mean}). The "$Mean$" label denotes versions of composite predictors constructed by taking the cross-sectional means across individual textual factors which predict employment either negatively or positively. The first two columns represent the baseline measure of $\eta_{i,t}$ constructed using all patents; the next two columns drop ICT patents, defined to be those falling under the instruments/information or electronics categories; finally, the last two columns form $\eta_{i,t}$ only using ICT patents.

Table 2.18: Breakthrough patents most related to tasks performed by order-fulfillment clerks

US. Pat. #	Distance ($\tilde{\rho}$)	Issue Year	Title
5,696,906	0.933	1997	Telecommunication user account management system and method
5,627,973	0.915	1997	Method and apparatus for facilitating evaluation of business opportunities for supplying goods and/or services to potential customers
5,689,705	0.896	1997	System for facilitating home construction and sales
5,592,560	0.885	1997	Method and system for building a database and performing marketing based upon prior shopping history
5,687,212	0.885	1997	System for reactively maintaining telephone network facilities in a public switched telephone network
5,628,004	0.881	1997	System for managing database of communication of recipients
5,621,812	0.880	1997	Method and system for building a database for use with selective incentive marketing in response to customer shopping histories
5,638,457	0.880	1997	Method and system for building a database for use with selective incentive marketing in response to customer shopping histories
5,659,469	0.879	1997	Check transaction processing, database building and marketing method and system utilizing automatic check reading
5,592,378	0.874	1997	Computerized order entry system and method
5,787,405	0.896	1998	Method and system for creating financial instruments at a plurality of remote locations which are controlled by a central office
5,802,513	0.884	1998	Method and system for distance determination and use of the distance determination
5,717,596	0.878	1998	Method and system for franking, accounting, and billing of mail services
5,797,002	0.873	1998	Two-way wireless system for financial industry transactions
5,812,985	0.866	1998	Space management system
5,774,877	0.866	1998	Two-way wireless system for financial industry transactions
5,848,396	0.865	1998	Method and apparatus for determining behavioral profile of a computer user
5,790,634	0.865	1998	Combination system for proactively and reactively maintaining telephone network facilities in a public switched telephone system
5,734,823	0.864	1998	Systems and apparatus for electronic communication and storage of information
5,712,987	0.864	1998	Interface and associated bank customer database

Table 2.19: Worker Earnings and Technology Exposure, by Prior Income and Education / Age

	(1)	(2)	(3)	(4)	(5)
Income Percentile		Age		Education	
	25–35	35–45	45–55	College	No Coll
A. Cond. Mean, $E[g]$, by Horizon					
0 to 25-th	-1.05	-1.72	-2.03	-2.14	-1.11
	(-2.57)	(-6.36)	(-5.06)	(-6.32)	(-3.58)
25 to 50-th	-0.28	-0.87	-2.24	-0.92	-1.11
	(-0.97)	(-2.89)	(-3.18)	(-2.72)	(-4.40)
50 to 75-th	-0.57	-0.68	-2.43	-0.82	-1.60
	(-1.46)	(-2.09)	(-3.32)	(-2.02)	(-4.36)
75 to 95-th	-0.39	-0.49	-2.22	-0.60	-2.13
	(-0.53)	(-1.56)	(-4.73)	(-1.37)	(-7.52)
95 to 100-th	-3.07	-1.85	-2.86	-1.98	-3.50
	(-4.22)	(-3.07)	(-3.51)	(-4.10)	(-4.47)
B. Risk, Absolute Income Growth $E[\|g\|]$					
0 to 25-th	0.01	-0.55	0.06	-0.51	-0.18
	(0.04)	(-2.13)	(0.18)	(-1.74)	(-0.77)
25 to 50-th	0.20	-0.17	0.55	0.12	0.03
	(0.68)	(-0.81)	(0.83)	(0.48)	(0.11)
50 to 75-th	0.42	0.06	1.17	0.50	0.49
	(1.66)	(0.31)	(1.86)	(1.85)	(1.44)
75 to 95-th	0.61	0.12	1.51	0.66	1.01
	(1.56)	(0.45)	(3.84)	(2.87)	(2.94)
95 to 100-th	2.38	1.44	2.38	2.01	2.09
	(3.05)	(3.03)	(3.92)	(6.02)	(2.76)
C. Skewness, Prob. Large Income Decline $p(g < p^{10})$					
0 to 25-th	0.18	0.25	0.68	0.40	0.20
	(0.99)	(1.93)	(4.37)	(2.5)	(1.39)
25 to 50-th	0.07	0.16	0.73	0.24	0.28
	(0.3)	(0.86)	(2.42)	(1.37)	(1.34)
50 to 75-th	0.18	0.16	0.91	0.26	0.56
	(1.26)	(1.08)	(3.05)	(1.82)	(4.03)
75 to 95-th	0.07	0.05	0.90	0.15	0.83
	(0.28)	(0.27)	(3.99)	(0.87)	(6.28)
95 to 100-th	2.42	0.68	1.46	1.14	1.53
	(4.52)	(2.24)	(4.43)	(4.56)	(4.89)

Note: Panel A shows the estimated slope coefficients (times 100) from equation (2.21) in the main text: the dependent variable is worker earnings growth over a 5-year horizon; the main independent variable of interest is a worker's technology exposure $\eta_{i,t}$. The slope estimate $\beta(h)$ is allowed to vary with the worker's prior income rank and age (columns (1) to (3)) or education (columns (4) to (5)). Panel B shows the slope coefficients of a variant of the above specification where we replace the dependent variable $g_{i,t:t+h}$ with its absolute value $|g_{i,t:t+h}|$. Panel C replaces the dependent variable with a dummy that takes the value of one if $g_{i,t:t+h}$ lies in the bottom 10-th percentile. See notes to Tables 2.3 to 2.5 for additional details.

Table 2.20: Worker Earnings and Technology Exposure, by Education

Horizon	Education	(1)	(2)	(3)	(4)
3 Years	College	-0.975	-0.880	-0.924	-0.911
		(0.26)	(0.246)	(0.292)	(0.274)
	No College	-1.32	-1.27	-1.30	-1.30
		(0.153)	(0.168)	(0.253)	(0.233)
	No College - College	-0.347	-0.391	-0.378	-0.388
		(0.209)	(0.197)	(0.199)	(0.192)
5 Years	College	-1.35	-1.17	-1.20	-1.14
		(0.291)	(0.282)	(0.362)	(0.343)
	No College	-1.68	-1.55	-1.54	-1.50
		(0.216)	(0.225)	(0.31)	(0.291)
	No College - College	-0.336	-0.387	-0.332	-0.3603
		(0.2315)	(0.225)	(0.230)	(0.225)
10 Years	College	-1.24	-1.08	-1.31	-1.21
		(0.352)	(0.343)	(0.504)	(0.486)
	No College	-1.82	-1.67	-1.85	-1.76
		(0.281)	(0.262)	(0.404)	(0.393)
	No College - College	-0.581	-0.581	-0.544	-0.550
		(0.283)	(0.274)	(0.290)	(0.283)
Fixed Effects:					
Industry		X	X		
Occupation		X		X	
Year		X			
Industry x Year				X	X
Occupation x Year			X		X

Table 2.21: Worker Earnings and Technology Exposure, by Age

Worker Age	Horizon	(1)	(2)	(3)	(4)
3 Years	25-35	-0.479	-0.396	-0.389	-0.388
		(0.268)	(0.244)	(0.224)	(0.209)
	35-45	-0.903	-0.83	-0.866	-0.857
		(0.18)	(0.159)	(0.181)	(0.164)
	45-55	-1.98	-1.91	-1.96	-1.95
		(0.392)	(0.42)	(0.542)	(0.524)
	(45-55) - (25-35)	-1.50	-1.51	-1.57	-1.56
		(0.428)	(0.441)	(0.463)	(0.462)
5 Years	25-35	-0.874	-0.706	-0.683	-0.64
		(0.243)	(0.227)	(0.271)	(0.251)
	35-45	-1.23	-1.08	-1.09	-1.04
		(0.148)	(0.148)	(0.221)	(0.204)
	45-55	-2.44	-2.29	-2.3	-2.25
		(0.535)	(0.542)	(0.656)	(0.634)
	(45-55) - (25-35)	-1.56	-1.59	-1.62	-1.61
		(0.489)	(0.494)	(0.516)	(0.512)
10 Years	25-35	-0.986	-0.823	-1.01	-0.918
		(0.291)	(0.266)	(0.397)	(0.38)
	35-45	-1.43	-1.26	-1.47	-1.37
		(0.238)	(0.228)	(0.39)	(0.377)
	45-55	-2.54	-2.40	-2.60	-2.51
		(0.682)	(0.676)	(0.861)	(0.832)
	(45-55) - (25-35)	-1.56	-1.58	-1.59	-1.59
		(0.565)	(0.564)	(0.579)	(0.574)
Fixed Effects:					
Industry		X	X		
Occupation		X		X	
Year		X			
Industry x Year				X	X
Occupation x Year			X		X

Table 2.22: Worker Earnings and Technology Exposure, by Prior Income

Horizon	Income Rank	(1)	(2)	(3)	(4)
3 Years	$[0, 25)$	-1.34	-1.21	-1.28	-1.24
		(0.236)	(0.239)	(0.207)	(0.216)
	$[25, 50)$	-0.934	-0.811	-0.883	-0.849
		(0.145)	(0.167)	(0.212)	(0.205)
	$[50, 75)$	-1.10	-0.974	-1.05	-1.01
		(0.241)	(0.244)	(0.304)	(0.289)
	$[75, 95)$	-1.14	-1.00	-1.09	-1.05
		(0.28)	(0.266)	(0.26)	(0.256)
	$[95, 100]$	-2.31	-2.20	-2.25	-2.24
		(0.421)	(0.422)	(0.374)	(0.366)
5 Years	$[0, 25)$	-1.76	-1.51	-1.56	-1.49
		(0.234)	(0.258)	(0.267)	(0.281)
	$[25, 50)$	-1.26	-1.02	-1.08	-1.01
		(0.186)	(0.21)	(0.281)	(0.282)
	$[50, 75)$	-1.43	-1.19	-1.24	-1.18
		(0.33)	(0.323)	(0.395)	(0.383)
	$[75, 95)$	-1.42	-1.17	-1.24	-1.17
		(0.328)	(0.315)	(0.345)	(0.339)
	$[95, 100]$	-2.71	-2.48	-2.52	-2.47
		(0.578)	(0.574)	(0.516)	(0.52)
10 Years	$[0, 25)$	-2.05	-1.74	-2.00	-1.85
		(0.36)	(0.337)	(0.395)	(0.393)
	$[25, 50)$	-1.14	-0.858	-1.1	-0.964
		(0.338)	(0.349)	(0.474)	(0.472)
	$[50, 75)$	-1.19	-0.90	-1.16	-1.01
		(0.414)	(0.40)	(0.578)	(0.541)
	$[75, 95)$	-0.986	-0.686	-0.967	-0.807
		(0.445)	(0.436)	(0.527)	(0.492)
	$[95, 100]$	-2.46	-2.20	-2.42	-2.28
		(0.589)	(0.594)	(0.611)	(0.596)
Fixed Effects:					
Industry		X	X		
Occupation		X		X	
Year		X			
Industry x Year				X	X
Occupation x Year			X		X

Table 2.23: Technology And Employment Over the Long Run (1850-2010)–Heterogenous effects by age

	A. Full Sample	B. Sub-samples		
		1850–1920	1930–1960	1970–1990
Age (20–29) × Technology Exposure, $\eta_{i,t}$	-0.71***	-2.24***	-0.073	-0.58**
	(-3.46)	(-4.08)	(-0.19)	(-2.33)
Age (30–39) × Technology Exposure, $\eta_{i,t}$	-0.57***	-1.70***	-0.098	-0.50**
	(-3.41)	(-3.28)	(-0.31)	(-2.44)
Age (40–49) × Technology Exposure, $\eta_{i,t}$	-1.10***	-2.13***	-0.62*	-1.03***
	(-6.20)	(-4.40)	(-1.88)	(-4.85)
Observations	6,512	2,232	1,989	2,291
R^2 (Within)	0.066	0.074	0.055	0.090
Controls				
Age Group X Year FE	Y	Y	Y	Y
Lagged Dependent Variable	Y	Y	Y	Y
P-val (40–49) - (20–29)	0.015	0.776	0.129	0.002

Note: The table above reports results from regressions of the form

$$\frac{1}{k}\Big(\log Y_{i,a',t+k} - \log Y_{i,a,t}\Big) = \alpha_0 + \alpha_t + \beta(k,a)\eta_{i,t} + \rho\,(\log Y_{i,t} - \log Y_{i,t-k}) + \epsilon_{i,t}$$

for $k = 20$ years for Census years spanning from 1850-2010, where we allow the coefficients to vary by age. The dependent variable $Y_{i,a,t}$ tracks employment by workers in age group a; for example, to compute the 20-year growth rate in employment in 1900 for workers aged 20–29 in occupation i, we compare it to the employment of 40–49 year old workers in occupation i in 1920. The main variable of interest is $\eta_{i,t}$, our technology exposure measure (normalized to unit standard deviation). Employment growth rates is in annualized percentage terms. Standard errors are clustered by occupation and corresponding t-stats are shown in parentheses. Observations are weighted by occupation employment share at time t. Census year 1870 does not show up in the first column of the 20-year subsample regressions because the 1890 Census records no longer exist.

Chapter 3

Intermediation Frictions in Equity Markets

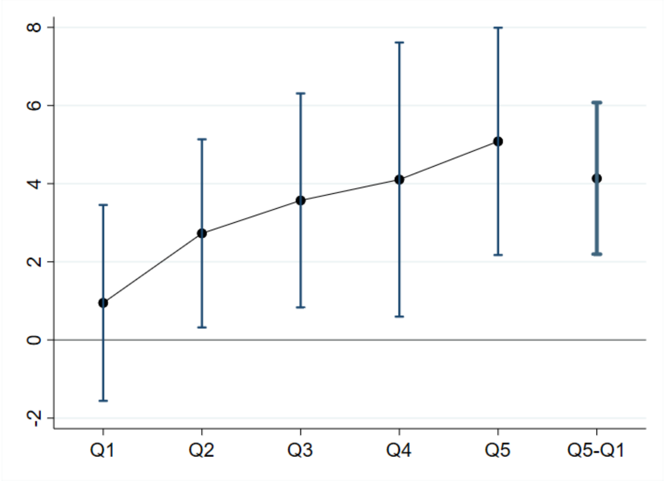
Empirical evidence has recently accumulated in favor of asset pricing models with frictions and that feature sophisticated financial intermediaries as the marginal investors. This has been particularly so in complex asset classes. At the same time, the relative importance of these sorts of theories for explaining stock price movements has been questioned, even among proponents of intermediary asset pricing models. There is cause for such skepticism, due to the comparative ease of household stock market participation relative to other more complex asset markets (for example credit default swaps, mortgage-backed securities, or even options and foreign exchange), which require far more expertise in order to participate. Despite the success of intermediary-based empirical asset pricing models–such as Adrian, Etula, and Muir (2014) and He, Kelly, and Manela (2017)–in explaining cross-sections of returns on stocks and other asset classes, such tests cannot rule out that a household-based pricing kernel holds for stocks because households also participate heavily in equity markets, both directly and indirectly, alongside financial institutions.

However, if households and institutional investors have differential preferences for direct holding of certain stocks for any reason unrelated to the true distributions of future cashflows, whether for heterogeneous beliefs or differential trading costs, dispersion in intermediation independent of fundamentals can arise naturally in the cross-section, even when households are not prevented from trading directly. This dispersion leads to similar patterns in asset covariances with shocks to intermediaries as in models where intermediaries face constraints and households are impeded from trading directly. The basic prediction is that for two otherwise similar assets, the more intermediated asset exhibits larger price response and risk premia variation due to shifts in intermediary risk-bearing capacity.

I find evidence strongly in support of this implication within the equity asset class. After accounting for firm characteristics, excess returns on stocks that are held more by some of the largest and most active institutional investors in equity markets (mutual funds, hedge funds, and other investment advisors) covary more with theoretically-motivated empirical proxies for shocks to intermediary risk tolerance. My main empirical proxy combines the two primary empirical measures of shocks to an intermediary pricing kernel proposed in the literature. Adrian, Etula, and Muir (2014) propose shocks to broker-dealer book leverage, while He, Kelly, and Manela (2017) use shocks to the market equity capital ratio of Federal Reserve primary dealer bank holding companies. I simply standardize both of these measures and take the average of the two, similarly to the approach in Haddad and Muir (2018), who argue that doing so provides a good proxy for average financial sector willingness to take risk. I also include tests with these measures separately, which provide evidence consistent with my main proxy that combines the two. I further show that another credible proxy for shocks to financial sector risk-bearing capacity–the excess return on the financial sector–displays the same empirical pattern of increased exposure to shocks to intermediary risk-bearing capacity along the dimension of increased intermediation.

Figure 3-1 illustrates some of my findings for betas on contemporaneous shocks to intermediary capital. Stocks sorted on a measure of intermediation that holds stock fundamentals constant have monotonically increasing betas on intermediary capital shocks; a portfolio formed on stocks from the top quintile of my intermediation measure has a beta of about 5.1 on intermediary capital shocks, while the beta on the lowest quintile is about 0.9. Moreover, intermediary shocks significantly explain the spread in returns between the top and bottom portfolios, with a t-stat of 4.21.

Figure 3-1: Coefficients on Intermediary Shocks Over Portfolios Formed on Intermediation Quintile



This figure shows the coefficient estimates on the average of the standardized Federal Reserve primary dealer equity capital ratio shocks from He, Kelly, and Manela (2017) and the broker-dealer book leverage growth shocks from Adrian, Etula, and Muir (2014) for five portfolios formed on the intermediation measure $\epsilon_{i,t}$ (which is constructed so as to be uncorrelated with fundamental stock characteristics; section 3.3.1 discusses this in more detail), as well as the coefficient on the intermediary shocks for the top minus bottom quintile spread. The sample is quarterly and comprises 1980q2 to 2017q3. The confidence bands represent 95% confidence intervals computed from Newey-West standard errors.

