Authors: John Haltiwanger, Henry Hyatt, and James Spletzer

Appendices

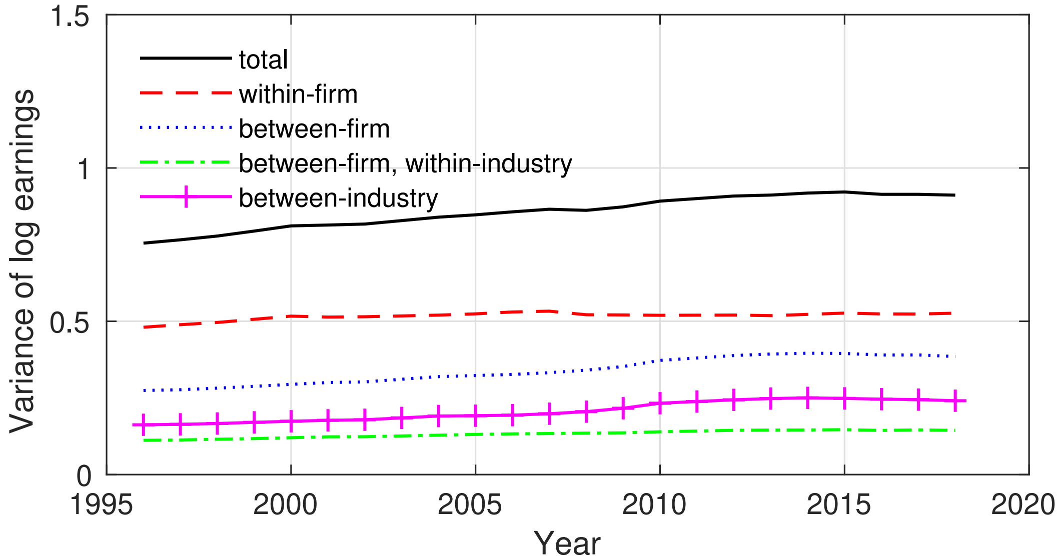
A Supplemental tables and figures

This appendix includes supplemental tables and figures for the results highlighted in the main text (Figures A1 to F6 and Tables A1 to A6).

Figure A1: Descriptive statistics

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > $3770 in EINs with 20 or more employees. 2003 and 2011 are linearly interpolated.

Figure A2: Variance decomposition by year, 1996-2018



Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > $3770 in EINs with 20 or more employees. 2003 and 2011 are linearly interpolated.

Table A1: Variance decomposition, following Song et al. (2019)

	Interval 1: 1996-2002	Interval 2: 2004-2010	Interval 3: 2012-2018	Growth: 1 to 3
Total var$(y_t^{i,j})$	0.794	0.862	0.915	0.121
Between-firm $(\bar{y}_t^{j,k} - \bar{y}_t)$	35.4%	38.3%	42.0%	85.1%
var$(\overline{\theta}^{j,k})$	10.4%	10.8%	12.2%	23.8%
var$(\psi^{j,k})$	7.3%	8.2%	7.9%	11.6%
var$(\overline{X}^{j,k}\beta)$	1.0%	0.9%	1.2%	2.7%
2cov$(\overline{\theta}^{j,k}, \psi^{j,k})$	11.7%	12.6%	13.8%	27.2%
2cov$(\overline{\theta}^{j,k}, \overline{X}^{j,k}\beta)$	2.3%	2.6%	3.3%	10.3%
2cov$(\overline{X}^{j,k}\beta, \psi^{j,k})$	2.7%	3.1%	3.6%	9.5%
Within-firm var$(y_t^{i,j,k} - \bar{y}_t^{j,k})$	64.6%	61.7%	58.0%	14.9%
var$(\theta^i - \overline{\theta}^{j,k})$	42.6%	40.8%	38.5%	11.7%
var$(X_t^i\beta - \overline{X}^{j,k}\beta)$	7.7%	5.8%	7.5%	6.3%
var$(\varepsilon_t^{i,j,k})$	16.1%	14.9%	13.5%	-3.5%
2cov$(\theta^i - \overline{\theta}^{j,k}, X_t^i\beta - \overline{X}^{j,k}\beta)$	-2.2%	-0.1%	-1.8%	0.7%
2cov$(\theta^i - \overline{\theta}^{j,k}, \varepsilon_t^{i,j,k})$	0.2%	0.2%	0.1%	-0.2%
2cov$(X_t^i\beta - \overline{X}^{j,k}\beta, \varepsilon_t^{i,j,k})$	0.1%	0.1%	0.1%	-0.1%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees. See discussion in text and Equation (5) for definitions.

