

# Sources of Increasing Earnings Inequality: Reconciling Survey and Administrative Data

John Haltiwanger<sup>1</sup>   Henry Hyatt<sup>2</sup>   James Spletzer<sup>3</sup>

<sup>1</sup>Department of Economics, University of Maryland

<sup>2</sup>Center for Economic Studies, U.S. Census Bureau

<sup>3</sup>Retired

NBER Conference on Wage Dynamics in the 21st Century

September 16, 2022

Opinions expressed herein are those of the authors alone and do not reflect the views of the U.S. Census Bureau.

All results have been reviewed to ensure that no confidential data are disclosed (CBDRB-FY22-159, CBDRB-FY22-311).

We would especially like to thank Thomas Lemieux for many helpful discussions about the replication files from his *Journal of Economic Perspectives* article with Florian Hoffmann and David Lee “Growing Income Inequality in the United States and Other Advanced Economies,” and we are also grateful for the easy-to-use replication files these authors provided on the *JEP* website.

We compare two very different perspectives on rising earnings inequality.

A large literature using household survey data emphasizes:

- ▶ rising dispersion across education and occupation groups.
- ▶ industry effects are modest or offsetting.

A more recent literature using matched employer-employee admin data emphasizes:

- ▶ rising dispersion between firms (and industries)
- ▶ rising between firm and industry dispersion is accounted for by sorting and segregation.

Our analysis uses linked micro admin and survey data to reconcile these approaches.

# Industries and increasing inequality

What is the role of industry in increasing inequality?

### **Administrative records data**

*Most of the rise in overall earnings inequality is accounted for by rising between-industry inequality.*

- ▶ Haltiwanger, Hyatt, and Spletzer (2022)

*The dominant driver of the rising inequality of both earnings and wage rates in Italy is the growing heterogeneity of pay across industries.*

- ▶ Briskar, Di Porto, Rodriguez Mora, and Tealdi (2022)

### **Survey data**

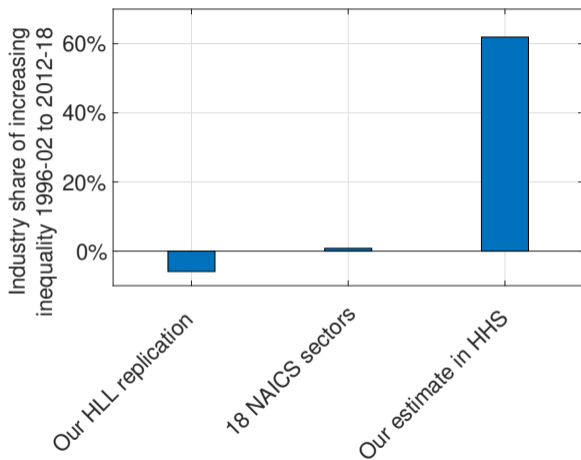
*The between-group variance component linked to industry has been declining over time.*

- ▶ Hoffmann, Lee, and Lemieux (2020)

*Using the CPS, we show that since the 1980s there has been a decline of about one third in the dispersion of industry wage premia.*

- ▶ Stansbury and Summers (2020)





- ▶ We conducted a number of exercises to compare HLL with our findings in HHS.
- ▶ Our first exercise concerns the industry classification method.
- ▶ In the late 1990s, the North American Industrial Classification System (NAICS) replaced the Standard Industrial Classification (SIC)
- ▶ Using HLL's method (next slide) and 18 NAICS sectors yields an industry contribution of 0.8% rather than the -5.9% when using 12 SIC aggregate industries

**Our replication of HLL:** For the seven-year intervals {1996-02, 2012-18}, we estimate a human capital earnings ( $y_i$  for worker  $i$ ) equation in three steps:

$$y_i = \text{AgeEduc}_i \beta_1 + \varepsilon_i \quad (1)$$

$$y_i = \text{AgeEduc}_i \beta_1 + \text{Occupation}_i \beta_2 + \varepsilon_i \quad (2)$$

$$y_i = \text{AgeEduc}_i \beta_1 + \text{Occupation}_i \beta_2 + \text{Industry}_i \beta_3 + \varepsilon_i \quad (3)$$

Estimating equation (1) provides the  $R^2$  from including age and education alone. The contribution of occupation is the  $R^2$  from equation (2) minus the  $R^2$  from (1). The contribution of industry is the  $R^2$  from equation (3) minus the  $R^2$  from (2).