Predictive tests using state variables proposed in the literature to capture intermediary willingness to take risk also exhibit a pattern in line with the mechanisms illustrated in intermediary asset pricing models. Namely, state variables that proxy for higher (lower) willingness to take risk predict returns more negatively (positively) for similar stocks that are more intermediated. Following He, Kelly, and Manela (2017), I use the squared market leverage ratio of Federal Reserve primary dealer bank holding companies and the book leverage ratio of broker dealers obtained from the Flow of Funds accounts in predictability tests, taking the average of the standardized versions of these two state variables as my main proxy for intermediary risk appetite at the current date.[1] I also show that these measures

[1] As He, Kelly, and Manela (2017) point out, these two state variables predict stock market returns with

perform well when included separately, and another proxy motivated by theory, financial sector stock market wealth share, predicts returns more negatively for the more intermediated stocks as implied by the theory.

The predictability tests suggest that discount rates on more intermediated stocks respond more to shocks to intermediary risk-bearing capacity, a fundamental feature in models of intermediary asset pricing. Such implications are discussed in more detail in section 3.1, where I present a simple model in which intermediary risk tolerance can shift as a result of shocks to some underlying state variables. Shocks to these state variables implicitly represent shocks to the capitalization of intermediaries, which in turn cause financial constraints on intermediaries to be more binding. This basic mechanism is inspired by numerous papers from the intermediary asset pricing literature.

I include an additional test confirming a feature in the cross-section of return predictability which is consistent with theory, though it is not explicitly laid out in my simple static model. I find that the predictive coefficients for the return spread between high- and low-intermediation portfolios are positive but declining with the time horizon of the monthly returns being predicted, and the R^2 is also decreasing with the time horizon. This suggests shocks to intermediaries induce temporary distortions in relative discount rates between more and less intermediated stocks, with such distortions reverting over time as intermediary capital recovers.[2]

The proxies for intermediary risk tolerance shocks proposed by He, Kelly, and Manela (2017) and Adrian, Etula, and Muir (2014) focus on a set of levered institutions–namely dealer banks and other broker-dealers–that have been argued in the literature to occupy a place of central importance in financial markets and as marginal investors in pricing numerous asset classes; however, they are not the same set of institutions whose stock holdings I measure (though there is some overlap). The empirical evidence I present in section 3.3 implies that these shocks also affect the risk-bearing capacity of the mutual funds, hedge funds, and other

opposite sign. Hence, when I take the average I take the negative of the broker-dealer leverage ratio so that the composite measure predicts returns with a positive sign.

[2] This mechanism is outlined theoretically in Gromb and Vayanos (2018), for example.

investment advisors whose holdings are included in my analysis. Therefore in section 3.4 I suggest reasons that the risk-bearing capacity of these classes of financial institutions are interlinked.

In my primary tests I construct a measure of stock-level intermediation that holds firm size, pre-ranking CAPM beta, book-to-market, firm profitability, and investment constant, while retaining wide variation in intermediation.[3] I do this by running cross-sectional regressions of stock percentage intermediated on stock characteristics and sorting on the residuals. In the appendix I demonstrate that this cross-sectional regression specification isolates a component of intermediary holdings that is unrelated to fundamentals and along which the stock price response is monotonically increasing with shocks to intermediary risk-bearing capacity. In particular, this specification arises in a setting related to the characteristics-demand setup of Koijen and Yogo (2019) when households and financial institutions believe that expected asset cashflows and covariances are linear in characteristics, but they may disagree about first moments. I also show that my findings don't depend crucially on the set of characteristics considered, so long as stock size is accounted for. In robustness checks I show that constructing a measure of intermediation that is uncorrelated with dozens of additional stock characteristics leaves results unchanged. These findings similarly hold after just controlling for stock size in the cross-sectional regressions.

Analysis at the individual stock-level via panel regressions corroborate portfolio-level evidence–more intermediated individual stocks have increasing betas on contemporaneous capital shocks and their returns are more predictable by the current capitalization of financial intermediaries. Consider two stocks with the exact same characteristics, with one being fully intermediated and the other owned entirely by households. My estimates indicate that a one-standard deviation negative shock to intermediary risk-bearing capacity decreases the return on the fully intermediated stock by about 8-10% on an annualized basis relative to the non-intermediated stock. Meanwhile, predictive regressions imply that a one-standard deviation decrease in the time t intermediary risk tolerance increases the expected return

[3] I construct the accounting-based characteristics following Koijen and Yogo (2019), who in turn follow the steps outlined in Fama and French (2015).

on the intermediated stock by about 11 to 12% on an annualized basis relative to the stock owned entirely by households.

Theoretical support for my empirical strategy is demonstrated in a simple economic setting introduced in section 3.1. The model shows that if households are relatively more willing to hold one asset for any reason unrelated to the true distribution of cash flows, assets that are less preferred by households become more intermediated and have risk premia that respond more to shocks to the intermediaries' risk tolerance. In my setting this increased intermediary willingness to hold certain assets comes because households have either heterogeneous expectations errors or view direct investing in certain assets as relatively more or less costly. The empirical implication is that relatively more intermediated assets that are similar on fundamentals should have prices that move more with contemporaneous intermediary shocks and risk premia that are more predictable by state variables representing intermediary risk tolerance.

3.0.1 Contribution to the Literature

The literature connecting the marginal value of wealth of financial intermediaries to asset price movements has grown rapidly in the aftermath of the financial crisis of 2008-2009, which brought such theories to the forefront. Since then, theories of frictional intermediation have found empirical support in many asset classes.[4] Adrian, Etula, and Muir (2014) and He, Kelly, and Manela (2017) show using classical asset pricing tests that proxies for the marginal utility of a representative intermediary successfully price large cross-sections of portfolios spanning multiple asset classes. Such tests imply that intermediaries are marginal investors in many markets.[5] These asset pricing tests do not necessarily mean that intermediation frictions

[4] For example, markets for credit-default swaps (Siriwardane, 2018 and Mitchell and Pulvino, 2012); convertible bonds (Mitchell, Pedersen, and Pulvino, 2007); foreign exchange (Du, Tepper, and Verdelhan, 2018); life insurance (Koijen and Yogo, 2015); treasuries (Haddad and Sraer, 2018, and Anderson and Liu, 2018); and, mortgage-backed securities (Krishnamurthy, 2010 and Gabaix, Krishnamurthy, and Veron, 2007), to name just a few examples.

[5] See also Muir (2017), Chen, Joslin and Ni (2016), Adrian, Moench, and Shin (2014), Kargar (2019), and Ma (2019) for further empirical evidence on the connection between the health of financial intermediaries and asset prices.

matter for price movements in all markets, because they do not preclude the possibility that households are jointly marginal with sophisticated intermediaries in certain asset classes. They also do not rule out that intermediaries' investment decisions merely directly reflect the preferences of households on whose behalf they make their investment decisions.

Haddad and Muir (2018) address this issue by constructing empirical tests designed to detect whether intermediaries matter for asset price movements or if they merely act as a veil to pass on household preferences. Their estimates imply that intermediation frictions do matter, especially in credit default swap, foreign exchange, commodities, and sovereign bond markets. On the other hand, they argue that equities are the least likely asset class to find price movements due to intermediation frictions (though they cannot rule out their presence). Moreover, their focus is on making comparisons across broad asset class representative portfolios; my focus is on heterogeneity in responses to shocks to intermediaries within the equity asset class.

Other papers in this literature have expressed skepticism towards the relevance of these theories in explaining price movements in equity markets. While He, Kelly, and Manela (2017) find that their proxy for a representative intermediary stochastic discount factor performs reasonably well in describing cross-sections of equity returns, they also argue that equity may be the asset class least fitting to their setting and suggest that intermediaries may act as a veil that merely passes through the preferences of households in equity markets. Similarly in the theoretical literature He and Krishnamurthy (2013) think of their model in the context of complex asset markets such as mortgage backed securities as opposed to equities. A recent intermediary asset pricing paper that does focus on equity markets is Koijen and Yogo (2019), who estimate a characteristics-based demand system for heterogeneous financial intermediaries in equity markets; however, they don't attempt to test how their findings relate to friction-based intermediary asset pricing.[6]

I contribute to this literature by demonstrating that theories of frictional intermediation do appear to matter for asset price movements in equity markets. In particular, my within-

[6] I draw from the set of stock characteristics that they use to create my primary measure of stock-level intermediation (as detailed in section 3.3).

asset class findings complement the between-asset class comparisons of Haddad and Muir (2018). This paper also helps address the question posed by Cochrane (2011), which is to explain *why* certain assets covary more with particular risk factors, rather than just examining the cross-sectional asset pricing performance of factor models without any accounting for determinants of the risk exposures. This paper demonstrates that risk factor loadings are in part determined endogenously merely by the agents who own a given asset.

In a related paper, Cho (2019) finds that stocks with higher arbitrage position (determined by abnormally high/low short interest in a stock) explains betas on shocks to the Adrian, Etula, and Muir (2014) leverage factor in the post-1993 period when hedge funds became more active in equity markets. I also focus on equity markets, but consider the holdings of a much larger class of financial institutions and for multiple definitions of intermediary shocks; analyze effects both at the portfolio and individual stock level; and, include contemporaneous and predictive tests using shocks to and levels of the state-variables implied by intermediary asset pricing models.

Cho (2019) contextualizes his findings within the Kondor and Vayanos (2019) setting of minimal frictions. By contrast, I prefer the friction-based interpretation, since the empirical measures proposed by Adrian, Etula, and Muir (2014) and He, Kelly, and Manela (2017) are constructed to proxy for mechanisms described in friction-based models–Brunnermeier and Pedersen (2009) and Adrian and Boyarchenko (2012) in the case of Adrian, Etula, and Muir (2014), and He and Krishnamurthy (2013) and Brunnermeier Sannikov (2014) in the case of He, Kelly, and Manela (2017). In Adrian, Etula, and Muir (2014) the underlying friction comes from time-varying margin constraints, while in He, Kelly, and Manela (2017) the friction entails an equity capital constraint imposed by investors in the equity of the intermediary because of moral hazard problems in delegation to professional asset managers. These frictions naturally lead to time-varying intermediary risk-bearing capacity, which is the key mechanism I focus on in my model to derive the predictions that I test in the data.

Besides the primary connection with the theoretical and empirical literature in intermediary asset pricing, this paper also has connections with research areas such as limits to arbitrage[7]

[7] See for example Shleifer and Vishny (1995) and Duffie (2010)

and the effects of institutional ownership on asset prices.[8]

In section 3.2 I describe the data used and sampling criteria. Section 3.3 describes in detail my empirical strategy and presents my empirical findings. In section 3.3.1 I explain in more detail the construction of my stock-level intermediation measure and why the intermediary shocks suggested by He, Kelly, and Manela (2017) and Adrian, Etula, and Muir (2014) can be considered proxies for changes in financial risk-bearing capacity. Sections 3.3.2-3.3.4 show my main findings, including portfolio- and stock-level analysis and robustness checks. Section 3.4 features a discussion on my empirical findings, including an examination of why the shocks to levered intermediaries proposed in the literature may be directly connected to the marginal utility/risk-bearing capacity of the financial institutions (mutual funds, hedge funds, and other large investment advisors) whose asset holdings I include; finally, section 3.5 provides some brief concluding remarks.

3.1 An Economic Setting For Empirical Tests

Theories linking asset price movements to intermediary health broadly divide into equity constraint models where financial constraints bind when intermediaries' net worth is low;[9] and, another a class of models where constraints explicitly limit the amount of leverage or risk that intermediaries can take on.[10] To set the stage for my empirical tests I present a simple model that takes the middle ground between these two broad classes of intermediary asset pricing models by allowing risk-bearing capacity to vary due to underlying state variables, which could be proxies for net worth shocks or changes in leverage/margin constraints. The intended interpretation is that these shifts in willingness to take on risk come from constraints that exist due to underlying agency frictions in delegation to intermediaries, which is a unifying theme in these models. Though I present my theoretical predictions using a slightly

[8] See for example Gompers and Metrick (2001), Nagel (2005) and Basak and Pavlova (2013)

[9] See Bernanke and Gertler (1989) and Holmstrom and Tirole (1997) for early examples, and more recently Brunnermeier and Sannikov (2014) and He and Krishnamurthy (2013).

[10] Examples from this literature include Brunnermeier and Pedersen (2009), Adrian and Shin (2014), and Garleanu and Pedersen (2011).

different setting, the intuition and consequent empirical implications in this section draw from the models of He and Krishnamurthy (2018) and Haddad and Muir (2018).

There are two agents, a representative institutional investor/intermediary (labelled "I") and a sophisticated household that can access stock markets directly (labelled "H"). The intermediary invests on behalf of an unmodeled household sector that cannot (or chooses not to) invest directly in the stock market. There are N risky assets each in net supply 1 with payoffs that are jointly normally distributed

$$D \sim N(\mu, \Sigma) \tag{3.1}$$

For simplicity I assume here that Σ is diagonal.[11] Agents have constant absolute risk-aversion utility. Intermediaries have coefficient of constant absolute risk aversion $\gamma_I(\omega)$, which is a function of a state variable (or variables) ω. In the same manner households have coefficient of absolute risk aversion $\gamma_H(\zeta)$, which I allow to be a function of a state variable (or variables) ζ. There is a risk-free asset with exogenously fixed gross rate of return R_f.

I make the following assumption that generates heterogeneity in intermediation independent of asset payoffs:

Key Assumption: Households invest as if $D \sim N(\mu + \lambda, \Sigma)$ for some vector λ.

The interpretation of λ is to reflect potentially heterogeneous expectations errors across stocks by households, or more broadly that households could have preferences for holding certain stocks for reasons unrelated to cashflow distributions (i.e. differences in perceived costliness of holding certain stocks). In more general terms, the presence of λ captures a feature that is present in many intermediary asset pricing models, which is that households' expertise in direct investing is limited in some way relative to intermediaries'.

Both agents maximize the expected utility of period 1 wealth. Given CARA utility, the total wealth invested in risky assets is independent of initial wealth, and the normality of

[11] I relax this assumption in a characteristics-based extension on the model in Appendix 3.6.

returns yields the familiar mean-variance criterion for portfolio choice for agent j:

$$\theta_j = \frac{1}{\gamma_j}\Sigma^{-1}(\mu_j - R_f P) \tag{3.2}$$

where γ_j is agent j's absolute risk aversion and μ_j is agent j's beliefs about expected cashflows.

The market clearing condition is

$$\mathbf{1} = \theta_I + \theta_H \tag{3.3}$$

Now denote by $\rho_j \equiv \frac{1}{\gamma_j}$, the risk-tolerance of agent j. Plugging in (3.2) for $j = I, H$ and solving for P gives

$$P = \frac{\rho_I(\omega)\mu + \rho_H(\zeta)(\mu + \lambda) - \Sigma\mathbf{1}}{\left(\rho_I(\omega) + \rho_H(\zeta)\right) R_f} \tag{3.4}$$

The percent held by intermediaries can be expressed as

$$\theta_I = \rho_I(\omega)\Sigma^{-1}\left[\frac{-\lambda\rho_H(\zeta) + \Sigma\mathbf{1}}{\rho_I(\omega) + \rho_H(\zeta)}\right] \tag{3.5}$$

Therefore the percent intermediated θ_I is strictly decreasing in λ (and vice versa). Consider a local shock to P by taking the total derivative:

$$\mathrm{d}P = \frac{\rho_I^{'}(\omega)(\Sigma\mathbf{1} - \lambda\rho_H(\zeta))}{R_f(\rho_I(\omega) + \rho_H(\zeta))^2}\mathrm{d}\omega + \frac{\rho_H^{'}(\zeta)(\Sigma\mathbf{1} + \lambda\rho_I(\omega))}{R_f(\rho_I(\omega) + \rho_H(\zeta))^2}\mathrm{d}\zeta \tag{3.6}$$

$$\equiv \beta_\omega d\omega + \beta_\zeta d\zeta \tag{3.7}$$

This equation leads to one of the key implications of the model:

Proposition 1 *Suppose $\rho_I^{'}(\omega) > 0$. The component of the total derivative dP due to changes in ω, β_ω, is strictly increasing in θ_I (percent intermediated).*

Proof: Follows immediately from the positivity of $\rho_I^{'}(\omega)$, the fact that λ is strictly decreasing in θ_I, and that the first term on the right-hand side of (3.6) is strictly decreasing in λ.

Note that (3.6) resembles a regression of the local change in stock price ("stock return" for a CARA investor) on shocks to ω and ζ. In other words, proposition 1 implies that the beta on a shock that increases (decreases) the intermediaries' risk tolerance is increasing (decreasing) in the percent intermediated. This is emphasized by He and Krishnamurthy (2018), and is the first theoretical implication that I test in the data. Equation (3.6) also underscores an issue highlighted by Haddad and Muir (2018), which is the potentially confounding effect of shocks that change the risk tolerance of households (if $\rho^{'}(\zeta) \neq 0$ and shocks to ω and ζ are correlated). As such, in my empirical implementation I include shocks that could potentially proxy for changes in household-level risk aversion.

Observe also the loading on the shock to household risk tolerance:

$$\frac{\rho_H^{'}(\zeta)(\Sigma\mathbf{1} + \lambda\rho_I(\omega))}{R_f(\rho_I(\omega) + \rho_H(\zeta))^2}\mathrm{d}\zeta \tag{3.8}$$

If $\rho_H^{'} > 0$, then this is increasing in λ and hence is decreasing in percent intermediated. The rate of decrease depends upon the slope $\rho_H^{'}$. In my empirical tests I find that betas on non-intermediary risk factors are relatively flat in the dimension of increased intermediation, implying that the slope $\rho_H^{'}$ is also relatively flat.

Prices in (3.4) are increasing in λ. Returning to the example of two similar stocks with $\lambda_1 < \lambda_2$; the price P_2 is higher than P_1, and this difference is decreasing in $\rho(\omega)$. The implications of this fact are summarized in the following proposition:

Proposition 2 *Consider two assets such that $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ and $\lambda_1 < \lambda_2$. Let $E[R_{p,i}] = \mu_i - R_f P_i$ denote the risk premium on asset i. Then the difference in the risk premium on asset 1 and asset 2 decreases with ω, i.e. $\partial\left(E[R_{p,1} - R_{p,2}]\right)/\partial\omega < 0$.*

Proposition (2) states that for two similar stocks, state variables proxying for higher (lower) intermediary risk tolerance predict returns more negatively (positively) for the stock that is more intermediated. In order to test Proposition (2) empirically, I regress the excess returns of high minus low intermediation stock portfolios on predetermined proxies for intermediary risk-bearing capacity. Similarly for Proposition (1), I regress stock returns on portfolios sorted on quintiles of my intermediation measure, as well as the high minus low intermediation portfolio excess returns, on contemporaneous shocks to risk-bearing capacity. In section 3.3.1

I describe in detail the construction of my intermediation measure, which is constructed so as to hold stock fundamentals constant while isolating the effects of increasing intermediation (lower λ). As is discussed briefly in section 3.3.1 and in more detail in Appendix 3.6, this exact empirical specification arises in a setting closely related to Koijen and Yogo (2019), where I assume that the representative household and intermediary's assessments of fundamental asset means and covariances are linear in characteristics.