Table A2: Variance decomposition, following Song et al. (2019), aggregated

	Interval 1: 1996-2002	Interval 2: 2004-2010	Interval 3: 2012-2018	Growth: 1 to 3
Total variance	0.794	0.862	0.915	0.121
Between-firm	35.4%	38.3%	42.0%	85.1%
Firm segregation	13.7%	14.4%	16.8%	36.8%
Firm pay premium	7.3%	8.2%	7.9%	11.6%
Firm sorting	14.4%	15.7%	17.4%	36.7%
Within-firm	64.6%	61.7%	58.0%	14.9%
Person effect	48.2%	46.5%	44.3%	18.8%
Residual	16.4%	15.2%	13.7%	-3.9%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees. See Equation (5) for definitions.

Table A3: Industry-enhanced variance decomposition, in detail

	Interval 1: 1996-2002	Interval 2: 2004-2010	Interval 3: 2012-2018	Growth: 1 to 3
Total var($y_t^{i,j,k}$)	0.794	0.862	0.915	0.121
Between-firm, within-industry $\bar{y}_t^{j,k} - \bar{y}_t^k$	14.0%	14.7%	15.3%	23.1%
var($\overline{\theta}^{j,k} - \overline{\theta}^k$)	5.2%	5.4%	5.7%	9.0%
var($\psi^{j,k} - \overline{\psi}^k$	3.1%	3.4%	3.1%	2.9%
var($\overline{X}^{j,k}\beta - \overline{X}^k\beta$)	0.6%	0.4%	0.6%	0.7%
2cov$[(\overline{\theta}^{j,k} - \overline{\theta}^k), (\psi^{j,k} - \overline{\psi}^k)$	3.9%	4.0%	4.3%	7.2%
2cov($\overline{\theta}^k$,$\overline{X}^k\beta$)	0.6%	0.8%	0.8%	2.0%
2cov$[(\psi^{j,k} - \overline{\psi}^k), (\overline{X}^{j,k}\beta - \overline{X}^k\beta)]$	0.7%	0.7%	0.8%	1.4%
Between-industry var($\bar{y}_t^k - \bar{y}_t$)	21.4%	23.6%	26.8%	61.9%
var($\overline{\theta}^k$)	5.3%	5.4%	6.5%	14.8%
var($\overline{\psi}^k$)	4.2%	4.8%	4.8%	8.7%
var($\overline{X}^k\beta$)	0.5%	0.5%	0.7%	2.1%
2cov($\overline{\theta}^k, \overline{\psi}^k$)	7.8%	8.6%	9.4%	19.9%
2cov($\overline{\theta}^k$,$\overline{X}^k\beta$)	1.6%	1.8%	2.5%	8.3%
2cov($\overline{\psi}^k$,$\overline{X}^k\beta$)	2.0%	2.4%	2.8%	8.1%
Within-firm var($y_t^{i,j,k} - \bar{y}_t^{j,k}$)	64.6%	61.7%	58.0%	14.9%
var($\theta^i - \overline{\theta}^{j,k}$)	42.6%	40.8%	38.5%	11.7%
var($X_t^i\beta - \overline{X}^{j,k}\beta$)	7.7%	5.8%	7.5%	6.3%
var($\varepsilon_t^{i,j,k}$)	16.1%	14.9%	13.5%	-3.5%
2cov($\theta^i - \overline{\theta}^{j,k}$,$X_t^i\beta - \overline{X}^{j,k}\beta$)	-2.2%	-0.1%	-1.8%	0.7%
2cov($\theta^i - \overline{\theta}^{j,k}, \varepsilon_t^{i,j,k}$)	0.2%	0.2%	0.1%	-0.2%
2cov($X_t^i\beta - \overline{X}^{j,k}\beta, \varepsilon_t^{i,j,k}$)	0.1%	0.1%	0.1%	-0.1%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings $>$ \$3770 in EINs with 20 or more employees.