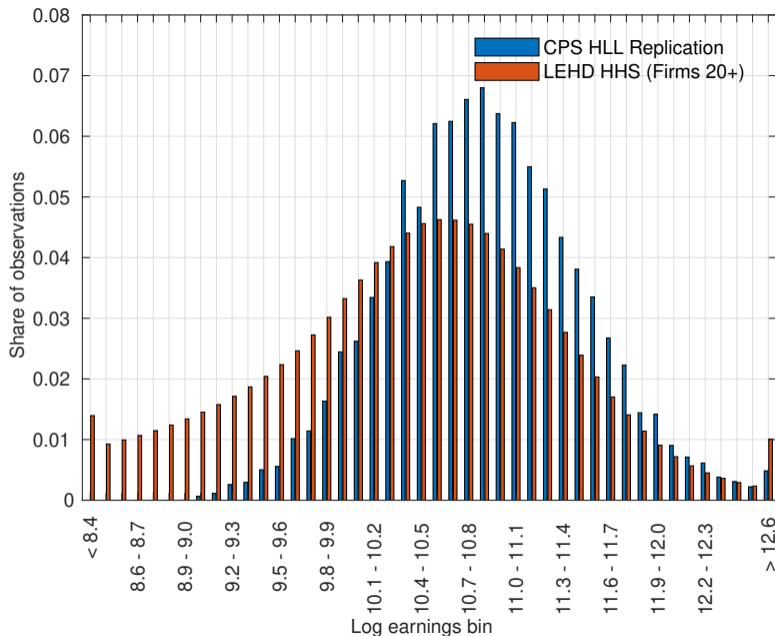
**Our between-industry in HHS:** We use the simple decomposition

$$\underbrace{\text{var}(y_{i,k} - \bar{y})}_{\text{earnings variance}} = \underbrace{\text{var}(y_{i,k} - \bar{y}_k)}_{\text{within-industry dispersion}} + \underbrace{\text{var}(\bar{y}_k - \bar{y})}_{\text{between-industry dispersion}} \quad (4)$$

where  $\bar{y}_k$  is average earnings in industry  $k$





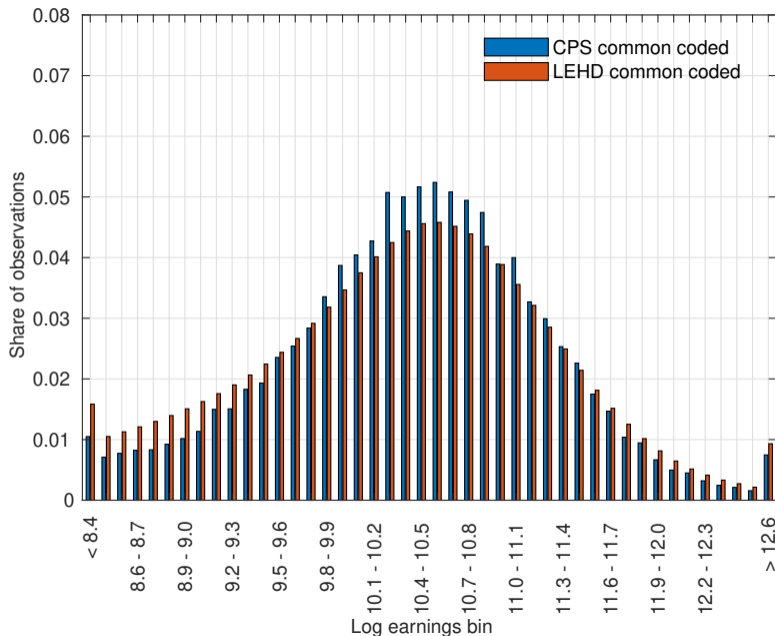


**HLL:** annual real earnings > \$7840, and:

- ▶ weeks worked > 49
- ▶ usual hours  $\geq 40$
- ▶ real hourly wage > \$4

**HHS:** annual earnings > \$3770

There are **additional differences** in which ages, job types, etc. are included.