Appendix 3.7 provides a minor extension on the model that examines the empirical implications for the case when household risk tolerance also responds in the same direction to shocks to the intermediary state variable(s) ω as intermediary risk tolerance. I show that in this case the presence of an increasing price response to intermediary shocks must come through the intermediary risk aversion channel and not through the shocks to household risk aversion, which actually works against finding an effect. I present this as a third proposition:

Proposition 3 *Suppose that household risk tolerance is also a function of the same state variable(s) ω as intermediary risk tolerance and the partial derivative of $\rho_H(\omega, \zeta)$ with respect to ω is positive. Then, holding all else constant, the presence of increasing price responses to ω shocks for more intermediated assets must be driven by shocks to intermediary risk tolerance and not by shocks to household risk tolerance.*

Proof: See Appendix 3.7.

This logic also extends to the setting where $\rho_H(\zeta)$ does not depend directly on ω as in equation (3.6). If shocks to ζ and ω are positively correlated and $\rho_H^{'} > 0$, the exclusion of shocks to ζ actually work *against* finding an effect, because the coefficient on dζ is decreasing in percent intermediated while the coefficient on dω is increasing. As Haddad and Muir (2018) point out, it is likely that financial institutions' risk tolerance shocks are positively correlated with those of households, so this seems to be the relevant case empirically.

Observe that the model implies the spread in betas are due to discount rate effects: price appreciation in a more intermediated stock occurs due to positive shocks to intermediaries' willingness to take risk, absent any fundamental information about stock cashflows. Though discount rate and cashflow components of returns cannot be observed perfectly, the combined presence of return predictability using pre-determined state variables and price movements induced by contemporaneous shocks to the same state variables would constitute strong evidence that the effects are driven through the discount rate component of returns. Because of this I include both contemporaneous and predictive tests of the model's implications.

The choice to have absolute risk aversion vary as a function of underlying state variables is obviously critical to the model's predictions and deserves further attention. Since the coefficient of relative risk aversion is related to the coefficient of absolute risk aversion by $w_I \gamma_I = \alpha_I$ (where w_I is the agent's wealth, γ_I is the absolute risk aversion, and α_I the relative risk aversion) allowing γ_I to vary as a function of wealth or wealth share captures effects resembling the wealth effects present in intermediary asset pricing models with constant relative risk aversion of specialists. He and Krishnamurthy (2013) is one such example. In this model wealth shocks lead to changes in risk premia, as the distribution of wealth shifts between agents with different willingness or ability to bear risk. These effects have outsize influence in the constrained region of the model, when equity capital constraints bind and intermediaries require price concessions in order to bear aggregate risk.

The presence of risk aversion is not required for intermediaries to exhibit time-varying risk-bearing capacity so long as there are binding constraints. Brunnermeier and Sannikov (2014) work with risk-neutral agents and find that specialists' wealth share is a critical state variable, generating large spikes in risk premia in the constrained region just as in He and Krishnamurthy (2013). Adrian, Etula, and Muir (2014) point out that in a setting resembling Brunnermeier and Pedersen (2009) with margin constraints, time variation in the margin constraint can lead to non-trivial state pricing where risk-neutral intermediaries value a dollar of wealth relatively more when the Lagrange multiplier on the margin constraint is higher and the value of relaxing the constraint is larger. When margin constraints are tighter, intermediaries invest as if they were more risk averse. Adrian, Etula, and Muir (2014) argue that their leverage measure (which is the reciprocal of margin) proxies for the tightness of leverage constraints and hence risk-bearing capacity. In this sense, having risk-tolerance shift due to intermediary shocks is a sort of reduced-form way of capturing the price effects of such mechanisms. Furthermore, allowing households' absolute risk aversion to vary as a function of state variables can capture features related to time-variation in household risk aversion, as would be found in a habit model, for example.

In summary, the crux of the model's predictions are this: if (1) intermediary risk tolerance is time-varying and we have suitable proxies for this time variance; (2) households make expectations errors (or have direct investment costs) that are different across stocks; and, (3) there is variation in households' expectations errors (or direct investment costs) across otherwise similar assets; then we should be able to detect the effects detailed in propositions 1 and 2. The justification behind point (1) comes from the literature on friction-based intermediary asset pricing models. Point (2) can be seen as resulting from limited rationality/information processing capacity of households relative to more sophisticated institutional investors; similar features are present in numerous asset pricing models. I argue point (3) by demonstrating in section 3.3.1 that I can construct a measure that holds fundamental stock information constant yet still generates a large spread in average intermediation.

3.2 Data Sources And Sample Construction

Before proceeding to the empirical implementation I first describe the datasets used and sampling procedure followed. Individual monthly firm stock returns are from CRSP. The sample is restricted to ordinary common shares (share codes 10 or 11) and that trade on the NYSE, Amex, or Nasdaq (exchange codes 1, 2, or 3). Institutional holdings data for individual stocks come from the Thomson-Reuters Institutional Holdings Database (S34 file). Due to well-documented errors in the S34 database institutional type classifications, I use the corrected type codes that are provided by Koijen and Yogo (2019) to classify institutions into mutual funds and other investment advisors (which category prominently includes the largest hedge funds).[12] I download the quarterly holdings data from 1980q1 to 2017q2. My primary set of stock characteristics are originally derived from Compustat, but are taken directly

[12] For a more detailed description of this data see Gompers and Metrick (2001), or more recently Koijen and Yogo (2019)

from Koijen and Yogo (2019), whose paper on characteristics-based demand of financial institutions also utilizes the Thomson-Reuters database. The characteristics are derived from the Fama-French 5-factor model, and include past 5-year stock CAPM beta, log book equity as a proxy for size, gross profitability, and asset growth. I further include the book-to-market ratio as the ratio of book equity to market cap from a year previous. As in Koijen and Yogo (2019), accounting characteristics are obtained as of at least 6 months and no more than 24 months prior to the given date in order to ensure the data are publically available at the time of portfolio formation.

Besides the Koijen and Yogo (2019) characteristics, in a robustness check I add to the set of stock characteristics dozens of financial ratios obtained from the Wharton Research Data Services financial ratios suite. I also obtain the quarterly and monthly series of shocks to Federal Reserve primary dealer capital introduced in He, Kelly, and Manela (2017) and available on Asaf Manela's website. As an additional intermediary variable I obtain the leverage of broker dealers introduced in Adrian, Etula, and Muir (2014).[13] The monthly Fama-French risk factors plus momentum factor are also downloaded from Ken French's website.

The sample construction proceeds as follows. Each quarter I take the intersection of the entire CRSP universe of stocks meeting share code and exchange code criteria described above with the Koijen and Yogo (2019) stock characteristics data, excluding any missing matches within a quarter. As done by many previous studies, I further exclude microcap stocks from the sample each quarter (defined to be stocks beneath the NYSE 20th percentile in market cap) and stocks with price less than $5 in order to focus on the set of stocks where large financial institutions are able to trade most freely. As is common practice, I additionally exclude financial stocks (stocks with SIC code between 6000 and 6999). This restriction is even more practical in my setting because the relationship between stock price movements and intermediary risk-bearing capacity is highly endogenous for financial stocks. In terms of market cap, these restrictions drop a small portion of the CRSP equity universe–my

[13] Thanks to Koijen and Yogo (2019), He, Kelly, and Manela (2017), and Adrian, Etula, and Muir (2014) for making their data readily available.

sampling retains on average about 97% of total market capitalization of non-financial stocks on the CRSP tape. These stocks constitute the primary quarterly sample. I also convert the monthly Fama-French five factors and momentum to their respective quarterly versions. Unless otherwise noted, the sample period for regressions spans 1980q2 to 2017q3.

3.3 Empirical Strategy and Results

3.3.1 Constructing Measure of Intermediation

The characteristics-based extension on the model discussed in Appendix 3.6 suggests that stocks with similar characteristics but higher intermediary holdings should have higher betas on intermediary capital shocks. As such, I construct a measure of intermediation intended to be unrelated to key stock characteristics that proxy for information regarding cashflow distributions. Let $X_{i,t}$ be a vector of stock characteristics that are informative about the distribution of time $t+1$ cashflows of asset i. At each time t I run the following cross-sectional regression:

$$\text{Percent Intermediated}_{i,t} = \alpha_t + \beta_t X_{i,t} + \epsilon_{i,t} \tag{3.9}$$

In Appendix 3.6 I illustrate that the exact specification in (3.9) arises under a framework that expands upon my setting in section 3.1 and includes features very similar to the characteristics-based demand setting of Koijen and Yogo (2019). In particular, I show that if investors believe that the payoff covariance matrix can be decomposed into fundamental risk factor loadings that are linear in characteristics and the average asset payoffs are also linear in the characteristics, then (3.9) arises when households and institutional investors agree on the covariance matrix but households have some residual demand unrelated to the characteristics due to their different assessment of the first moment of asset payoffs. Under these assumptions, $\epsilon_{i,t}$ identifies a component of intermediary holdings that are due to variation in λ in the model and are uncorrelated with characteristics that provide information on the moments of asset cashflows. More broadly this approach is done to ensure that the cross-sectional spread in asset price response to intermediary risk-bearing capacity shocks is not driven primarily by differential fundamental exposures to other risk factors.

If the risk tolerance of the financial institutions who are active in equity markets is time-varying and moves due to changes in empirical proxies of financial intermediary capital, sorting on $\epsilon_{i,t}$ should induce variation in betas on shocks to intermediary capital, and current intermediary capital should contain information about the expected returns of high $\epsilon_{i,t}$ assets relative to low $\epsilon_{i,t}$ assets. Specifically, high $\epsilon_{i,t}$ assets should have larger contemporaneous price response due to shocks to proxies for intermediary capitalization/intermediary risk tolerance and greater return predictability by the level of intermediary capital, as outlined in propositions 1 and 2.

Here Percent Intermediated$_{i,t}$ denotes the percentage of shares held by mutual funds, hedge funds, and other investment advisors. I focus on these institution types because they include the set of financial intermediaries that are the largest and most active in equity markets, though results are unaffected by additionally including any or all other 13F institutional investor types. One may also consider taking net positions by subtracting out aggregate short interest from Percent Intermediated$_{i,t}$. Average short interest on most stocks is small enough that this does not change my findings in any meaningful way, so I focus on just the long positions as presented on the 13F reports.

Regression equation (3.9) decomposes intermediary holdings into three components: $\beta X_{i,t}$, holdings due to firm fundamentals; α_t, holdings due to rise in average intermediation over time;[14] and, $\epsilon_{i,t}$, holdings unrelated to fundamentals (possibly reflecting households' unobserved expectations errors or perceived direct holding costs). In the empirical tests to follow I show in a variety of settings that $\epsilon_{i,t}$ is strongly related to intermediary capital betas on both predictive and contemporaneous variables.

Implementing (3.9) requires that $X_{i,t}$ be strongly related to asset fundamentals that are informative about the distribution of future cashflows. Following Koijen and Yogo (2019), I focus on a set of stock characteristics derived from the Fama and French (2015) empirical asset pricing model that is known to have significant explanatory power for the cross-section of stock returns, and hence presumably provides considerable fundamental information regarding

[14] This has been well-documented. See Stambaugh (2014), for example.

asset cash flows.[15] The empirical implementation of (3.9) includes the following set of stock characteristics in $X_{i,t}$: a second degree polynomial in log book equity; gross profitability to book equity; annual growth in firm assets (as a proxy for investment); book-to-market ratio using one-year lagged market cap; and, 5-year rolling monthly pre-ranking CAPM beta (requiring at least 24 observations to be included). These are derived from the sorting characteristics used to construct the risk factors in the Fama and French (2015) model. I use these characteristics as constructed by Koijen and Yogo (2019), who in turn construct them from Compustat so as to align with the procedure in Fama and French (2015).

In robustness checks I demonstrate that the set of characteristics included in (3.9) is not particularly important, so long as you control for stock size. My proxy for stock size is log book equity rather than market equity because market equity is a more endogenous equilibrium outcome that is affected by intermediary and household demand for the asset (in unreported regressions I find that findings maintain when using market equity instead of book equity for my size proxy). My model presented in section 3.1 demonstrates why it is important to control for stock size. The empirical tests described in propositions 1 and 2 require holding means and variances of underlying cashflows constant. Since I normalize net supply to one, the price of a given asset has the interpretation of the stock market cap and the means and variances are the means and variances of cashflows for owning the entire set of shares for a given stock. Hence μ and Σ are highly dependent on the stock size.

In Table 3.1 I estimate the Fama-Macbeth time series average coefficients from each cross-sectional regression (3.9). By far the strongest predictor of intermediary holdings is stock size, as proxied by log book equity and log book equity squared, although each of the other characteristics is statistically significant in explaining institutional holdings. These institutions tend to overweight large stocks, profitable stocks, and stocks with high asset growth and CAPM betas, and tend to underweight value stocks. Note also that the average

[15] Hou, Xue, and Zhang (2015) argue that their empirical asset pricing model, which is closely related to the model Fama and French (2015), performs particularly well in describing the cross-section of returns when micro-cap stocks are not over-weighted in portfolio formation. Thus (3.9) is likely to be more relevant among the set of the larger, more liquid stocks (non-micro cap stocks and stocks with share price above $5) that I consider in my analysis.

cross-sectional R^2 is only around .11, which still leaves a substantial portion of intermediary holdings unexplained each period. I use this unexplained portion to proxy for variation in the parameter λ (and hence θ_I, percent intermediated) from the model that is unrelated to stock fundamentals.

3.3.2 Empirical Results For Portfolios Sorted on Intermediation

Though theoretical propositions in section 3.1 required holding stock fundamentals constant, as a practical matter the empirical predictions in propositions 1 and 2 still hold as long as two assets look similar enough but have a wide spread in λ (and hence in percent intermediated). In terms of model parameters, if $\lambda_1 << \lambda_2$, $\mu_1 \approx \mu_2$ and $\sigma_1^2 \approx \sigma_2^2$, then the empirical implications of propositions 1 and 2 should still hold; namely, that asset 1 is much more intermediated than asset 2 and should have a higher beta on shocks to intermediary risk-bearing capacity and should be more predictable by state variables capturing intermediary risk tolerance.

I organize around this idea in two ways: first, by forming equal-weighted portfolios on the quintiles of the residual institutional holding measure $\epsilon_{i,t}$ and then by running stock-level panel regressions interacting intermediary shocks with $\epsilon_{i,t}$. The portfolios are rebalanced quarterly. Figure 3-3 shows that the portfolio formation does quite well in holding characteristics constant while inducing variation in intermediation–the average institutional holdings quintile is just below five at each point in time for the top $\epsilon_{i,t}$ quintile portfolio, while it is just above one for the bottom $\epsilon_{i,t}$ quintile portfolio. Meanwhile the average quintile of the rest of the characteristics all hover around three for both portfolios. Thus these portfolios look almost exactly the same on key stock characteristics that form the basis of the Fama-French (2015) asset pricing model, which is known to describe well the cross-section of stock returns.

Table 3.2 shows the means and medians of stock characteristics for each of the five portfolios formed on quintiles of $\epsilon_{i,t}$, including percent holdings by mutual funds, hedge funds, and investment advisors; log of market and book equity; book-to-market ratio; asset growth; profitability/book equity; and pre-ranking CAPM beta estimated over the past 60 months (and a minimum of 24 months). In line with the graphical evidence in Figure 3-3, the means and medians of each characteristic besides percent intermediated are extremely close for the top- and bottom-quintile portfolios formed on $\epsilon_{i,t}$, and are also fairly close for the middle three portfolios (though they tend to have slightly higher profitability and book/market). Meanwhile there is a large spread in average percent intermediated between the top and bottom quintile portfolios, with 62% intermediated at the top and only 19% intermediated at the bottom. Thus Table 3.2 provides further confirmation that sorting on $\epsilon_{i,t}$ isolates variation in holdings by financial institutions while holding other stock fundamentals more or less constant, particularly when comparing the top and bottom quintile portfolios.

The top and bottom $\epsilon_{i,t}$ portfolios also have a high degree of comovement, as can be seen graphically in Figure 3-2. The correlation in excess returns on the two portfolios is 0.96. Table 3.3 shows the means, standard deviations, and Sharpe Ratios of the five portfolios formed on $\epsilon_{i,t}$. Focusing on the top and bottom quintiles, the annualized excess return standard deviations of 42.33% and 38.97% of the top and bottom quintile portfolios are also close to one another, as well as their Sharpe ratios, which are respectively 0.25 for the top quintile and 0.22 for the bottom quintile. As implied by the model (though not a highlighted feature), the top quintile portfolio has higher returns, though the spread is not very large at 1.868 percent per year and carries a t-stat of 1.880 that is marginally significant at the 10% threshold.

The model implies that the portfolios should have monotonically increasing exposures to shocks to intermediary risk-bearing capacity. I use four proxies for this (from here I abbreviate the references to He, Kelly, and Manela (2017) and Adrian, Etula, and Muir (2014) as HKM and AEM, respectively): shocks to primary dealer market equity capital ratio from HKM; broker dealer book leverage shocks from AEM; value-weighted excess returns on the financial sector (stocks with SIC codes between 6000 and 6999); and, my primary proxy, which standardizes the HKM and AEM measures individually, takes the average of the two, and then standardizes this average to zero mean and unit variance.