Table A4: Industry-enhanced variance decomposition

	Interval 1: 1996-2002	Interval 2: 2004-2010	Interval 3: 2012-2018	Growth: 1 to 3
Total variance	0.794	0.862	0.915	0.121
Variance, as percent of total:				
Between-firm, within-industry	14.0%	14.7%	15.3%	23.1%
Firm segregation	6.3%	6.6%	7.0%	11.6%
Firm pay premium	3.1%	3.4%	3.1%	2.9%
Firm covariance	4.6%	4.7%	5.1%	8.6%
Between-industry	21.4%	23.6%	26.8%	61.9%
Industry segregation	7.4%	7.8%	9.7%	25.2%
Industry pay premium	4.2%	4.8%	4.8%	8.7%
Industry covariance	9.9%	11.0%	12.3%	28.0%
Within-firm	64.6%	61.7%	58.0%	14.9%
Person effect and observables	48.2%	46.5%	44.3%	18.8%
Residual	16.4%	15.2%	13.7%	-3.9%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > $3770 in EINs with 20 or more employees. See discussion in text and Equation (5) for definitions.

Table A5: Industry contributions to between-industry variance growth, top 10 and bottom 3 industries

4-digit		Relative earnings:		Employment share:		Bet.-ind.	Bet.-ind.
NAICS	Industry title	average	change	average	change	var. growth	var. share
7225	Restaurants and Other Eating Places	-0.739	-0.027	4.9%	2.0%	0.013	16.9%
4529	Other General Merchandise Stores	-0.539	-0.051	1.4%	1.5%	0.005	6.8%
5191	Other Information Services	0.798	0.699	0.2%	0.3%	0.004	5.8%
5415	Computer Systems Design	0.663	0.012	1.7%	0.9%	0.004	5.6%
5112	Software Publishers	1.009	0.186	0.5%	0.2%	0.004	5.6%
5511	Management of Companies	0.471	0.201	2.0%	-0.1%	0.004	5.0%
4451	Grocery Stores	-0.378	-0.194	2.4%	0.0%	0.004	4.7%
6221	General Medical & Surg. Hospitals	0.205	0.170	4.5%	0.5%	0.003	4.2%
6241	Individual and Family Services	-0.490	-0.155	0.8%	0.6%	0.003	3.5%
5239	Other Financial Invest. Activities	0.834	0.388	0.3%	0.1%	0.003	3.3%
6231	Skilled Nursing Care Facilities	-0.375	0.079	1.5%	-0.1%	-0.001	-1.5%
4521	Department Stores	-0.593	-0.142	1.6%	-1.1%	-0.001	-1.5%
3341	Computer Manufacturing	0.911	0.191	0.5%	-0.4%	-0.001	-1.6%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees. Average log earnings for industry k are relative to the economy average. Industry k's contribution to between-industry variance growth is in terms of Equation (2).