## Common coding applied to both datasets

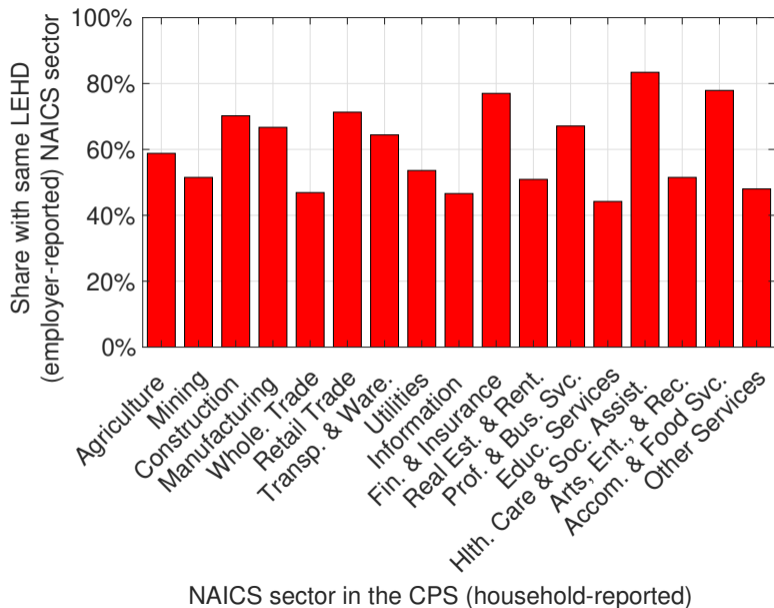
Annual real earnings > \$3770

Other common coding:

- ▶ Exclude self-employed
- ▶ Age 20-60
- ▶ Exclude longest job if government
- ▶ Use PCE (2013=100)
- ▶ Any firm size







In our linked CPS-LEHD dataset, at the NAICS sector level, there is less than 50% agreement in:

- ▶ Wholesale Trade (46.9%)
- ▶ Information (46.6%)
- ▶ Educational Services (44.6%)
- ▶ Other Services (48%)







# Pay premia vs. between-industry differences: the roles of sorting and segregation

Song, Price, Guvenen, Bloom, and von Wachter (2019) build on Abowd, Kramarz, and Margolis (1999) and Card, Heining, and Kline (2013) to **measure between-firm inequality** in terms of the following:

- ▶ **Pay premia:** some firms offer greater earnings to any worker
- ▶ **Sorting:** high-paying firms employ more highly paid workers
- ▶ **Segregation:** highly paid workers concentrate among each other

In HHS, we extend the Song et al. (2019) framework in order to measure how these components of inequality **occur between vs. within industries**

In this paper, we show how to apply this method to the CPS to measure industry-level sorting and segregation by age, education, and occupation

**Pay premia vs. between-industry:** We re-write the human capital earnings equation used by HLL (introducing a subscript for industry  $k$ ) as

$$y_{i,k} = Z_{i,k}\beta_Z + \text{Industry}_{i,k}\beta_3 + \varepsilon_{i,k}, \quad (5)$$

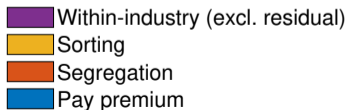
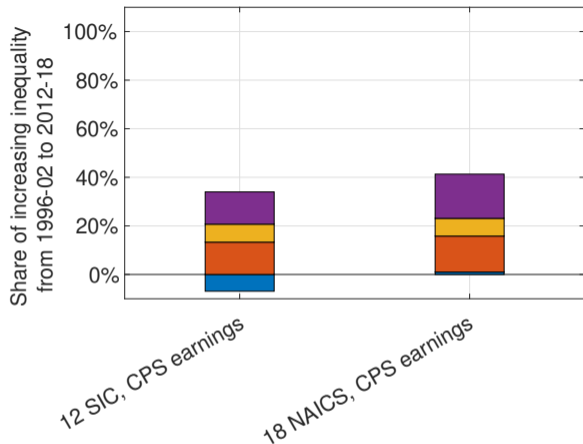
where  $Z$  concatenates the  $\text{AgeEduc}_i$  and  $\text{Occupation}_i$  vectors, and  $\beta_Z$  concatenates the marginal effects vectors  $\beta_1$  and  $\beta_2$ .

Define  $\overline{Z_k\beta_Z}$  as the industry mean of  $Z_{i,k}\beta_Z$ .