The justification for combining the AEM and HKM shocks is to take a weighted average of financial sector risk-bearing capacity using the most prominent proxies proposed in the literature, analogous to Haddad and Muir (2018). Moreover, both Kargar (2019) and Ma (2019) demonstrate that a heterogeneous intermediary SDF can be constructed as a function of shocks to two state variables that are closely related to the AEM and HKM measures. Such an SDF arises when different classes of intermediaries face heterogeneous financial constraints or have different risk aversion, and yields leverage patterns that are simultaneously consistent with the findings of both HKM and AEM.[16] I further include returns on the financial sector, as this measure is directly related to the wealth share shocks that are important in equity-constraint based intermediary asset pricing models.

I test proposition 1 by running regressions of the form

$$R^e_{i,t+1} = \alpha_i + \beta_{1,i} F_{t+1} + \beta_{2,i}(\mathrm{Mkt}^{NonFin}_{t+1} - R_{f,t}) + \nu_{i,t} \tag{3.10}$$

individually for F_{t+1} = Intermediary Shock_{t+1}, Capital Shock_{t+1}, Leverage Shock_{t+1}, and Ex Ret. $(\text{Fin})_{t+1}$ and also for i equal to the excess returns over the risk free rate for the five intermediation ($\epsilon_{i,t}$) quintile portfolios and the high minus low $\epsilon_{i,t}$ spread portfolio. Here Intermediary Shock_{t+1} refers to the combined AEM and HKM measure; Capital Shock_{t+1} represents the HKM shock; Leverage Shock_{t+1}, the AEM shock; and finally, Ex Ret. $(\text{Fin})_{t+1}$, the excess return on the financial sector. I control for a version of the value-weighted market risk factor that includes just the returns to nonfinancial stocks. I include this control for several reasons. First, the AEM and HKM models present asset pricing tests controlling for market risk. Second; as illustrated in equation (3.6), it's important to control for shocks that could proxy for changes in the risk aversion of households, and market returns relate to time-variation in risk-aversion for certain classes of models, such as in a habit model. The joint inclusion of shocks to intermediary risk-bearing capacity and non-financial stocks also directly relates asset price movements to financial and non-financial wealth share shocks. The non-financial market risk factor has a correlation of 0.99 with the value-weighted market risk factor from Ken French's website.

[16] HKM find that bank holding companies have countercyclical leverage, while AEM finds that broker-dealers have procyclical leverage.

Table 3.4 shows the results of the contemporaneous portfolio tests using the combined Intermediary Shock$_{t+1}$ measure. The same findings are illustrated graphically in Figure 3-4. Strikingly, there is a strong monotonically increasing relationship in the betas on the intermediary shock and no pattern whatsoever in the non-financial market return betas. The t-stat of 4.45 on the quintile 5 minus quintile 1 intermediation spread portfolio is highly significant. This monotonic pattern is directly in line with the theoretical implications presented in section 3.1 and also with He and Krishnamurthy (2018), who show that shocks to intermediary risk tolerance for similar but more intermediated assets should have relatively higher betas on intermediary risk tolerance shocks than on household wealth shocks. Since the intermediary shock is scaled to unit standard deviation and returns are in annualized percent form, the coefficient of 4.13 in column (6) of Table 3.4 means that the return on the high intermediation portfolio increases by 4.13% relative to the low intermediation portfolio on an annualized basis in response to a one-standard deviation intermediary shock.

The empirical patterns illustrated in Figure 3-4 continue to hold when examining each individual proposed intermediary shock. Figure 3-5 demonstrates this. Loadings are increasing from bottom to top quintile and the top minus bottom quintile spread has a significant loading for each of the four intermediation risk-bearing capacity shocks. The exposures increase monotonically for all measures. Note also that combining the information in AEM leverage shocks and HKM capital shocks leads to a more significant coefficient on the top minus bottom quintile spread.

As a final piece of evidence for the contemporaneous portfolio regressions, in Figure 6 I run regressions separating the HKM capital shocks and AEM leverage shocks within the same specification:

$$R^e_{i,t+1} = \alpha_i + \beta_{1,i} F_{t+1} + \beta_{2,i}(\text{Mkt}^{NonFin}_{t+1} - R_{f,t}) + \nu_{i,t} \tag{3.11}$$

The high minus low intermediation excess return is significantly positive for both risk factors, with monotonicity in betas for both the capital shocks and the leverage shocks. Thus the HKM capital and AEM leverage factors continue to display patterns in line with proposition

1 when included together in the same regression.

I next turn to my predictability tests that relate to proposition 2, which is that otherwise similar but more intermediated stocks should have excess returns that are more negatively (positively) predictable by state variables that represent higher (lower) risk-bearing capacity of financial institutions. To do this, I run regressions of overlapping quarterly high minus low $\epsilon_{i,t}$ excess returns at the monthly frequency on a set of intermediary state variables. Regressions take the following form:

$$R^{Q5}_{t\to t+3} - R^{Q1}_{t\to t+3} = \alpha + \beta_1 X_t + \beta_2 Z_t + \nu_t \tag{3.12}$$

Here X_t represents any of my proxies for state variables related to time t risk tolerance of intermediaries and Z_t is a set of control predictors. Following HKM, I use the squared market leverage of Federal Reserve primary dealers as a predictor. HKM show that the conditional risk premium is a nonlinear function of the underlying capital ratio state variable, which is proportional to $(1/\text{capital ratio})^2$ in a simplified version of the He and Krishnamurthy (2013) model. Since this variable relates to lower risk tolerance of intermediaries, it should predict returns with positive sign. In predictability tests HKM also point out that the theory of AEM implies that the broker-dealer leverage ratio should predict returns negatively, so I use this as a second state variable. In line with the contemporaneous regressions, my primary predictor is a combined state variable that takes the average of the standardized squared primary dealer leverage and the negative of the standardized broker dealer leverage. I label this combined state variable η. I also use the squared primary dealer leverage and broker dealer leverage individually, as well as the share of stock market wealth held by financial stocks as my final proxy for X_t in equation (3.12) above. To be in agreement with theory, the financial sector wealth variable should predict high minus low intermediation portfolio excess returns with negative sign, as a high wealth share state corresponds with higher risk tolerance.

I also control for several return predictors from the literature. I obtain the cyclically-adjusted price to earnings ratio from Robert Shiller's website and the consumption-wealth ratio ("cay") from Lettau and Ludvigson (2001) from Sydney Ludvigson's website. I also add the 10-year minus 3-month t-bill rate term spread as a control in the predictive regressions and the investor sentiment measure from Baker and Wurgler (2006) obtained from Jeffery Wurgler's website. I include cay and sentiment as potential proxies for household willingness to take risk. The cyclically-adjusted price/earnings ratio are included because of their common use as leading indicators of aggregate macroeconomic conditions and as return predictors. Because the broker-dealer leverage ratio and the consumption/wealth ratio are available quarterly I hold them constant within a quarter for each month in the sample, but I use the monthly versions for the the rest of the predictors. Due to autocorrelation induced by overlapping observations I compute standard errors using the method of Newey-West with 4 lags.[17] Table 3.5 shows the regressions for the combined state variable (labeled "η"), while Tables 3.6 and 3.7 present the same results using the HKM/AEM predictors separately and the financial sector wealth share, respectively.

The predictability tests support proposition 2–in each case the proxy for intermediary risk tolerance has the appropriate sign and is statistically significant in predicting the quarterly returns on the high minus low intermediary ownership (high minus low $\epsilon_{i,t}$) portfolio excess returns, with t-stats hovering just above or below 3 depending on the specification. The R^2 of η in predicting the quarterly high-minus low intermediation portfolio return is 4%. The inclusion of the other predictors actually enhances the power of η to predict the spreads, as the highest t-stat on η attains in the last column where all predictors are included.

Note also that the other predictors all enter with statistically insignificant sign, except for the cyclically adjusted price to earnings ratio ("P/E") in the last column, which is significant at the 10% level with positive sign. As will be demonstrated in the next section, this positive sign disagrees with the negative coefficient on this variable in the stock-level analysis. The only predictors that demonstrate consistently strong predictive performance for the high minus low intermediation spread return across all specifications are those related to the health

[17] The standard errors on the coefficients of interest tend to decrease when including more lags than this, so I choose 4 as the lag length to be conservative, while still correcting for the autocorrelation from overlapping observations.

of financial intermediaries.

Since excess returns are expressed in annualized percentage terms in these regressions and predictors are standardized, the coefficients in row 1 of Table 3.5 imply that a one-standard deviation increase in η (i.e. a decrease in intermediary risk tolerance) translates into a roughly 3-6% increase in expected returns going forward for the top intermediated quintile portfolio over the bottom quintile portfolio. Similar magnitudes are estimated in Tables 3.6 and 3.7. Table 3.6 demonstrates that the predictability of the high minus low intermediation portfolio maintains when separating the HKM/AEM predictors, with the two having the appropriate positive and negative signs, respectively. Meanwhile Table 3.7 shows that the financial stock market wealth share significantly predicts the spread with negative sign in all specifications, consistent with theoretical models where intermediaries are less constrained when their wealth share is high.

As discussed in section 3.1, the combined presence of greater return return predictability and outsize price movements to contemporaneous shocks is important for empirically testing the theory. The loadings on shocks should come because of movements in discount rates; since the price response to both the pre-determined level of and the growth rate in the state variables is larger in the more intermediated portfolio, this supports the discount rate channel as the driving force behind these patterns.

I include a final test to examine the presence of a theoretical mechanism in the cross-section of return predictability outlined in Gromb and Vayanos (2018). In their model when the capital of constrained arbitrageurs depletes, the expected returns increase relatively more on the assets where arbitrageurs take larger positions. This causes the increased spread to self-correct over time as intermediary capital recovers due to the increased expected returns on their positions. To test for such effects, I run regressions of the form

$$R^{Q5}_{t+k} - R^{Q1}_{t+k} = \alpha + \beta_k \eta_t + \nu_t \tag{3.13}$$

where t is at the monthly horizon and k varies from 1-month ahead to 18-months ahead. Figure 3-7 plots $\widehat{\beta_k}$ and its 90% confidence interval as well as the R^2 for $k = 1, ..., 18$. As implied by the theory, the coefficients $\widehat{\beta_k}$ decrease with k, as does the R^2. Thus the quarterly horizon used in the previous predictability tests from this section features much of the overall high minus low intermediation portfolio spread return predictability of η_t. This is consistent with temporary relative asset price distortions that are corrected over time as constraints on intermediaries relax when capitalization improves, in line with Gromb and Vayanos (2018), and more broadly with models where intermediary capital moves slowly because of constraints that become more binding when intermediaries are poorly capitalized.[18]

3.3.3 Stock-Level Panel Regressions

This section demonstrates that the portfolio-level evidence from the previous section extends to the individual stock level. My stock level empirical tests take the following form for the contemporaneous regressions:

$$R_{i,t+1} - R_{f,t} = \alpha_0 + \beta_1 F_{t+1} \times \epsilon_{i,t} + \beta_2 W_{t+1} \times \epsilon_{i,t} + \alpha_t + \alpha_i + \nu_{i,t+1} \tag{3.14}$$

Here F_{t+1} is any of the contemporaneous shocks to intermediary risk tolerance. Finding $\beta_1 > 0$ implies that betas on shocks to financial institutions increase with the component of intermediary holdings that is uncorrelated with characteristics of the stock. Thus for the contemporaneous shocks used in the previous section $\beta_1 > 0$ is in line with the theory. I control for value-weighted non-financial market excess returns and also add specifications that include the Fama-French (2015) factors plus the momentum factor for W_{t+1} in (3.14). I also add time fixed-effects to control for common shocks to the cross-section as well as including stock fixed effects. Replacing the time fixed effects with uninteracted risk factors yields estimates that are essentially identical.

Once again agreeing with the theory, Table 3.8 shows that $\beta_1 > 0$ for all intermediary shocks considered and is strongly significant for all specifications except in the case of the AEM

[18] See for example Duffie (2010) for a theoretical summary and Mitchell, Pedersen, and Pulvino (2007) for early empirical evidence in convertible bond markets, or more recently Siriwardane (2019) in credit default swap markets.

leverage shocks, which show consistent positive sign but have p-values that are significant at just the 10% level for each of the specifications. However, note in the first row of Table 3.8 that the intermediary shocks which combines the information embedded in the AEM and HKM factors yields a much stronger estimate than just including the HKM or AEM factors alone. The financial sector excess return provides more evidence in agreement with the theory, as it also has positive and significant coefficient on the residual intermediation interaction term across specifications.

The economic magnitude of these estimates are fairly large. Consider two stocks with the exact same characteristics, except one is entirely owned by mutual funds/hedge funds/investment advisors, and the other is owned entirely by households. Looking at the coefficients in the first row of Table 3.8, the returns to the fully intermediated stock increase by 8-10% per year relative to the unintermediated stock on an annualized basis in response to a one-standard deviation shock to the HKM/AEM averaged intermediary factor. Point estimates on these coefficients are also quite precise, with t-stats ranging from 4.6 to 5.3.[19]

The only included non-intermediary risk factor whose betas significantly increase with $\epsilon_{i,t}$ is the Fama-French robust minus weak profitability factor. This feature is also present in portfolio regressions where I control for the Fama-French (2015) factors plus momentum in the section 3.3.4, though I don't explicitly report the coefficient estimates in those specifications. The reason for this increased exposure to the profitability factor is not immediately clear, though I find in unreported regressions that the addition of the profitability factor increases the magnitude and significance of the coefficients on the intermediary risk factors. It's also important to note that my model does not preclude the possibility that other risk factors have increasing exposure across level of intermediation Still, this case is interesting because I have already averaged out stock-level profitability. It is possible that intermediary marginal utility loads more on the profitability factor relative to households, though such an investigation is out of the scope of this paper.[20]

[19] Standard errors are clustered by date to account for cross-sectional correlation in the residuals and are also adjusted for one lag of autocorrelation.

[20] In a previous working version of the paper, Cho (2019) argues that the capital of arbitrageurs such as hedge funds loads positively on the RMW profitability factor.

Next I perform stock-level predictive regressions, which are specified as

$$R_{i,t+1} - R_{f,t} = \alpha_0 + \delta_1 X_t \times \epsilon_{i,t} + \delta_2 Z_t \times \epsilon_{i,t} + \alpha_t + \alpha_i + \nu_{i,t+1} \tag{3.15}$$

For the predictive panel regressions δ_1 should be less than zero for broker dealer squared leverage and the financial sector wealth share and should be positive for primary dealer squared leverage and the combined predictor η. I include the same controls for Z_t in the predictive regressions as in the portfolio regressions from the previous section. As in the contemporaneous regressions I include time and stock fixed effects.

Table 3.9 paints a similar picture for the predictive regressions for individual stocks as in the portfolio level analysis from the preceding section. Interaction terms on the combined state variable η are strongly positive across specifications, as is primary dealer squared leverage. Significantly negative coefficients obtain on the broker dealer leverage and financial sector wealth share interactions with $\epsilon_{i,t}$, in accordance with the theory and the empirics in the previous section. Meanwhile, alternative predictors don't seem to have predictability that relates strongly with intermediation, except in the case of the cyclically-adjusted price to earnings ratio, which has negative sign and is significant in the specification in column 6. However, recall that the "P/E" coefficient in the portfolio regressions had an opposite positive sign, so that the estimates for the cyclically-adjusted price to earnings ratio are inconsistent across portfolio and stock-level settings. Meanwhile the intermediary state variables have statistically significant coefficients with consistent signs across specifications, and have comparable magnitudes as well.

To put the predictability regressions in economic terms, once again consider the hypothetical stock that is entirely owned by mutual funds/hedge funds/other investment advisors relative to a stock owned entirely by households but with the same characteristics. The first row of Table 3.9 implies that the intermediated stock has a risk premium that is 11 to 12% on an annualized basis higher relative to the unintermediated stock when η increases by one standard deviation.

For a final stock-level test, I examine the relationship between residual intermedation $\epsilon_{i,t}$

and rolling stock betas on intermediary shocks. I first compute rolling betas for each stock i and each intermediary shock:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i \mathrm{F}_t + \beta_i^M \left(\mathrm{Mkt}_t^{NonFin} - \mathrm{Rf}_t \right) + \delta_{i,t} \tag{3.16}$$

individually for $F =$ Capital Shock, Leverage Shock, Intermediary Shock, and Ex Ret (Fin). The parameter β_i, t is estimated at each time t using a rolling window of plus or minus 15 quarters, including the given quater. I then run the panel regression

$$\widehat{\beta}_{i,t-15 \to t+15} = \alpha_0 + \beta_1 \epsilon_{i,t} + \beta_2 Z_{i,t} + \alpha_t + \alpha_i + \nu_{i,t} \tag{3.17}$$

The controls $Z_{i,t}$ include profitability, investment, CAPM beta, book/market, second-degree polynomial in log market cap and log book equity, in addition to stock and time fixed effects. I require the estimated betas to have at least 20 observations in order to include the observation in (3.17). Because of the overlapping windows I double cluster the standard errors by stock and time. Table 3.10 shows that the individual stock betas centered around time t on each of the intermediary shocks are each strongly increasing intermediation measure $\epsilon_{i,t}$, with t-stats ranging from 2.67 to 3.85. Thus the component of intermediary holdings unrelated to characteristics has strong explanatory power for time-variation in betas even at the individual stock level. The coefficient of 6.7 on $\epsilon_{i,t}$ in column 1 for the combined AEM/HKM intermediary shock is comparable (albeit slightly lower) to the magnitudes found in Table 3.8, and has the interpretation that holding stock characteristics constant, the return response of a completely intermediated stock to a one-standard deviation intermediary shock is 6.67 percentage points higher on an annualized basis relative to a comparable but completely household-owned stock.

3.3.4 Additional Tests and Robustness

A natural question arises concerning whether or not results depend crucially on the characteristics included in, or excluded from, the regression (3.9) to back out the residual intermediation component $\epsilon_{i,t}$. Though this can't be ruled out perfectly, I examine the empirical robustness of my findings to the inclusion of many more characteristics or alternatively to just controlling for size. To do this, I download the set of stock financial ratios provided by the Wharton Research Data Services Financial Ratios Suite. This set of stock characteristics was used previously by Kozak, Nagel, and Santosh (2019) to construct a stochastic discount factor from a large number of potential cross-sectional return predictors.