Table A6: Sources of industry contributions to between-industry variance growth, top 10 and bottom 3 industries

4-digit NAICS	Industry title	Bet.-ind. var. share	Segregation	Pay premia	Sorting	Shift share: earnings	Shift share: employment
7225	Restaurants and Other Eating Places	16.9%	40.3%	12.5%	47.1%	15.5%	84.5%
4529	Other General Merchandise Stores	6.8%	36.2%	16.3%	47.5%	15.0%	85.0%
5191	Other Information Services	5.8%	28.0%	22.3%	49.7%	51.5%	48.5%
5415	Computer Systems Design	5.6%	62.3%	1.7%	35.9%	6.5%	93.5%
5112	Software Publishers	5.6%	43.0%	11.4%	45.6%	45.5%	54.5%
5511	Management of Companies	5.0%	49.8%	8.4%	41.8%	103.5%	-3.5%
4451	Grocery Stores	4.7%	28.3%	21.3%	50.4%	99.5%	0.5%
6221	General Medical & Surg. Hospitals	4.2%	19.5%	28.7%	51.8%	92.9%	7.1%
6241	Individual and Family Services	3.5%	43.5%	11.5%	45.0%	45.8%	54.2%
5239	Other Financial Invest. Activities	3.3%	46.6%	10.0%	43.4%	64.4%	35.6%
6231	Skilled Nursing Care Facilities	-1.5%	39.7%	13.3%	47.0%	82.0%	18.0%
4521	Department Stores	-1.5%	24.3%	28.1%	47.6%	-242.4%	342.4%
3341	Computer Manufacturing	-1.6%	11.5%	34.8%	53.6%	-145.9%	245.9%

Notes: Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings $>$ \$3770 in EINs with 20 or more employees. Industry k's contribution to between-industry variance growth is in terms of Equation (2). The shift-share calculations follow Equation (3).

Table A7: Summary statistics on employment share and contribution to earnings inequality growth, LEHD

	Contribution	Initial emp. share
1st percentile	-1.4%	0.0%
10th percentile	-0.1%	0.0%
Median	0.1%	0.2%
90th percentile	1.0%	0.7%
99th percentile	5.6%	2.4%
Mean	0.3%	0.3%
Min	-1.6%	0.0%
Max	16.9%	4.2%
Skewness	6.9	4.2
Kurtosis	71.5	25.7

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees.

B Earnings percentiles

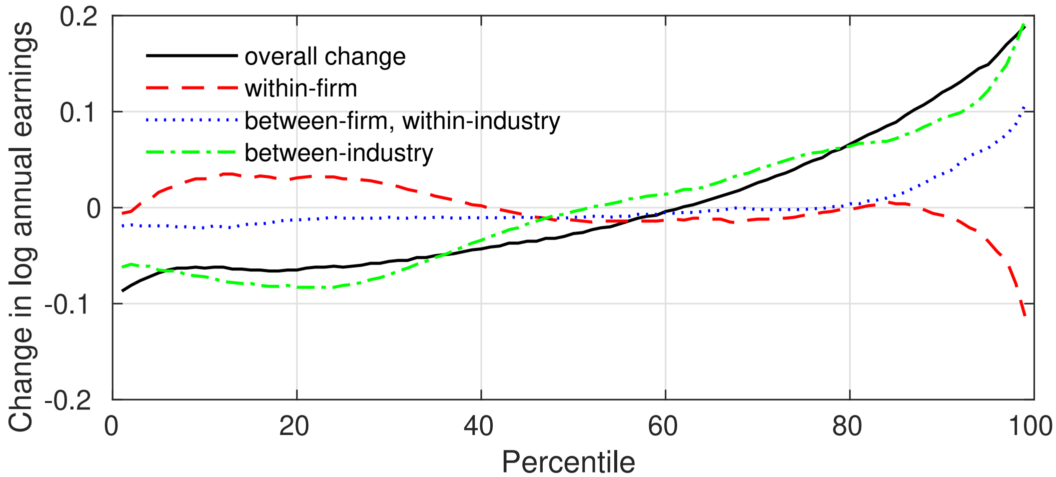
The statistics in Table 1 demonstrate that 4-digit NAICS industry accounts for almost two-thirds of the growth of earnings variance. In this section, we present a descriptive analysis to learn where in the earnings distribution industry is important. We first estimate annual earnings for each of the percentiles 1 to 99 for the first (1996-2002) and the third (2012-2018) 7-year intervals, and then calculate the difference between the first and third intervals for each percentile, as shown in Figure B1.[1] In our analytical sample, comparing the first and the third intervals, annual earnings declined by more than five log points for the first 34 percentiles, and declined for the first 61 percentiles. However, earnings at the top increased substantially. Earnings in the top 23 percentiles increased by more than 5 log points (5.1%), and earnings in the top 13 percentiles increased by more than 10 log points (10.5%).[2]