Taking variances of both sides of the human capital earnings equation results in:

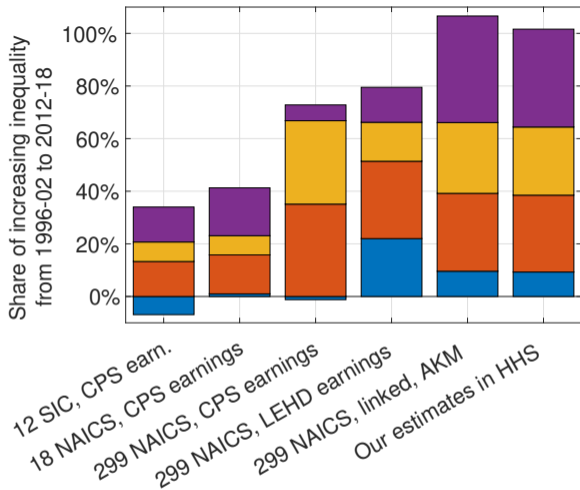
$$\underbrace{\text{var}(y_{i,k})}_{\text{earnings variance}} = \underbrace{\text{var}(Z_{i,k}\beta_Z - \overline{Z_k\beta_Z})}_{\text{within-industry dispersion from age, education, and occupation}} + \underbrace{\text{var}(\overline{Z_k\beta_Z})}_{\text{between-industry segregation}} + \underbrace{\text{var}(\text{Industry}_{i,k}\beta_3)}_{\text{between-industry pay premium}} + \underbrace{2\text{cov}(\overline{Z_k\beta_Z}, \text{Industry}_{i,k}\beta_3)}_{\text{between-industry sorting}} + \underbrace{\text{var}(\varepsilon_{i,k})}_{\text{residual dispersion (within-industry)}} \quad (6)$$

## Using the CPS alone



- ▶ We start by applying our decomposition to our replication of the HLL analysis dataset
- ▶ Pay premium explains -6.9% of the increase using 12 SIC aggregates, 1.0% using 18 NAICS sectors
- ▶ Segregation explains 13.3%-14.8%
- ▶ Sorting explains 7.3%-7.4%
- ▶ Within-industry dispersion by age, education, & occupation explains 13.3% (18.2%) using 12 SIC aggregates (18 NAICS sectors).





## If we estimate AKM and implement our decomposition from HHS

- ▶ Using person and firm effects rather than age, education, & occupation
- ▶ Pay premium explains 9.3%-9.6%
- ▶ Segregation explains 29.2%-29.6%
- ▶ Sorting explains 25.9%-26.9%
- ▶ Within-industry worker and firm effects explain 37.2%-40.5%
- ▶ Small offsetting effect of residual

# Industries and occupations

We use public domain data from the Occupational Employment and Wage Statistics (OEWS) to construct a dataset of 287 4-digit NAICS industries  $k$  by 22 occupations  $j$ .

We estimate the following equation for the intervals 2002-03 and 2015-16:

$$y_{j,k} = \text{Occupation}_{j,k}\beta_2 + \text{Industry}_{j,k}\beta_3 + \varepsilon_{j,k}. \quad (7)$$

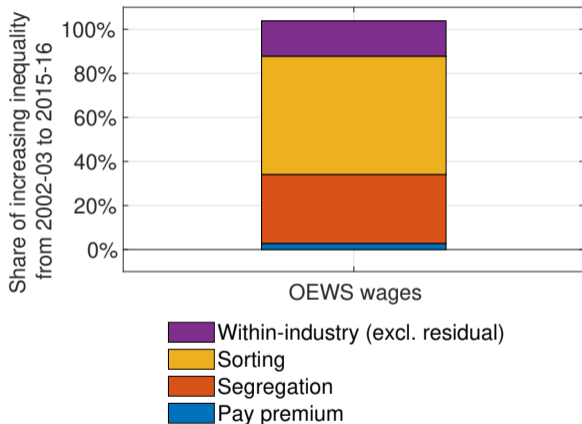
Taking (employment-weighted) variances of both sides of equation (7) yields:

$$\underbrace{\text{var}(y_{j,k})}_{\text{earnings variance}} = \underbrace{\text{var}(\text{Occupation}_{j,k}\beta_2 - \overline{\text{Occupation}_k\beta_2})}_{\text{within-industry dispersion from occupation}} + \underbrace{\text{var}(\overline{\text{Occupation}_k\beta_2})}_{\text{between-industry segregation}} + \underbrace{\text{var}(\text{Industry}_{j,k}\beta_3)}_{\text{between-industry pay premia}} + \underbrace{2\text{cov}(\overline{\text{Occupation}_k\beta_2}, \text{Industry}_{j,k}\beta_3)}_{\text{between-industry sorting}} + \underbrace{\text{var}(\varepsilon_{j,k})}_{\text{residual dispersion (within-industry)}} \quad (8)$$