Though I obtain the full set of 73 financial ratios from WRDS, I restrict the set of characteristics to 40 out of the 73 due to data availability restrictions that I impose.[21] Using the categories provided by WRDS, the 40 ratios that remain comprise 6 valuation ratios, 13 profitability ratios, 4 capitalization ratios, 7 financial soundness ratios, 3 solvency ratios, 3 efficiency ratios, and 4 other ratios. I supplement the original set of characteristics included in (3.9), which consisted of a second degree polynomial in log book equity; gross profitability to book equity; annual growth in firm assets; book-to-market ratio using one-year lagged market cap; and, 5-year rolling monthly pre-ranking CAPM beta (requiring at least 24 observations to be included) with these 40 financial ratios and examine if including the additional characteristics substantially changes anything. On the other end, I also check the robustness of my results to the inclusion of just the second degree polynomial in log book equity. Using these alternative sets of characteristics I re-estimate $\epsilon_{i,t}$ and re-form the quarterly quintile portfolios.

Further robustness checks include value-weighting the portfolios using one-year lagged market cap (and which uses value-weighted cross-sectional regressions to back out $\epsilon_{i,t}$); dropping the financial crisis from the sample (defined using the dates calculated by the NBER as beginning after the business cycle peak in the end of the fourth quarter of 2007 and ending after the business cycle trough in the second quarter of 2009); and, controlling for the Fama-French factors plus momentum as in the stock-level panel regressions from the last section. The contemporaneous regressions are found in Table 3.11 and the predictive regressions are in Table 3.12. Table 3.11 uses my primary proxy for contemporaneous shocks to risk-bearing capacity, the "Intermediary Shock" (which is the average of the standardized

[21] I outline the process I use for selecting these characteristics in detail in Appendix 3.8.

AEM/HKM shocks). Meanwhile in the Table 3.12 predictive regressions, I focus on the state variable η, which is my main proxy for the time t risk-bearing capacity and is constructed as the average of the standardized primary dealer squared leverage ratio and the negative of standardized broker-dealer leverage. The tables report regression coefficients for the high minus low intermediation quintile spread portfolio.

Table 3.11 demonstrates that the intermediary shock significantly explains the spread in returns between high and low residual intermediation portfolios no matter the specification or the set of characteristics included. Interestingly, without controlling for the other characteristics the non-financial market risk factor also strongly loads on the returns to the spread portfolio, but this is not the case in any of the other specifications. Value-weighting changes little, nor does controlling for the Fama-French (2015) risk factors plus the momentum factor. In the last column we do see that both dropping the financial crisis and including the additional risk factors increases estimation noise substantially and reduces the t-stat on the intermediary shock to 1.75. Still, testing the theoretical prediction that the loading is positive entails a one-tailed rather than a two-tailed test, which would still imply significance at the 5% level for this coefficient. The point-estimates for both specifications excluding the crisis are also lower, suggesting that the financial crises was an important event in determining the endogenous covariance with shocks to intermediaries, and points to the magnified effects of shocks to risk-bearing capacity during times when intermediaries were likely financially constrained, consistent with features in the constrained regions of the models of He and Krishnamurthy (2013) and Brunnermeier and Sannikov (2014).

The predictive regressions in Table 3.12 have the same features. Increasing η (or decreasing intermediary risk tolerance) predicts higher returns going forward on the top intermediation portfolio relative to the low intermediation portfolio. As in Table 3.11, the specification which only includes log book equity features more significant coefficients on the control predictors, but this almost entirely goes away in the other specifications. The coefficient on η remains quite stable, strongly significant, and positive for all specifications, with value-weighting portfolios, including more stock characteristics, or dropping the crisis hardly affecting the estimates nor the significance. Interestingly, the coefficient on η goes up a bit in the specification that drops the financial crisis, though the estimation error does increase and the predictability as measured by the R-squared decreases, from 8.7% to 7.5%.

While one can never account for all information regarding a stock, Tables 3.11 and 3.12 illustrate that the empirical patterns are robust to conditioning on a wide range of characteristics so long as stock size is taken into account. It should also be noted that unobserved characteristics would tend to bias against finding an effect. This is because stocks whose cashflows have naturally higher covariance with shocks to intermediaries provide very poor hedges against bad times for financial institutions, and so observed holdings are unlikely to be driven by some underlying institutional preference for high-intermediary shock beta stocks. Thus unobserved stock information would tend to result in understating rather than overstating these effects.

3.4 Discussion of Empirical Results

What are the reasons that the risk-bearing capacity of mutual funds, hedge funds, and other investment advisors depends on shocks to bank holding companies of Federal Reserve Primary dealers (and the broker-dealer sector in general)? The connection for hedge funds is most readily apparent, as hedge funds are levered institutional investors who depend heavily on capital provision by dealer banks for their ability to trade actively in equity markets. For example, Aragon and Strahan (2012) list the top prime brokers to hedge funds in the years 2002-2008 leading up to the financial crisis; the vast majority of the top ten institutions and all of the top five each year were also Federal Reserve primary dealers at the time. Cho (2019) also argues that hedge fund capital depends on the AEM broker-dealer leverage. When these institutions become distressed, capital availability declines and hedge funds in turn also become distressed. In line with this, Ben-David, Franzoni, and Moussawi (2011) demonstrate that hedge funds were forced to delever when their institutional capital providers withdrew capital via margin calls and redemptions.

Ben-David, Franzoni, and Moussawi (2011) show that mutual funds also suffered redemptions in the crisis period, although they were not as severe as those of hedge funds. Since most mutual funds do not use leverage, they are not as dependent on levered institutions such as dealer banks for obtaining capital. However, there are a few reasons to believe that mutual funds' ability to trade may be impeded when levered dealer banks/broker dealers and hedge funds become distressed. When dealer banks reduce their exposure to equity markets, whether directly through their trading desks or indirectly by seeking redemptions from hedge funds, mutual funds lose natural counterparties to their trades when market liquidity dries up. Thus as argued in He and Krishnamurthy (2013), in a liquidation crisis the distress of levered institutions may be the most relevant for determining price movements.

In line with this argument, Nagel (2012) documents that returns to liquidity provision in equity markets dramatically spiked during the financial crisis as levered financial institutions in distress required high price concessions in return for offering liquidity. As mutual funds represent the largest class of institutional investors in equity markets, entering and exiting positions requires shifting large amounts of capital. Consequently mutual funds' ability to trade in equity markets is likely to be highly dependent on the health of levered institutions who supply market liquidity for their trades, even if they directly obtain most of their investment capital from households rather than dealer banks/broker-dealers. Moreover, as noted by Gompers and Metrick (2001), the distribution of asset managers in the 13F data is highly skewed so that the holdings are dominated by a relatively small set of very large institutional investors. These investors with large concentrated positions are likely to value the market liquidity provision of levered broker dealers much more relative to small and dispersed individual retail investors whose trades are comparably miniscule in size.

Another more direct connection for at least some mutual funds and investment advisors comes from the fact that dealer banks/broker-dealers often directly operate equity-focused funds through asset management subsidiaries whose holdings would be classified under mutual funds or investment advisors in the 13F data. In fact, nearly every historical Federal Reserve primary dealer in the post-1980 period has a subsidiary fund manager in the 13F data that is identified as a mutual fund or investment advisor using the Koijen and Yogo (2019) corrected type codes. Hence shocks to their bank holding companies would likely have directly diminished the willingness of such funds to take risk via internal capital markets.

The empirical facts that are documented in this paper are all broadly in accord with the mechanisms detailed above. As explained in proposition 3 of section 3.1, unless household risk tolerance shocks are negatively correlated with risk tolerance shocks of mutual funds/hedge funds/investment advisors, increasing price responses to intermediary shocks along the dimension of increased intermediation *must* come because these institutions' ability to take on risk is inordinately affected by these shocks. Thus in any case my findings imply that the largest institutional investors in equity markets are directly affected by shocks to dealer banks and other broker-dealers; the discussion above simply offers some potential explanations as to why this is the case.

3.5 Conclusion

Building off of theoretical and empirical work that features constrained intermediaries as marginal investors, I show that the asset holdings of financial institutions generate higher covariances of more intermediated stocks with shocks to intermediary risk-bearing capacity, via temporary differential movements in discount rates. After accounting for stock fundamentals, stocks that are held more by intermediaries covary more with shocks to intermediaries' ability to take on risk, and state variables capturing the health of financial intermediaries predict better the returns of the more intermediated stocks than the less intermediated stocks, again conditional on stocks having similar characteristics. These effects are large in economic magnitude. Two alike stocks would have annualized conditional risk premia that are 11-12% higher at the quarterly horizon if owned entirely by financial institutions instead of households following a one standard deviation drop in intermediary risk tolerance. Furthermore, the beta on shocks to intermediary risk-bearing capacity on a portfolio formed on the most intermediated stocks is more than 5 times higher than the beta on the least intermediated portfolio, despite the two portfolios being comprised of stocks that are on average of the same size, book/market, investment (asset growth), profitability, and CAPM betas.

Previous empirical papers testing frictional intermediary asset pricing theories have tended to focus on asset markets that are comparatively difficult for households to access. By contrast, I demonstrate that effects predicted by intermediary asset pricing models persist even among equities, which is perhaps the easiest asset class for households to directly access. In this sense the findings in this paper may provide a lower bound for the relative importance of intermediary asset pricing in other asset classes.

The empirical evidence presented in this paper suggests that even the risk-bearing capacity of large institutional investors who tend to avoid taking on leverage, such as mutual funds, still depends on the health of levered dealer banks. Accordingly, future research may examine this connection in more depth, including quantifying how much the ability of large institutional investors to trade or bear risk in equity markets directly depends upon the liquidity provision of levered institutions such as dealer banks, broker-dealers, and hedge funds.

Bibliography

[1] Adrian, Tobias, Erkko Etula, and Tyler Muir, 2014, Financial intermediaries and the cross-section of asset returns, *The Journal of Finance* 69, 2557–2596.

[2] Adrian, T., Shin, H.S., 2014. Procyclical leverage and value-at-risk. *Review of Financial Studies* 27 (2), 373–403

[3] Adrian, T. , Moench, E. , Shin, H.S. , 2014. Dynamic leverage asset pricing. Working Paper. Federal Reserve Bank of New York

[4] Aragon, George A. and Philip E Strahan, 2012. Hedge funds as liquidity providers: Evidence from the Lehman bankruptcy, *Journal of Financial Economics*, 103 (2012) 570–587.

[5] Anderson, Chris, and Weiling Liu, 2018, The Shadow Price of Intermediary Constraints, Working paper.

[6] Baker, Malcolm, and Jeffrey Wurgler, 2006, Investor sentiment and the cross-section of stock returns, *Journal of Finance* 61, 1645–1680.

[7] Basak, S., Pavlova, A., 2013. Asset prices and institutional investors. *American Economic Review* 103 (5), 1728–1758.

[8] Ben-David, Itzhak, Franzoni, Francesco, and Rabih Moussawi, 2012. Hedge Fund Stock Trading in the Financial Crisis of 2007–2009. *The Review of Financial Studies*, 25 (1), 1–54.

[9] Bernanke, B., Gertler, M. , 1989. Agency costs, net worth, and business fluctuations. American Economic Review 79, 14–31 .

[10] Brav, A., G. M. Constantinides, and C. C. Geczy. Asset pricing with heterogeneous consumers and limited participation: Empirical evidence. *Journal of Political Economy* 110 (2002): 793–824.

[11] Brunnermeier,Markus, and Yuliy Sannikov, 2014, A macroeconomic model with a financial sector, *American Economic Review*, 104, 379–421.

[12] Brunnermeier, Markus, and Pedersen,L.H., 2009, Market liquidity and funding liquidity, *Review of Financial Studies*, 22 (6), 2201-2238

[13] Chen, Hui, Scott Joslin, and Sophie X Ni, 2016, Demand for crash insurance, intermediary constraints, and risk premia in financial markets, Forthcoming *Review of Financial Studies*.

[14] Cho, Thummim, 2019, Turning Alphas into Betas: Arbitrage and the Cross-Section of Risk, *Journal of Financial Economics*, forthcoming.

[15] Cochrane, John. 2011. Presidential Address: Discount Rates *The Journal of Finance* 66, (4) 1047-1108.

[16] Duffie, Darrell, 2010, Presidential address: Asset price dynamics with slow-moving capital, *The Journal of Finance* 65, 1237–1267.

[17] Fama, Eugene F. and Kenneth R. French. 2015. A Five-Factor Asset Pricing Model. *Journal of Financial Economics* 116 (1):1–22.

[18] Gabaix, X. , Krishnamurthy, A. , Vigneron, O. , 2007. Limits of arbitrage: theory and evidence from the mortgage-backed securities market. Journal of Finance 62 (2), 557–595.

[19] Gârleanu, Nicolae, Stavros Panageas, and Jianfeng Yu. 2015. Financial Entanglement: A Theory of Incomplete Integration, Leverage, Crashes, and Contagion. *American Economic Review*, 105(7): 1979-2010.

[20] Gompers, Paul A. and Andrew Metrick. 2001. Institutional Investors and Equity Prices, *Quarterly Journal of Economics* 116 (1):229–259.

[21] Gromb, D. and Vayanos, D. (2018), The Dynamics of Financially Constrained Arbitrage. *The Journal of Finance*, 74: 371-399.

[22] Haddad, Valentin and Tyler Muir, 2018, Do intermediaries matter for aggregate asset prices? Working Paper.

[23] Haddad, Valentin and David Sraer, 2018, The Banking View of Bond Risk Premia. Working Paper.

[24] He, Zhiguo, Kelly, Bryan, and Asaf Manela, 2017 Intermediary asset pricing: New evidence from many asset classes, *Journal of Financial Economics* 126, 1-35

[25] He, Zhiguo, and Arvind Krishnamurthy, 2012. A model of capital and crises. The Review of Economic Studies 79 (2), 735–777

[26] He, Zhiguo, and Arvind Krishnamurthy, 2013, Intermediary asset pricing, *American Economic Review* 103, 732–770.

[27] He, Zhiguo, and Arvind Krishnamurthy, 2018, forthcoming, Intermediary asset pricing and the financial crisis, *Annual Review of Financial Economics*.

[28] Holmstrom, B, and Jean Tirole, Financial Intermediation, Loanable Funds, and The Real Sector, *The Quarterly Journal of Economics*, Volume 112, Issue 3, 1 August 1997, Pages 663–691

[29] Kewei, Hou, Xue, Chen, and Lu Zhang, Digesting Anomalies: An Investment Based Approach. *The Review of Financial Studies*, Volume 28, Issue 3, March 2015, Pages 650–705

[30] Kondor, P. and Vayanos, D. (2019), Liquidity Risk and the Dynamics of Arbitrage Capital. *The Journal of Finance*, 74: 1139-1173

[31] Koijen, Ralph, and Motohiro Yogo. *Journal of Political Economy*, 127(4), 2019, 1475–1515,

[32] Ralph S.J. Koijen and Motohiro Yogo. 2015, The cost of financial frictions for life insurers. *American Economic Review*, 105:445–475.

[33] Kozak Serhiy, Nagel Stefan, and Shrihari Santosh, 2019. Shrinking the cross-section, *Journal of Financial Economics*, forthcoming.

[34] Krishnamurthy, Arvind . How debt markets have malfunctioned in the crisis. Journal of Economic Perspectives, 24(1):3–28, March 2010

[35] Lettau, Martin, and Sydney Ludvigson, 2001, Resurrecting the (c)capm: A cross-sectional test when risk premia are time-varying, *Journal of Political Economy* 109, 1238–1287.

[36] Ma, Sai, 2017, Heterogeneous Intermediaries and Asset Prices. Working Paper.

[37] Mitchell, M. , Pedersen, L.H. , Pulvino, T. , 2007. Slow moving capital. American Economic Review, Papers and Proceedings 97, 215–220.

[38] Mitchell, M. , Pulvino, T. , 2012. Arbitrage crashes and the speed of capital. Journal of Financial Economics 104 (3), 469–490 .

[39] Muir, T. , 2017. Financial crises and risk premia. *Quarterly Journal of Economics* 132, 765–809.

[40] Nagel, Stefan, 2005, Short sales, institutional investors and the cross-section of stock returns, *Journal of Financial Economics* 78, 277–309.

[41] Nagel, Stefan, 2012, Evaporating Liquidity, *The Review of Financial Studies*, 25, (7), 2005–2039.

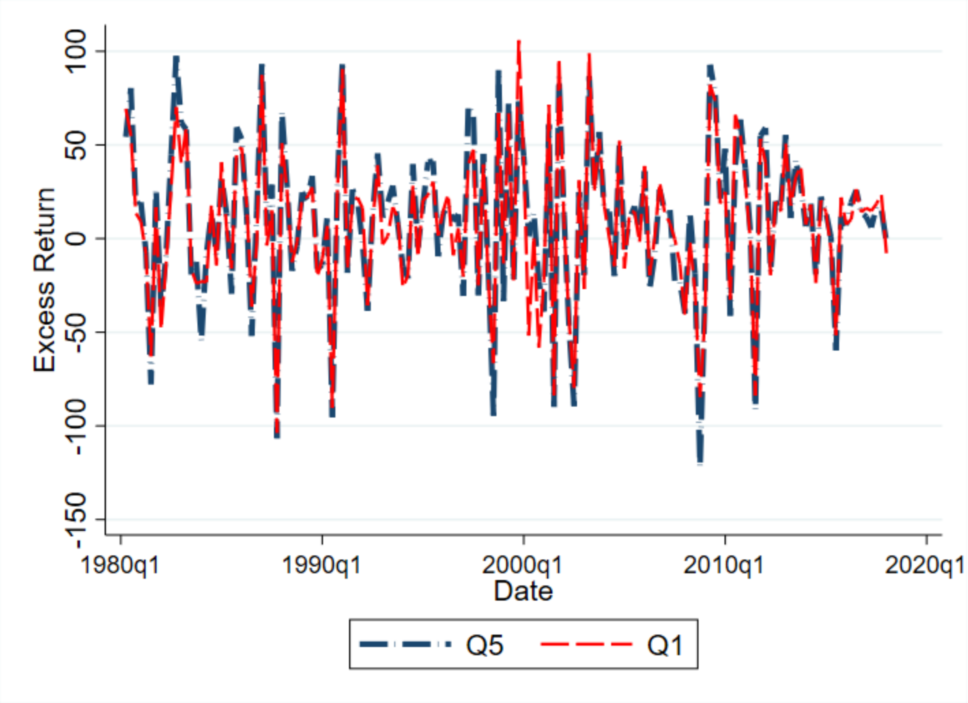
[42] Shleifer, A. and Vishny, R. W. 1997, *The Limits of Arbitrage*. The *Journal of Finance*, 52: 35–55.

[43] Siriwardane, Emil N. Limited Investment Capital and Credit Spreads, 2018. Forthcoming, *Journal of Finance*.

[44] Stambaugh, Robert F. 2014. Investment Noise and Trends. *Journal of Finance*, 69 (4), 1415-1453.

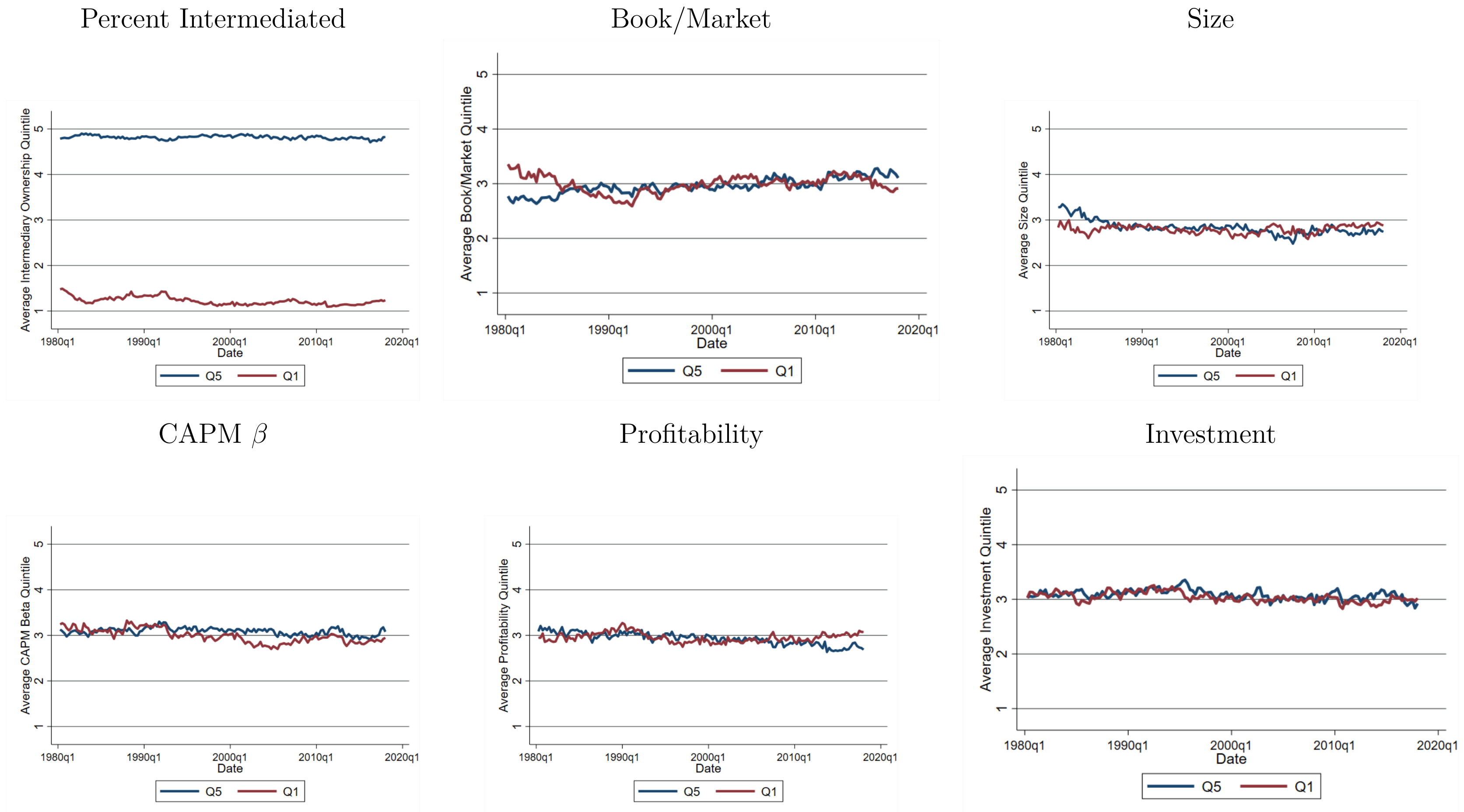
Figures

Figure 3-2: Annualized Excess Return For Top and Bottom Intermediation ($\epsilon_{i,t}$) Quintile Portfolios



This figure shows the time series of annualized quarterly excess returns on the top and bottom quintile equal-weighted portfolios formed on the intermediation measure $\epsilon_{i,t}$. Details on the construction of $\epsilon_{i,t}$ are presented in section 3.3.1. Sample spans 1980q2 to 2017q3.

Figure 3-3: Average Quintile of Given Characteristic For Top and Bottom Intermediation ($\epsilon_{i,t}$) Quintile Portfolios


Percent Intermediated

Book/Market

Size

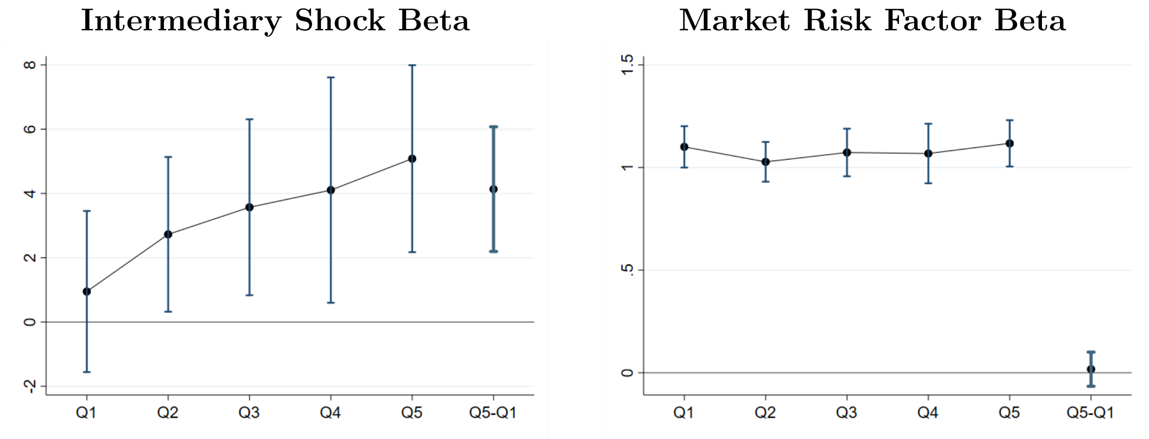
CAPM β

Profitability

Investment

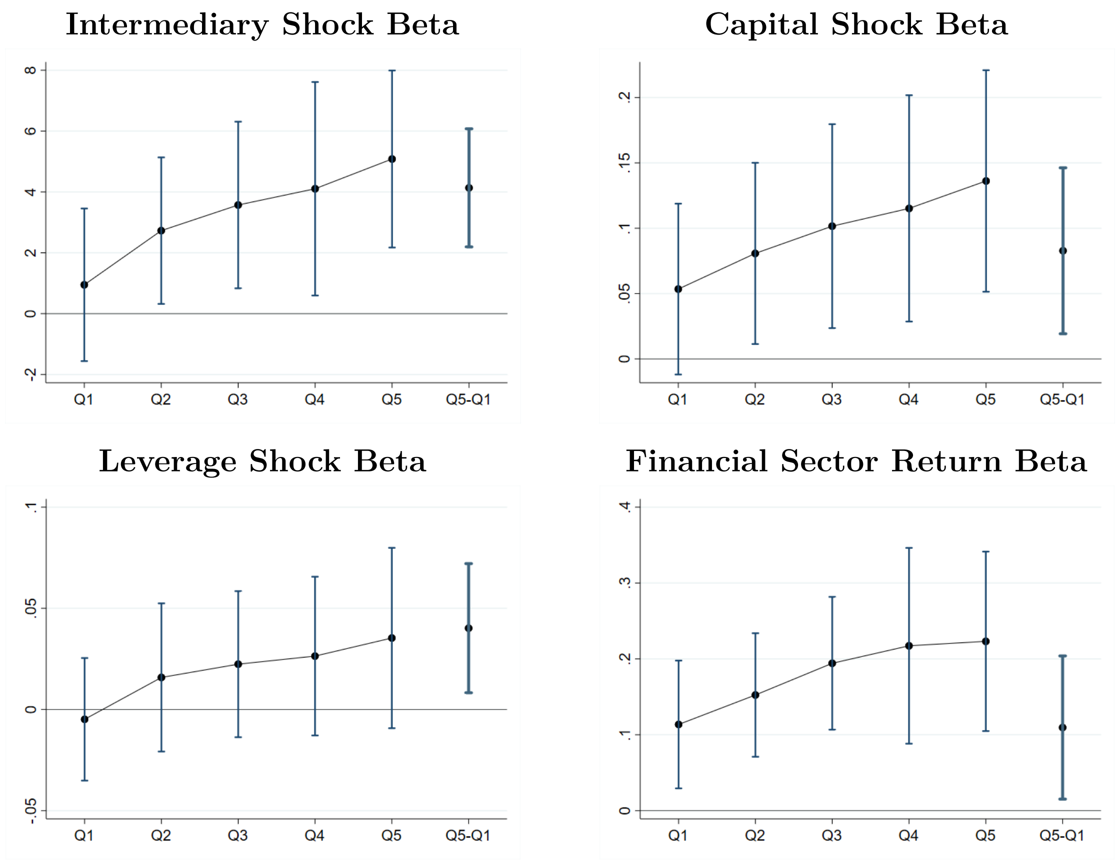
This figure shows average the quintiles over time for given characteristics for top and bottom quintile equal-weighted portfolios formed on the intermediation measure $\epsilon_{i,t}$. Details on the construction of $\epsilon_{i,t}$ are presented in section 3.3.1. Sample spans 1980q2 to 2017q3.

Figure 3-4: Coefficients on Intermediary Shocks and Market Risk Over Portfolios Formed on Intermediation Quintile



This figure plots regression coefficients as in (3.10) of the main text. The figure on the left shows the coefficient estimates on the average of the standardized Federal Reserve primary dealer equity capital ratio shocks from He, Kelly, and Manela (2017) and the broker-dealer book leverage growth shocks from Adrian, Etula, and Muir (2014) for five portfolios formed on the intermediation measure $\epsilon_{i,t}$ (which is constructed so as to be uncorrelated with fundamental stock characteristics; section 3.3.1 discusses this in more detail), as well as the coefficient on the intermediary shocks for the top minus bottom quintile spread. The figure on the right shows the corresponding betas on a version of the value-weighted market risk factor that excludes returns on financial stocks (SIC code between 6000 and 6999). The confidence bands represent 95% confidence intervals computed from Newey-West standard errors. The Intermediary Shock measure is standardized and returns are in annualized percent form. Sample spans 1980q2 to 2017q3.

Figure 3-5: Betas On Portfolios Sorted By Intermediation on Different Shocks To Intermediary risk-bearing Capacity


Intermediary Shock Beta

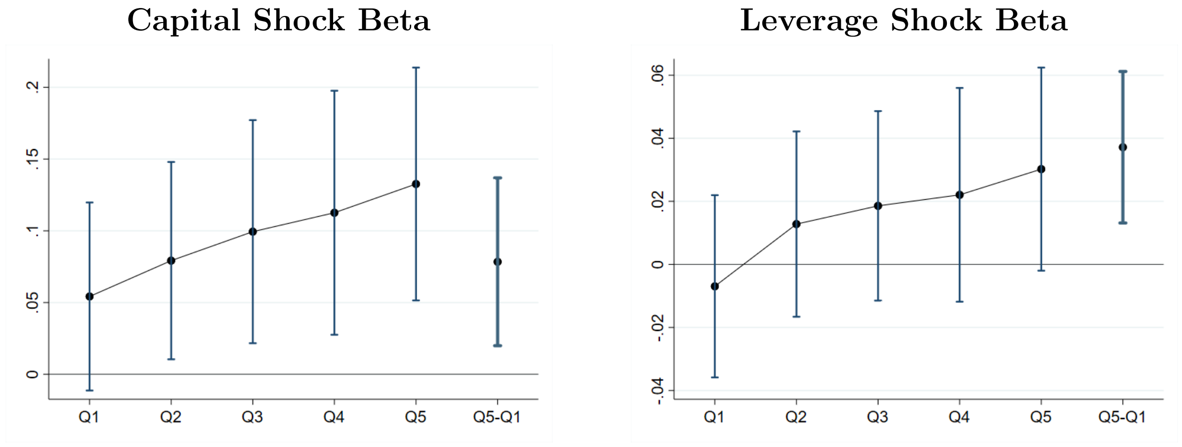
Capital Shock Beta

Leverage Shock Beta

Financial Sector Return Beta

This figure presents regressions estimates as in (3.10) of the main text for each of the proposed intermediary shocks for each of the five portfolios formed on quintiles of the intermediation measure $\epsilon_{i,t}$ constructed in section 3.3.1 of the main text, as well as the top minus bottom quintile portfolio spread. The capital and shocks refer to the Federal Reserve primary dealer equity capital ratio shocks proposed in He, Kelly, and Manela (2017), while the leverage shocks refer to the broker-dealer leverage shocks from Adrian, Etula and Muir (2014). Intermediary shock refers to the average of the standardized leverage and capital shocks. Financial sector return is the value-weighted return on the financial sector (stocks with SIC code between 6000 and 6999). Regressions control for a version of the value-weighted market risk factor that excludes financial stocks. The Intermediary Shock measure is standardized and returns are in annualized percent form. Error bands represent 95% Newey-West confidence intervals. Sample spans 1980q2 to 2017q3.

Figure 3-6: Betas On Portfolios Sorted By Intermediation On Capital and Leverage Shocks Included in Same Specification


Capital Shock Beta

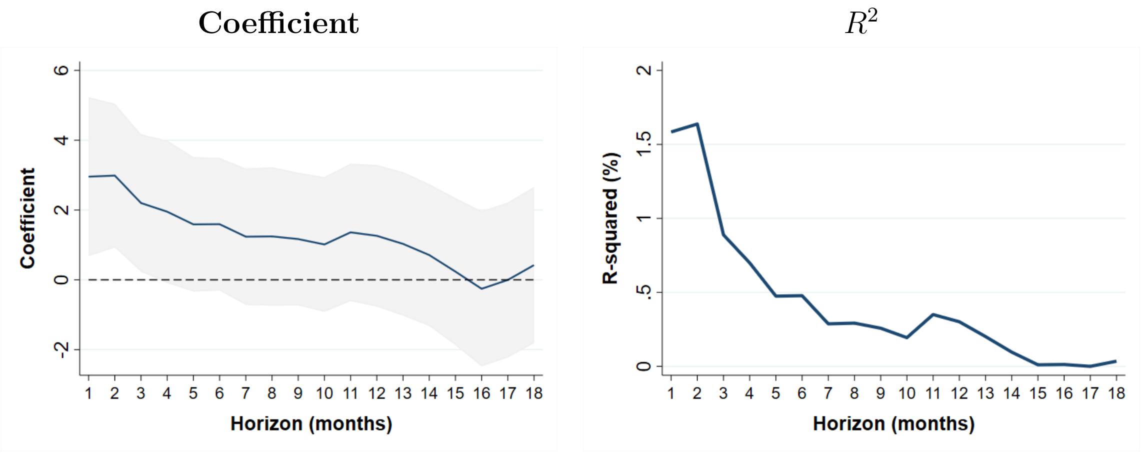
Leverage Shock Beta

This figure presents regressions estimates as in (3.11) of the main text:

$$R_t^e = \beta_{0,i} + \beta_{1,i}\text{Capital Shock}_t + \beta_{2,i}\text{Leverage Shock}_t + \beta_{3,i}\left(\text{Mkt}^{NonFin} - \text{Rf}\right)_t + \epsilon_{i,t} \qquad (3.18)$$

This plots show betas on capital and leverage shocks included together in the same specification and for each of the five portfolios formed on quintiles of the intermediation measure $\epsilon_{i,t}$ constructed in section 3.3.1 of the main text, as well as the top minus bottom quintile portfolio spread. The capital shocks refer to the Federal Reserve primary dealer equity capital ratio shocks proposed in He, Kelly, and Manela (2017), while the leverage shocks refer to the broker-dealer leverage shocks from Adrian, Etula and Muir (2014). Error bands represent 95% Newey-West confidence intervals. Sample spans 1980q2 to 2017q3.

Figure 3-7: Predictability of One Month High Minus Low Intermediation Spread Portfolio Returns On Intermediary risk-bearing Capacity At Different Monthly Horizons



Coefficient

R^2

This figure shows coefficients obtained from predictive regressions of the one month high minus low return spread for portfolios formed on top and bottom quintiles of intermedation measure $\epsilon_{i,t}$ (which is constructed in section 3.3.1 of the text) on predictor η_t at different monthly horizons. Regressions are of the form

$$R^{Q5}_{t+k} - R^{Q1}_{t+k} = \alpha + \beta_k \eta_t + \nu_t \tag{3.19}$$

as in equation (3.13) in the main text. The horizon k varies from 1 month to 18 months. The predictor η_t is the average of the standardized primary dealer squared leverage from He, Kelly, and Manela (2017) and the negative of standardized broker dealer leverage from Adrian, Etula, and Muir (2014). The gray shaded area corresponds to 90% Newey-West confidence intervals with one lag.

Tables

Table 3.1: Fama-Macbeth Regressions of Percent Stock Ownership By Intermediaries on Baseline Stock Characteristics

	Percent Intermediated$_{i,t}$
Log Book Equity	0.058^{***}
	(15.13)
Log Book Equity Sq.	-0.0045^{***}
	(-10.52)
Profitability	0.024^{*}
	(1.92)
CAPM Beta	0.047^{***}
	(5.39)
Asset Growth	0.028^{***}
	(4.13)
Book/Market	-0.015^{***}
	(-6.53)
Observations	214448
Average R^2	0.11

This table shows the Fama-Macbeth time series average coefficients from the cross-sectional regression in (3.9):

$$\text{Percent Intermediated}_{i,t} = \alpha_0 + \alpha_t + \beta X_{i,t} + \epsilon_{i,t}$$

for stocks included in the sample. At each time t the top 1% of observations of Percent Intermediated$_{i,t}$ are winsorized to deal with outliers in the cross-section of institutional holdings. T-stats in parentheses are computed using Fama-Macbeth standard errors, robust to 8 lags of autocorrelation. Average R^2 refers to the time series average of the R-squared from each cross-sectional regression. The sample ranges from 1980q2 to 2017q3.