## Using OEWS aggregates

- ▶ 287 NAICS industries
- ▶ Pay premium explains 2.8%
- ▶ Segregation explains 31.3%
- ▶ Sorting explains 53.7%
- ▶ Within-industry occupation effects explain 16.0%
- ▶ Small offsetting effect of residual
- ▶ Between-industry occupation changes also dominate within-industry **in the CPS**
- ▶ Sorting and segregation reflect changes in the way workers are allocated across industries



# Conclusion

Findings from the CPS, OEWS, and our administrative records data:

1. A large share of inequality growth in recent decades has occurred at the industry level
2. The role of industry pay premia in increasing inequality is relatively small
3. Sorting and segregation are of first order importance when assessing the role of industries in inequality growth
  - ▶ whether we measure sorting and segregation using education, occupation, or AKM worker effects

# Appendix





Recall our formula for between- vs. within-industry variance:

$$\underbrace{\text{var}(y_{i,k} - \bar{y})}_{\text{earnings variance}} = \underbrace{\text{var}(y_{i,k} - \bar{y}_k)}_{\text{within-industry dispersion}} + \underbrace{\text{var}(\bar{y}_k - \bar{y})}_{\text{between-industry dispersion}}$$

The between-industry component can be computed for published aggregates of industry-level average earnings. We use the Quarterly Census of Employment and Wages (QCEW). The 50 state between-industry variance growth is lower than that of our 18 states in both the CPS and the QCEW, but this ratio is lower in the CPS (70.5%) than in the QCEW (79.6%).

[Back](#)

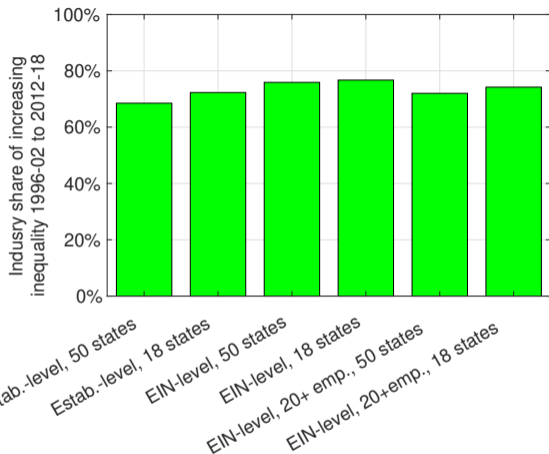
$$\underbrace{\text{var}(y_{i,k} - \bar{y})}_{\text{earnings variance}} = \underbrace{\text{var}(y_{i,k} - \bar{y}_k)}_{\text{within-industry dispersion}} + \underbrace{\text{var}(\bar{y}_k - \bar{y})}_{\text{between-industry dispersion}}$$

Differences between the CPS and QCEW are driven by two industries:

Data	CPS (Micro)		CPS (Agg)		QCEW (Agg)	
	50 states	18 states	50 states	18 States	50 States	18 States
Contribution to variance growth from 1996-02 to 2012-18:						
Retail Trade	0.0035	0.0054	0.0024	0.0041	0.0066	0.0078
Information	0.0054	0.0082	0.0050	0.0076	0.0030	0.0038
Ratio of 50 State to 18 State (%):						
Retail Trade	64.8%		59.3%		84.8%	
Information	65.9%		65.8%		78.7%	

[Back](#)



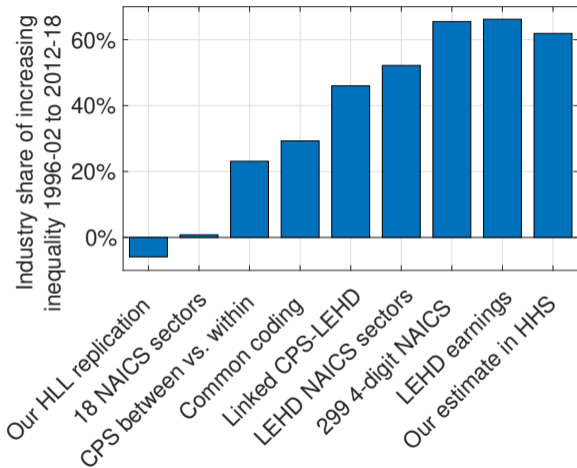


Use the Longitudinal Business Database to measure the share of inequality growth that is within-industry