Table 3.2: Summary Statistics of Stock Characteristics For Portfolios Sorted on Quintiles of Intermediation Measure $\epsilon_{i,t}$

Panel A: Portfolio Characteristic Means

	% Inst	Log(ME)	Log(BE)	BE/ME	Asset Growth	Prof/BE	CAPM β
Q1	.2	6.84	6.16	.88	.14	.22	1.17
Q2	.34	7.22	6.52	.92	.12	.23	1.11
Q3	.43	7.25	6.54	.92	.12	.23	1.16
Q4	.51	7.14	6.43	.87	.13	.23	1.18
Q5	.62	6.92	6.16	.88	.14	.21	1.18

Panel B: Portfolio Characteristic Medians

	% Inst	Log(ME)	Log(BE)	BE/ME	Asset Growth	Prof/BE	CAPM β
Q1	.17	6.7	5.82	.79	.14	.22	1.17
Q2	.32	7.19	6.24	.83	.12	.23	1.09
Q3	.42	7.26	6.28	.82	.12	.23	1.13
Q4	.52	7.14	6.2	.8	.13	.23	1.17
Q5	.65	6.93	5.86	.82	.14	.22	1.19

This table shows the means and medians of percent holdings by institutional investors (mutual funds, hedge funds, and investment advisors), log market equity, log book equity, book/market, asset growth (investment), profitability to book equity, and pre-ranking CAPM beta for the 5 portfolios formed on quintiles of the intermediation measure $\epsilon_{i,t}$. Details on the construction of $\epsilon_{i,t}$ are presented in section 3.3.1. Sample spans 1980q2 to 2017q3.

Table 3.3: Return Summary Stats For Portfolios Formed on Quintiles of Intermediation Measure $\epsilon_{i,t}$

	Q1	Q2	Q3	Q4	Q5	Q5-Q1
μ(Ex Ret)	8.76	10.14	10.46	11.2	10.63	1.79
t-stat	2.84	3.38	3.3	3.51	3.16	1.85
σ(Ex Ret)	38.97	37.05	39.33	40.02	42.33	12.22
Sharpe Ratio	.22	.27	.27	.28	.25	.15

This table reports the means, standard deviations, and Sharpe Ratios for the percent annualized excess returns for portfolios formed on quintiles of intermediation measure $\epsilon_{i,t}$. Details on the construction of $\epsilon_{i,t}$ are presented in section 3.3.1. Sample spans 1980q2 to 2017q3.

Table 3.4: Regressions of Quintile-Sorted Portfolios Formed by Intermediation Measure $\epsilon_{i,t}$ on Contemporaneous Intermediary Shocks

	Q1	Q2	Q3	Q4	Q5	Q5-Q1
Intermediary Shock	0.948	2.727**	3.570**	4.104**	5.082***	4.134***
	(0.75)	(2.24)	(2.58)	(2.31)	(3.45)	(4.21)
$\text{Mkt}^{NonFin} - \text{Rf}$	1.101***	1.028***	1.073***	1.068***	1.118***	0.017
	(21.59)	(20.96)	(18.28)	(14.55)	(19.56)	(0.41)
Observations	150	150	150	150	150	150
R^2	0.88	0.90	0.89	0.86	0.87	0.13

This table shows regressions of the intermediation measure $\epsilon_{i,t}$ portfolio quintile excess returns (and top minus bottom quintile spread) on risk factors as in (3.10) of the main text:

$$R^e_{i,t+1} = \alpha_i + \beta_{1,i}\text{Intermediary Shock}_{t+1} + \beta_{2,i}(\text{Mkt}^{NonFin}_{t+1} - R_{f,t}) + \nu_{i,t} \tag{3.20}$$

The first row of this table shows the coefficient estimates on the average of the standardized Federal Reserve primary dealer equity capital ratio shocks from He, Kelly, and Manela (2017) and the broker-dealer book leverage growth shocks from Adrian, Etula, and Muir (2014) for five portfolios formed on the intermediation measure $\epsilon_{i,t}$ (which is constructed so as to be uncorrelated with fundamental stock characteristics; section 3.3.1 discusses this in more detail), as well as the coefficient on the intermediary shocks for the top minus bottom quintile spread. The second row shows the betas on a version of the value-weighted market risk factor that excludes returns on financial stocks (SIC code between 6000 and 6999). The sample is quarterly and comprises 1980q2 to 2017q3. Newey-West t-stats are in parentheses. Quarterly excess returns are in annualized percent form and the intermediary shock is standardized. Sample spans 1980q2 to 2017q3.

Table 3.5: Predictability Regressions of High minus Low Intermediation Spread Portfolio Returns On Intermediary risk-bearing Capacity

	(1)	(2)	(3)	(4)	(5)	(6)
η	2.828***	5.506***	2.806***	2.826***	2.818***	6.236***
	(2.87)	(3.57)	(2.83)	(2.84)	(2.87)	(3.93)
P/E		3.479				4.460*
		(1.53)				(1.90)
cay			0.401			0.658
			(0.49)			(0.77)
10Y-3Mo				-0.075		0.735
				(-0.07)		(0.80)
sentiment					1.418	1.726
					(1.37)	(1.56)
Observations	450	450	450	450	450	450
R^2	0.041	0.066	0.042	0.041	0.051	0.087

This table shows coefficients obtained from regressing overlapping quarterly returns at the monthly frequency of the high minus low intermediation portfolio excess returns on η, the average of the standardized primary dealer squared leverage from He, Kelly, and Manela (2017) and the negative of standardized broker dealer leverage from Adrian, Etula, and Muir (2014):

$$R^{Q5}_{t\rightarrow t+3} - R^{Q1}_{t\rightarrow t+3} = \alpha + \beta_1\eta_t + \beta_2 Z_t + \nu_t \tag{3.21}$$

Controls include P/E, the cyclically-adjusted price to earnings ratio; cay, the consumption-wealth ratio from Lettau and Ludvigson (2001); the 10 year minus 3 month treasury term spread; and, the Baker and Wurgler (2006) sentiment index. Newey-West t-stats with four lags are presented in parentheses. All independent variables are standardized and excess returns are expressed in annualized percentage form. The high and low portfolios are formed on the intermediation measure $\epsilon_{i,t}$ (which is constructed so as to be uncorrelated with fundamental stock characteristics; section 3.3.1 discusses this in more detail). Sample spans 1980m4 to 2017m9.

Table 3.6: Predictability Regressions of High minus Low Intermediation Spread Portfolio Returns On HKM and AEM State Variables

	(1)	(2)	(3)	(4)	(5)	(6)
PD Lev. Sq.	2.023**	4.563***	2.035**	2.025**	2.133**	5.589***
	(2.22)	(3.52)	(2.22)	(2.23)	(2.41)	(3.98)
BD Lev.	-1.628*	-2.999***	-1.586*	-1.623*	-1.505	-3.273***
	(-1.79)	(-2.83)	(-1.72)	(-1.77)	(-1.63)	(-3.10)
P/E		3.938*				5.295**
		(1.67)				(2.13)
cay			0.425			0.878
			(0.52)			(1.04)
10Y-3Mo				-0.089		0.784
				(-0.08)		(0.86)
sentiment					1.458	1.895*
					(1.40)	(1.65)
Observations	450	450	450	450	450	450
R^2	0.041	0.070	0.042	0.041	0.052	0.096

This table shows coefficients obtained from regressing overlapping quarterly returns at the monthly frequency of the high minus low intermediation portfolio excess returns on the primary dealer squared leverage from He, Kelly, and Manela (2017) and the broker dealer leverage from Adrian, Etula, and Muir (2014):

$$R^{Q5}_{t\to t+3} - R^{Q1}_{t\to t+3} = \alpha + \beta_1 \text{PD Lev Sq}_t + \beta_2 \text{BD Lev}_t + \beta_3 Z_t + \nu_t \tag{3.22}$$

Controls include P/E, the cyclically-adjusted price to earnings ratio; cay, the consumption-wealth ratio from Lettau and Ludvigson (2001); the 10 year minus 3 month treasury term spread; and, the Baker and Wurgler (2006) sentiment index. All independent variables are standardized and excess returns are expressed in annualized percentage form. The high and low portfolios are formed on the intermediation measure $\epsilon_{i,t}$ (which is constructed so as to be uncorrelated with fundamental stock characteristics; section 3.3.1 discusses this in more detail). Sample spans 1980m4 to 2017m9.

Table 3.7: Predictability Regressions of High minus Low Intermediation Spread Portfolio Returns On Financial Sector Stock Market Wealth Share

	(1)	(2)	(3)	(4)	(5)	(6)
Fin. Share	-2.577***	-4.038**	-2.571***	-2.575***	-2.379***	-3.876**
	(-3.06)	(-2.22)	(-2.88)	(-3.05)	(-2.75)	(-2.14)
P/E		2.087				2.092
		(0.88)				(0.89)
cay			0.032			-0.034
			(0.04)			(-0.04)
10Y-3Mo				-0.086		0.322
				(-0.08)		(0.34)
sentiment					0.928	0.854
					(0.88)	(0.78)
Observations	450	450	450	450	450	450
R^2	0.034	0.045	0.034	0.034	0.038	0.049

This table shows coefficients obtained from regressing overlapping quarterly returns at the monthly frequency of the high minus low intermediation portfolio excess returns on the share of stock market wealth held in the financial sector (SIC codes between 6000 and 6999):

$$R^{Q5}_{t\rightarrow t+3} - R^{Q1}_{t\rightarrow t+3} = \alpha + \beta_1 \text{Fin. Share}_t + \beta_3 Z_t + \nu_t \tag{3.23}$$

Controls include P/E, the cyclically-adjusted price to earnings ratio; cay, the consumption-wealth ratio from Lettau and Ludvigson (2001); the 10 year minus 3 month treasury term spread; and, the Baker and Wurgler (2006) sentiment index. Newey-West t-stats with four lags are presented in parentheses. All independent variables are standardized and excess returns are expressed in annualized percentage form. The high and low portfolios are formed on the intermediation measure $\epsilon_{i,t}$ (which is constructed so as to be uncorrelated with fundamental stock characteristics; section 3.3.1 discusses this in more detail). Sample spans 1980q2 to 2017q3.

Table 3.8: Panel Regressions of Stock Excess Returns on Contemporaneous Intermediary Shocks Interacted With Intermediation Measure $\epsilon_{i,t}$

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Intermediary Shock × $\epsilon_{i,t}$	9.01^{***}			8.11^{***}			10.0^{***}		
	(5.32)			(4.96)			(4.63)		
Capital Shock × $\epsilon_{i,t}$		0.17^{***}			0.18^{***}			0.23^{***}	
		(3.51)			(2.65)			(3.96)	
Leverage Shock × $\epsilon_{i,t}$		0.048^{*}			0.048^{*}			0.062^{*}	
		(1.70)			(1.66)			(1.86)	
Ex Ret (Fin.) × $\epsilon_{i,t}$			0.19^{***}			0.22^{**}			0.31^{***}
			(3.17)			(2.48)			(2.94)
$\text{Mkt}^{NonFin} - \text{Rf} \times \epsilon_{i,t}$				0.058	-0.010	-0.049	0.085	-0.014	-0.10
				(0.66)	(-0.09)	(-0.39)	(0.97)	(-0.15)	(-0.76)
SMB × $\epsilon_{i,t}$							0.064	0.056	0.075
							(0.38)	(0.33)	(0.47)
HML × $\epsilon_{i,t}$							-0.17	-0.21	-0.14
							(-1.16)	(-1.45)	(-0.78)
CMA × $\epsilon_{i,t}$							-0.099	-0.082	-0.21
							(-0.60)	(-0.51)	(-1.26)
RMW × $\epsilon_{i,t}$							0.39^{***}	0.40^{***}	0.36^{***}
							(2.87)	(2.92)	(2.64)
UMD × $\epsilon_{i,t}$							-0.056	-0.032	-0.033
							(-0.60)	(-0.34)	(-0.37)
Stock Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	211255	211255	211255	211255	211255	211255	211255	211255	211255
R^2	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.24

This table shows estimates from panel regressions as in (3.14) of the main text:

$$R_{i,t+1} - R_{f,t} = \alpha_0 + \beta_1 F_{t+1} \times \epsilon_{i,t} + \beta_2 W_{t+1} \times \epsilon_{i,t} + \alpha_t + \alpha_i + \nu_{i,t+1} \tag{3.24}$$

Here F_{t+1} denotes shocks to intermediaries and W_{t+1} controls for other common shocks. The capital shocks refer to the Federal Reserve primary dealer equity capital ratio shocks proposed in He, Kelly, and Manela (2017), while the leverage shocks refer to the broker-dealer leverage shocks from Adrian, Etula and Muir (2014). Intermediary shock refers to the average of the standardized leverage and capital shocks. Financial sector return is the value-weighted return on the financial sector (stocks with SIC code between 6000 and 6999). Regressions control for a version of the value-weighted market risk factor that excludes financial stocks. Controls SMB, HML, CMA, RMW, UMD refer to the Fama-French (2015) risk factors and the up minus down momentum factor. In parentheses are t-stats that are clustered by time to adjust for cross-sectional correlation in the residuals and are also adjusted for one lag of autocorrelation. The Intermediary Shock measure is standardized and returns are in annualized percent form. Sample spans 1980q2 to 2017q3.

Table 3.9: Predictive Panel Regressions of Stock Excess Returns on Intermediary State Variables Interacted With Intermediation Measure $\epsilon_{i,t}$

	(1)	(2)	(3)	(4)	(5)	(6)
$\eta \times \epsilon_{i,t}$	12.3^{***}			11.0^{***}		
	(3.98)			(2.96)		
PD Lev. Squared × $\epsilon_{i,t}$		7.01^{***}			4.82^{**}	
		(3.04)			(2.30)	
BD Lev. × $\epsilon_{i,t}$		-8.38^{***}			-8.00^{***}	
		(-2.96)			(-2.70)	
Fin. Share × $\epsilon_{i,t}$			-8.17^{***}			-5.32^{*}
			(-2.75)			(-1.86)
P/E × $\epsilon_{i,t}$				-3.22	-4.29	-7.15^{**}
				(-0.83)	(-1.15)	(-2.17)
cay × $\epsilon_{i,t}$				1.24	1.33	0.54
				(0.53)	(0.56)	(0.22)
10Y-3Mo × $\epsilon_{i,t}$				-0.0083	-0.0074	-0.0077
				(-0.37)	(-0.34)	(-0.34)
sentiment × $\epsilon_{i,t}$				2.73	2.56	0.44
				(0.63)	(0.60)	(0.11)
Stock Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	211255	211255	211255	211255	211255	211255
R^2	0.24	0.24	0.24	0.24	0.24	0.24

This table shows estimates from running predictive regressions of quarterly $t+1$ stock excess returns on date t state variables interacted with the intermediation measure $\epsilon_{i,t}$ (which is constructed so as to be uncorrelated with fundamental stock characteristics; section 3.3.1 discusses this in more detail). The state variable η is the average of the standardized primary dealer squared leverage ratio from He, Kelly, and Manela (2017) and the negative of the standardized broker-dealer leverage ratio from Adrian, Etula, and Muir (2014). Fin. Share is the share of stock market wealth held in financial stocks. The Controls include P/E, the cyclically adjusted P/E ratio; cay, the consumption-wealth ratio from Lettau, and Ludvigson (2001); the 10 year minus 3 month treasury term spread; and, the Baker and Wurgler (2006) sentiment index. All independent variables are standardized and returns are in annualized percent form. Sample spans 1980q2 to 2017q3.

Table 3.10: Panel Regressions of Rolling Stock-Level Intermediary risk-bearing Capacity Betas on Intermediation Measure $\epsilon_{i,t}$

	Intermediary Shock	Capital Shock	Leverage Shock	Ex Ret (Fin.)
$\epsilon_{i,t}$	6.70***	0.14***	0.070***	0.17***
	(3.75)	(3.85)	(2.67)	(3.22)
Stock Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	188453	188453	188453	188453
R^2	0.53	0.54	0.50	0.55

This table shows regressions of rolling individual stock betas on the intermediation measure $\epsilon_{i,t}$ (which is constructed so as to be uncorrelated with fundamental stock characteristics; section 3.3.1 discusses this in more detail). Stock betas are obtained from regressions of the form

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i \mathrm{F}_t + \beta_i^M \left(\mathrm{Mkt}_t^{NonFin} - \mathrm{Rf}_t\right) + \delta_{i,t} \tag{3.25}$$

using a window of plus or minus 15 quarters. Stocks betas must have been estimated using at least 20 observations to be included in the sample. Reported coefficients are then estimated from panel regressions taking the form

$$\widehat{\beta}_{i,t-15\to t+15} = \alpha_0 + \beta_1 \epsilon_{i,t} + \beta_2 Z_{i,t} + \alpha_t + \alpha_i + \nu_{i,t} \tag{3.26}$$

Controls $Z_{i,t}$ include gross profitability, investment (asset growth), CAPM beta, book/market, second-degree polynomial in log market cap and log book equity, plus stock and time fixed effects. In parentheses are t-statistics double clustered by stock and quarter. Returns and risk factors are expressed in annualized percentage terms, with the exception of the intermediary shock, which is standardized to zero mean and unit variance. Sample spans 1980q2 to 2017q3.

Table 3.11: Robustness: Contemporaneous Portfolio Regressions

	Original		Add WRDS Ratios		Just log(BE)		Value-Weighted		Drop Crisis	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intermediary Shock	4.13***	4.43***	3.42***	3.34***	2.76**	4.38***	6.13***	7.23***	4.07**	3.27*
	(4.21)	(3.05)	(3.46)	(2.83)	(2.60)	(3.17)	(5.20)	(5.33)	(2.53)	(1.75)
$\text{Mkt}^{NonFin} - \text{Rf}$	0.017	0.046	-0.025	0.036	0.19***	0.14**	-0.11	-0.14*	0.016	0.070
	(0.41)	(0.86)	(-0.58)	(0.73)	(4.09)	(2.50)	(-1.47)	(-1.84)	(0.32)	(1.09)
Additional Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Observations	150	150	150	150	150	150	150	150	144	144
R^2	0.13	0.20	0.072	0.15	0.27	0.40	0.092	0.19	0.077	0.16

This table contains predictability regressions of high minus low excess returns for portfolios formed on the top and bottom quintiles of intermediation measure $\epsilon_{i,t}$ on risk factors

$$R^{Q5}_{i,t+1} - R^{Q1}_{i,t+1} = \alpha + \beta_1 \text{Intermediary Shock}_{t+1} + \beta_2(\text{Mkt}^{NonFin}_{t+1} - R_{f,t}) + \nu_t \tag{3.27}$$

The first two columns estimate the residual intermediation $\epsilon_{i,t}$ as done throughout the main text (and decribed in section 3.3.1); the next two add 40 financial ratios obtained from WRDS to the cross-sectional regression (3.9) from the main text; columns (5) and (6) include just a second degree polynomial in log book equity to estimate $\epsilon_{i,t}$. Columns (7)/(8) and (9)/(10) are, respectively, for value-weighted instead of equal-weighted portfolios and for a subsample that excludes the financial crisis (2008q1 through 2009q2). Odd columns control just for a version of the value-weighted market factor that excludes returns on financial stocks and even columns add the Fama-French (2015) non-market risk factors plus the momentum factor. The Intermediary Shock measure is formed as an average of the standardized shocks to primary dealer equity capital from He, Kelly, and Manela (2017) and broker-dealer leverage from Adrian, Etula, and Muir (2014). The Intermediary Shock measure is standardized and returns are expressed in annualized percent form. Newey-West t-stats are in parentheses. Sample spans 1980q2 to 2017q3.

Table 3.12: Robustness: Predictive Portfolio Regressions

	Original	Add WRDS Ratios	Just log(BE)	Value-Weighted	Drop Crisis
η	6.236***	5.535***	7.865***	5.604**	6.961***
	(3.93)	(3.57)	(4.62)	(2.56)	(3.10)
P/E	4.460*	3.753	5.810***	3.110	5.079*
	(1.90)	(1.58)	(2.91)	(1.04)	(1.79)
cay	0.658	0.862	1.700	-0.036	0.647
	(0.77)	(1.07)	(1.45)	(-0.03)	(0.74)
10Y-3Mo	0.735	0.304	1.513	1.415	0.737
	(0.80)	(0.47)	(1.36)	(0.98)	(0.78)
sentiment	1.726	1.479	0.394	3.655**	1.919*
	(1.56)	(1.40)	(0.33)	(2.50)	(1.67)
Observations	450	450	450	450	432
R^2	0.087	0.096	0.083	0.069	0.075

This table shows predictability regressions of high minus low excess returns for portfolios formed on the top and bottom quintiles of intermediation measure $\epsilon_{i,t}$ on on η, the average of the standardized primary dealer squared leverage from He, Kelly, and Manela (2017) and the negative of standardized broker dealer leverage from Adrian, Etula, and Muir (2014):

$$R^{Q5}_{t\to t+3} - R^{Q1}_{t\to t+3} = \alpha + \beta_1 \eta_t + \beta_2 Z_t + \nu_t \tag{3.28}$$

Regressions are for overlapping quarterly returns at the monthly frequency. The first column contains the original version for backing out residual intermediation $\epsilon_{i,t}$ (as used throughout the main text and described in 3.3.1), while the second column estimates the residual intermediation $\epsilon_{i,t}$ by adding 40 financial ratios obtained from WRDS to the cross-sectional regression (3.9) using to obtain $\epsilon_{i,t}$; the third column includes just a second degree polynomial in log book equity to estimate $\epsilon_{i,t}$. The last two columns are, respectively, for value-weighted instead of equal-weighted portfolios and for a subsample that excludes the financial crisis (2008m1 through 2009m6). Newey-West t-stats with four lags are in parentheses. All independent variables are standardized and returns are in annualized percent form. Sample spans 1980m4 to 2017m9.

3.6 Characteristics-Based Framework For Empirical Tests

I present here a simple setting (which borrows heavily but is slightly different from the characteristics-based demand setup of Koijen and Yogo, 2019) that leads to the empirical specification for backing out the residual intermediation measure $\epsilon_{i,t}$ that I use to form portfolios. As in the model from section 3.1, assume there are two investors, a representative institutional investor and a representative household with constant-absolute risk aversion utility and respective risk tolerance ρ_I and ρ_H (where for now I have suppressed any dependence of risk tolerances on underlying state variables). Assume that there are N assets in net supply 1 whose cashflows are distributed multivariate normal, $D \sim N(\mu, \Sigma)$. Similarly to Koijen and Yogo (2019), I assume that Σ can be decomposed as $\Sigma = \beta\beta' + \sigma^2 I$, where β contains asset factor loadings, σ^2 is idiosyncratic variance, and β is of dimension $N \times 1$. There is also a risk-free asset whose gross return R_f is fixed exogenously. Let X be a $N \times k$ matrix of stock characteristics. I assume that the representative household and institutional investor agree that

$$\beta = X\Pi + \pi \tag{3.29}$$

where Π is a $k \times 1$ vector and π is a constant $N \times 1$ vector. Hence fundamental loadings β are affine in characteristics.[22]

Now, assume that μ is linear in characteristics, but households and institutional investors may disagree on the mapping from characteristics to μ in the following manner. Households mistakenly believe that the mean μ follows

$$\mu_H = X\Phi_H + \phi_H + \epsilon_H \tag{3.30}$$

while institutional investors' (correct) estimate of the mean μ is given by

$$\mu_I = X\Phi_I + \phi_I \tag{3.31}$$

Here ϕ_H and ϕ_I are constant across assets and ϵ_H may differ across assets. The residual ϵ_H is the component of household's beliefs about the mean of the asset payoff distribution that are uncorrelated with the asset characteristics.

[22] Since any multifactor model of payoffs/returns implies a single factor model where the stochastic-discount factor is the lone factor, this essentially assumes that loadings on the SDF are affine in characteristics.

Given constant absolute risk aversion utility, the optimal demand for agent j is

$$\theta_j = \rho_j \Sigma^{-1}(\mu_j - R_f P) \tag{3.32}$$

Imposing market clearing ($\theta_I + \theta_H = 1$) gives the following expression for prices:

$$P = \frac{\rho_I \mu_I + \rho_H \mu_H - \Sigma \mathbf{1}}{R_f(\rho_I + \rho_H)} \tag{3.33}$$

Substituting out price using market clearing gives the following for intermediary demand (or percent intermediated):

$$\begin{aligned}
\theta_I &= \rho_I \Sigma^{-1} \left[\frac{\rho_H(\mu_I - \mu_H) + \Sigma \mathbf{1}}{\rho_I + \rho_H} \right] && (3.34)\\
&= \alpha \left(\beta\beta' + \sigma^2 I \right)^{-1} (X\Delta\Phi + \Delta\phi - \epsilon_H) + \delta \mathbf{1} && (3.35)\\
&= \frac{\alpha}{\sigma^2} \left(I + \frac{1}{\kappa}\beta\beta' \right) (X\Delta\Phi + \Delta\phi - \epsilon_H) + \delta \mathbf{1} && (3.36)\\
&= \frac{\alpha}{\sigma^2} \left(X\Delta\Phi + \Delta\phi - \epsilon_H + (X\Pi + \pi)\eta \right) + \delta \mathbf{1} && (3.37)\\
&= \frac{\alpha}{\sigma^2}(\Delta\phi + \pi\eta) + \delta\mathbf{1} + X\frac{\alpha}{\sigma^2}\left(\Delta\Phi + \Pi\eta\right) - \frac{\alpha}{\sigma^2}\epsilon_H && (3.38)\\
&\equiv a + XB + \tilde{\epsilon} && (3.39)
\end{aligned}$$

Where the terms in the above are defined as follows:

$$\alpha = \frac{\rho_I \rho_H}{\rho_I + \rho_H}, \ \delta = \frac{\rho_I}{\rho_I + \rho_H}, \ \kappa = -(\sigma^2 + \beta'\beta), \tag{3.40}$$

$$\Delta\Phi = \Phi_I - \Phi_H, \ \Delta\phi = \phi_I - \phi_H, \ \eta = \frac{1}{\kappa}\beta'(X\Delta\Phi + \Delta\phi - \epsilon_H), \tag{3.41}$$

$$B = \frac{\alpha}{\sigma^2}(\Pi\eta + \Delta\Phi), \ a = \frac{\alpha}{\sigma^2}(\Delta\phi + \pi\eta) + \delta\mathbf{1}, \text{ and } \tilde{\epsilon} = -\frac{\alpha}{\sigma^2}\epsilon_H \tag{3.42}$$

The relation between the second and third lines follows from the Woodbury matrix identity and then simplifying. Note that the constant η is obtained by multiplying β by X and ϵ_H, the current characteristics of all assets and the residual component of the household's estimate of the mean for all assets. The constants α and δ also depend on the current risk tolerance of the agents in the model. Hence the parameters in (3.39) can only be identifed with time-specific coefficients, which implies a cross-sectional regression as in (3.9) in the main text:

$$\text{Percent Intermediated}_{i,t} = \alpha_t + \beta_t X_{i,t} + \epsilon_{i,t} \tag{3.43}$$

I now show that under the assumptions above the residual $\epsilon_{i,t}$ recovers a component of intermediary demand along which the price response to intermediary risk tolerance shocks is strictly increasing. Returning to the equation for prices:

$$P = \frac{\rho_I \mu_I + \rho_H \mu_H - \Sigma \mathbf{1}}{R_f(\rho_I + \rho_H)} \tag{3.44}$$

$$= \frac{\rho_I(\omega)(X\Phi_I + \phi_I) + \rho_H(\zeta)(X\Phi_H + \phi_H + \epsilon_H) - \Sigma \mathbf{1}}{R_f(\rho_I(\omega) + \rho_H(\zeta))} \tag{3.45}$$

Now, letting ρ_I depend on the state variable ω and ρ_H on the state variable ζ as before, we can take the total derivative of price with respect to a local shock to these variables:

$$dP = \frac{\rho_I^{'}(\omega)\left(\rho_H(\zeta)(\Delta\Phi X + \Delta\phi - \epsilon_H) + \Sigma \mathbf{1}\right)}{R_f(\rho_I(\omega) + \rho_H(\zeta))^2} d\omega \tag{3.46}$$

$$- \frac{\rho_H^{'}(\zeta)\left(\rho_I(\omega)(\Delta\Phi X + \Delta\phi - \epsilon_H) + \Sigma \mathbf{1}\right)}{R_f(\rho_I(\omega) + \rho_H(\zeta))^2} d\zeta \tag{3.47}$$

$$\equiv \beta_\omega d\omega + \beta_\zeta d\zeta \tag{3.48}$$

Note that since we assume $\rho_I^{'}(\omega) > 0$, β_ω is strictly decreasing in ϵ_H, or equivalently is strictly increasing in $\tilde{\epsilon} = -\frac{\alpha}{\sigma^2}\epsilon_H$. Since the residuals $\epsilon_{i,t}$ in the regression equation (3.9) are analogous to $\tilde{\epsilon}$ in this setup, this implies that sorting on $\epsilon_{i,t}$ should induce variation in betas on proxies for shocks to intermediary risk tolerance. Note also that in the case where $\Phi_I = \Phi_H$ and $\phi_I = \phi_H$, so that $\Delta\Phi = 0$ and $\Delta\phi = 0$, we recover the expressions in section 3.1 by defining $\epsilon_H = \lambda$.

This setting also recovers the differential return predictability for high ϵ_H assets as in proposition 2. Define the risk premium on asset j by $E[R_{p,j}] = \mu_j - R_f P_j$, and suppose $X_2 = X_1$, so that asset characteristics are the same, but $\epsilon_{H,1} < \epsilon_{H,2}$ (or equivalently, $\tilde{\epsilon}_1 > \tilde{\epsilon}_2$, so that asset 1 is more intermediated) . Then

$$E[R_{p,1} - R_{p,2}] = \frac{\rho_H(\zeta)(\epsilon_{H,2} - \epsilon_{H,1}))}{R_f(\rho_I(\omega) + \rho_H(\zeta))} \tag{3.49}$$

which is positive and strictly decreasing in ω. Hence $\partial E[R_{p,1} - R_{p,2}]/\partial\omega < 0$ and the difference in expected returns for high minus low intermediated assets decreases (increases) when intermediaries are more (less) risk tolerant, as in proposition 2, implying that empirical proxies for current intermediary risk tolerance should negatively predict the return spread for high minus low $\epsilon_{i,t}$ assets.

3.7 Model Extension

Consider the following extension on the model from section 3.1–suppose that household risk tolerance ρ_H is a function of both the state variable ζ, which does not move intermediary risk tolerance, and ω, which does induce changes in intermediary risk tolerance. Then for local changes in ω and ζ

$$\mathrm{d}P = \frac{\rho_I^{'}(\omega)(\Sigma\mathbf{1} - \lambda\rho_H(\zeta)) + \rho_{H_\omega}(\zeta,\omega)(\Sigma\mathbf{1} + \lambda\rho_I(\omega))}{R_f(\rho_I(\omega) + \rho_H(\zeta,\omega))^2}\mathrm{d}\omega + \frac{\rho_{H_\zeta}(\zeta,\omega)(\Sigma\mathbf{1} + \lambda\rho_I(\omega))}{R_f(\rho_I(\omega) + \rho_H(\zeta,\omega))^2}\mathrm{d}\zeta \tag{3.50}$$

Proposition 3 follows easily from here. As proposition 3 assumes the partial derivative $\rho_{H_\omega}(\zeta,\omega) > 0$, then since $\rho_I^{'}(\omega)$ is multiplied by λ with negative sign and $\rho_{H_\omega}(\zeta,\omega)$ is multiplied by λ with positive sign, the two effects work in opposite direction for the coefficient on $d\omega$. Moreover, as percent intermediated is strictly decreasing in λ, the negative sign on $\rho_I^{'}(\omega)$ causes betas on shocks to ω to increase with intermediation, while $\rho_{H_\omega}(\zeta,\omega)$ does the opposite. Therefore if betas increase with intermediation (holding all else constant), it must be because of price responses to changes in intermediary risk tolerance.

It should be noted that this doesn't have to be the case if $\rho_{H_\omega}(\zeta,\omega) < 0$; however, it seems highly unlikely in practice that shocks to household and intermediary risk tolerance are negatively correlated. Indeed Haddad and Muir (2018) argue that if anything $\rho_{H_\omega}(\zeta,\omega) \geq 0$, as episodes where intermediaries become more risk averse are also likely to be periods of time where household risk aversion increases (the financial crisis of 2008-2009 being a particularly salient example).

3.8 Data Appendix

3.8.1 Construction of AEM Leverage Factor

As noted by Cho (2019), changes to the Federal Reserve Flow of Funds data have significantly altered the implied broker-dealer leverage ratio. Starting with the first quarter of 2014, repo assets (reverse repo) are included in assets and just repo liabilities, rather than net repo, are included in the liabilities section. In order to make my leverage factor consistent with the construction in the original Adrian, Etula, and Muir (2014) paper, I obtain the broker-dealer leverage from Table L128 of the 2013q4 Flow of Funds release. I then compute the leverage as

$$\text{Leverage}_t = \frac{\text{Total Financial Assets}_t}{\text{Total Financial Assets}_t - \text{Total Financial Liabilities}_t} \tag{3.51}$$

I then seasonally adjust as described in Adrian, Etula, and Muir (2014). Cho (2019) suggests that the following change allows one to extend the original AEM factor

$$\text{Leverage}_t = \frac{\text{Total Financial Assets}_t - \text{Repo Assets}_t}{\text{Total Financial Assets}_t - \text{Total Financial Liabilities}_t - \text{FDI in US}_t} \tag{3.52}$$

This accounts for changes to foreign direct investment reflected in liabilities in later releases of the Flow of Funds. However, I find that when I use the above for the most recent releases, the two methods (3.51) and (3.52) agree until the end of 2010 at which point broker dealer leverage begins an upward spike for (3.52) relative to (3.51), which spike becomes extreme to the point that leverage becomes negative towards the end of the sample. Due to this issue, I simply use (3.51) through the 2013q4 release and then extend the series using (3.51) with updated Flow of Funds data, which is also consistent with the extended leverage factor data posted on Tyler Muir's website. I further seasonally adjust the leverage growth series using expanding window regressions of leverage growth on quarterly dummies as in AEM to arrive at my final leverage factor.

3.8.2 Selection of WRDS Ratios For Final Sample

For my robustness checks in section 3.3.4 I obtain the 73 financial ratios from the Wharton Research Data Services Financial Ratios suite. I find that data availability are sparse, so I do the following:

1. When firm dividend yield and dividend/price ratios are missing, I assume they are equal to zero
2. I replace missing values for any variables with their lags as of up to 8 quarters previous
3. I then check the fraction of missing observations for stocks that overlap with my main sample. If this fraction is greater than 1% I exclude the ratio from the analysis.

This leaves the following ratios:
Enterprise Value Multiple, Price/Sales, Price/Cash flow, Dividend Payout Ratio, Net Profit

Margin, Operating Profit Margin Before Depreciation, Operating Profit Margin After Depreciation, Gross Profit Margin, Pre-tax Profit Margin, Cash Flow Margin, Return on Assets, Return on Equity, Return on Capital Employed, After-tax Return on Average Common Equity, After-tax Return on Invested Capital, After-tax Return on Total Stockholders Equity, Gross Profit/Total Assets, Common Equity/Invested Capital, Long-term Debt/Invested Capital, Total Debt/Invested Capital, Capitalization Ratio, Cash Balance/Total Liabilities, Total Debt/Total Assets, Total Debt/EBITDA, Long-term Debt/Total Liabilities, Cash Flow/Total Debt, Total Liabilities/Total Tangible Assets, Long-term Debt/Book Equity, Total Debt/Total Assets, Total Debt/Capital, Total Debt/Equity, Asset Turnover, Sales/Invested Capital, Sales/Stockholders Equity, Research and Development/Sales, Advertising Expenses/Sales, Labor Expenses/Sales, Accruals/Average Assets, Price/Book, and Dividend Yield.

Though the WRDS book/market ratio satisfies sampling criteria, I also exclude this variable because I already include a version of book/market in the regression. Finally, these variables are winsorized cross-sectionally at the 1% level to deal with outliers.