Average earnings at the establishment- or EIN-level, so omits variation between workers at the same employer

18 states have an industry share that is 0.8 to 3.4 percentage points higher than the national average

[Back](#)



- ▶ Replacing CPS earnings with LEHD earnings yields a similar between-industry share of the increase in earnings dispersion (66.2% vs. 65.5%)
- ▶ This is despite the higher variance of LEHD earnings (relative to CPS earnings), even after common coding and linking



Following Abowd, Kramarz, and Margolis (1999, AKM), assume that earnings  $y_t^{i,j,k,p}$  is the sum of the worker  $i$  effect  $\theta^{i,p}$ , a firm  $j$  in industry  $k$  effect  $\psi^{j,k,p}$ , and observable characteristics  $X_t^{i,p}$  (marginal effects  $\beta^p$ ).

$$y_t^{i,j,k,p} = X_t^{i,p} \beta^p + \theta^{i,p} + \psi^{j,k,p} + \varepsilon_t^{i,j,k,p} \quad (9)$$

We estimate this AKM equation separately by interval  $p$ .

For the purposes of this presentation (on following slide), we omit the superscript  $p$  and the effects of observable characteristics  $X_t^{i,p} \beta^p$ .







Data	CPS	Linked CPS-LEHD	Linked CPS-LEHD
Sample	HLL JEP	Common coded	Common coded
Earnings measure	CPS	CPS	LEHD
Industry measure	CPS 18	LEHD 299	LEHD 299
Within-industry:			
Age, education & occupation:	18.2%	6.0%	13.3%
Age and education	11.8%	9.0%	14.4%
<b>Occupation</b>	<b>4.1%</b>	<b>0.4%</b>	<b>0.0%</b>
<b>Covariance: age+educ &amp; occ.</b>	<b>2.2%</b>	<b>-3.6%</b>	<b>-1.0%</b>
Residual	58.8%	28.5%	20.6%
Between-industry:			
Segregation	14.8%	31.7%	14.8%
Age and education	3.8%	9.6%	5.6%
<b>Occupation</b>	<b>2.5%</b>	<b>5.4%</b>	<b>2.0%</b>
<b>Covariance: age+educ. &amp; occ:</b>	<b>8.4%</b>	<b>16.7%</b>	<b>7.0%</b>
Pay premia	1.0%	-1.2%	22.0%
Sorting			
Covariance: age+educ. & ind.	7.3%	35.1%	29.4%
<b>Covariance: industry &amp; occ.</b>	<b>5.3%</b>	<b>15.5%</b>	<b>10.5%</b>

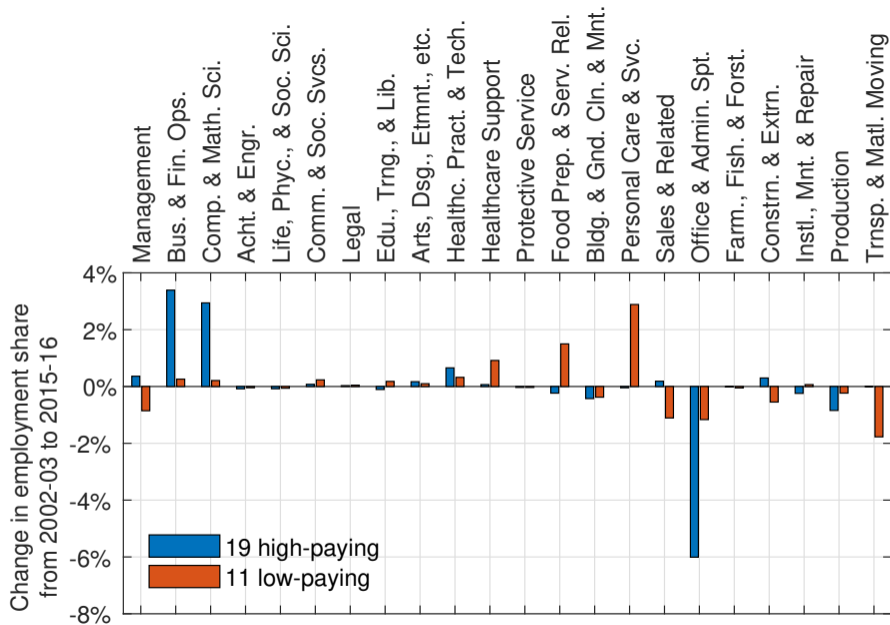
Change from 1996-02 to 2012-18

Within-industry occupation dispersion explains **-3.2%** to **6.3%** of the increase in earnings dispersion.

Between-industry sorting and segregation by occupation explains **16.2%** to **37.6%** of increasing dispersion

Most of the contribution of occupations to increasing inequality is through how they are allocated across industries





[Back](#)

## The top thirty industries

- ▶ **High-tech**: 11 of the 19 high-paying industries are high-tech in terms of STEM intensity as classified by Hecker (2005) and Goldschlag and Miranda (2016)
  - ▶ One-third of the increase in between-industry inequality
- ▶ **Mining**: 2 high-paying: oil and gas (also high-tech), drilling wells
- ▶ **Finance and Insurance**: 4 of the 19 high-paying industries
- ▶ **Management of Companies and Enterprises** : corporate headquarters
- ▶ **Health Care and Social Assistance**: 2 high-paying (physician offices, hospitals), 3 low-paying (in-home care, nursing homes, social services)
- ▶ **Support services**: 2 of the 11 low-paying industries
- ▶ **Retail, restaurants, and gyms**: 6 of the 11 low-paying industries
  - ▶ Another one-third of the increase in between-industry inequality



## High-tech: 11 of the 19 high-paying industries (II)

Industry title	Employment share:		Relative log earnings:		Share of bet.-ind. var. gth.
	average	change	average	change	
Other Information Services	0.2%	0.3%	0.798	0.699	5.8%
Architectur. & Enginr. Services	1.2%	0.1%	0.469	0.161	2.6%
Computer Systems Design	1.7%	0.9%	0.663	0.012	5.6%
Management & Scientific Serv.	0.9%	0.6%	0.381	0.069	1.8%
Scientific Research Services	0.8%	-0.1%	0.741	0.244	3.3%

*Notes:* Persons with annual real earnings > \$3770 in EINs with 20 or more employees. Average log earnings for industry  $k$  are relative to the economy average. The 1996-2002 and 2012-2018 intervals are averaged. Changes are the growth (or decline) from 1996-2002 to 2012-2018.

[Back](#)

## Mining: 2 of the 19 high-paying industries

Industry title	Employment share:		Relative log earnings:		Share of bet.-ind. var. gth.
	average	change	average	change	
Oil & Gas Extraction	0.3%	-0.1%	1.012	0.247	1.8%
Support Activities for Mining	0.5%	0.3%	0.374	0.191	1.4%

*Notes:* Persons with annual real earnings > \$3770 in EINs with 20 or more employees. Average log earnings for industry  $k$  are relative to the economy average. The 1996-2002 and 2012-2018 intervals are averaged. Changes are the growth (or decline) from 1996-2002 to 2012-2018.

[Back](#)





## Support services

Industry title	Employment share:		Relative log earnings:		Share of bet.-ind. var. gth.
	average	change	average	change	
Employment Services	3.9%	0.6%	-0.685	0.017	2.5%
Services to Buildings & Dwell.	1.1%	0.3%	-0.493	-0.002	1.1%

*Notes:* Persons with annual real earnings  $>$  \$3770 in EINs with 20 or more employees. Average log earnings for industry  $k$  are relative to the economy average. The 1996-2002 and 2012-2018 intervals are averaged. Changes are the growth (or decline) from 1996-2002 to 2012-2018.

[Back](#)



## Retail, restaurants, and gyms

Industry title	Employment share:		Relative log earnings:		Share of bet.-ind. var. gth.
	average	change	average	change	
Building Material & Supplies	0.9%	0.1%	-0.293	-0.180	1.5%
Grocery Stores	2.3%	0.0%	-0.378	-0.194	4.7%
Clothing Stores	0.7%	-0.0%	-0.607	-0.244	2.6%
Othr. Genrl. Merchandise Stores	1.4%	1.5%	-0.539	-0.051	6.8%
Othr. Amusement & Recreation	0.6%	0.1%	-0.594	-0.106	1.7%
Restaurants & Othr. Eat Places	4.9%	2.0%	-0.739	-0.027	16.9%

*Notes:* Persons with annual real earnings  $>$  \$3770 in EINs with 20 or more employees. Average log earnings for industry  $k$  are relative to the economy average. The 1996-2002 and 2012-2018 intervals are averaged. Changes are the growth (or decline) from 1996-2002 to 2012-2018.

[Back](#)