We use a simple decomposition to understand how the person, the firm, and the industry help account for the changing distribution of earnings. We can express the difference between earnings $y_t^{i,j,k,p}$ and average earnings $\bar{y}^p$ as

[1] For each 7-year interval p, we create percentiles $x \in \{1, 2, \ldots, 99\}$ for $y_t^{i,j,k,p} - \bar{y}^p$ where percentile X is defined as the mean of $y_t^{i,j,k,p} - \bar{y}^p$ for all workers between the $x - 1/2$ and the $x + 1/2$ percentiles.

[2] Throughout this paper, we convert any log differential x into a proportionate change using the expression $e^x - 1$. For small differences, log points (i.e., log differentials multiplied by 100) are approximately equal to the percent change.

Figure B1: Change in log real annual earnings, by percentile



Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees. See Equation (B1) for definitions.

$$\underbrace{y_t^{i,j,k,p} - \bar{y}^p}_{\text{relative earnings}} = \underbrace{y_t^{i,j,k,p} - \bar{y}^{j,k,p}}_{\text{within-firm}} + \underbrace{\bar{y}^{j,k,p} - \bar{y}^{k,p}}_{\substack{\text{between-firm,}\\ \text{within-industry}}} + \underbrace{\bar{y}^{k,p} - \bar{y}^p}_{\text{between-industry}} \tag{B1}$$

We estimate the mean of each of the terms on the right-hand side for each percentile of relative worker earnings ($y_t^{i,j,k,p} - \bar{y}^p$), noting that firm mean earnings $\bar{y}^{j,k,p}$, industry mean earnings $\bar{y}^{k,p}$, and the grand mean of earnings $\bar{y}^p$ are from the full sample of workers rather than calculated within each percentile.[3] To interpret this exercise, think of workers in the first percentile, who have earnings between the ½th and 1½th percentiles. We estimate how the earnings of these workers differ from the earnings of their firm ($y_t^{i,j,k,p} - \bar{y}^{j,k,p}$), how the earnings of their firm differ from the earnings of their industry ($\bar{y}^{j,k,p} - \bar{y}^{k,p}$), and how the earnings of their industry differ from the grand mean of earnings ($\bar{y}^{k,p} - \bar{y}^p$). We do this for each percentile in the first and third intervals, and then calculate the difference between the first and third intervals for each percentile.

For each percentile, the dashed line in Figure B1 is the person component $y_t^{i,j,k,p} - \bar{y}^{j,k,p}$, the dotted line is the firm component $\bar{y}^{j,k,p} - \bar{y}^{k,p}$, and the dash-dot line is the industry component $\bar{y}^{k,p} - \bar{y}^p$. We see that at the lower end of the earnings distribution, industry accounts for most of the decline. At the higher end of the earnings distribution, industry also plays a sizeable role in accounting for increasing earnings. Looking ahead to our subsequent results, Figure B1 suggests that industry plays a major role

[3] Importantly, we not computing percentiles of each of the terms. Rather we are reporting the mean of each of the terms in the decomposition for each percentile of the overall earnings distribution.

in understanding earnings change at both the lower and the upper ends of the earnings distribution.

Of interest is the role of the between-firm, within-industry component in Figure B1. This component $\bar{y}^{j,k,p} - \bar{y}^{k,p}$ has only a modest contribution to the changing earnings distribution for the first 87 percentiles. The absolute value of the dotted line is less than 2.5 log points (2.5%) for each of the first 87 percentiles. From the 88th to the 99th percentiles, the between-firm, within-industry component increases monotonically to a value of 10.7 log points (11.3%) for the highest percentile.

C The between-firm, within-industry component of rising earnings inequality

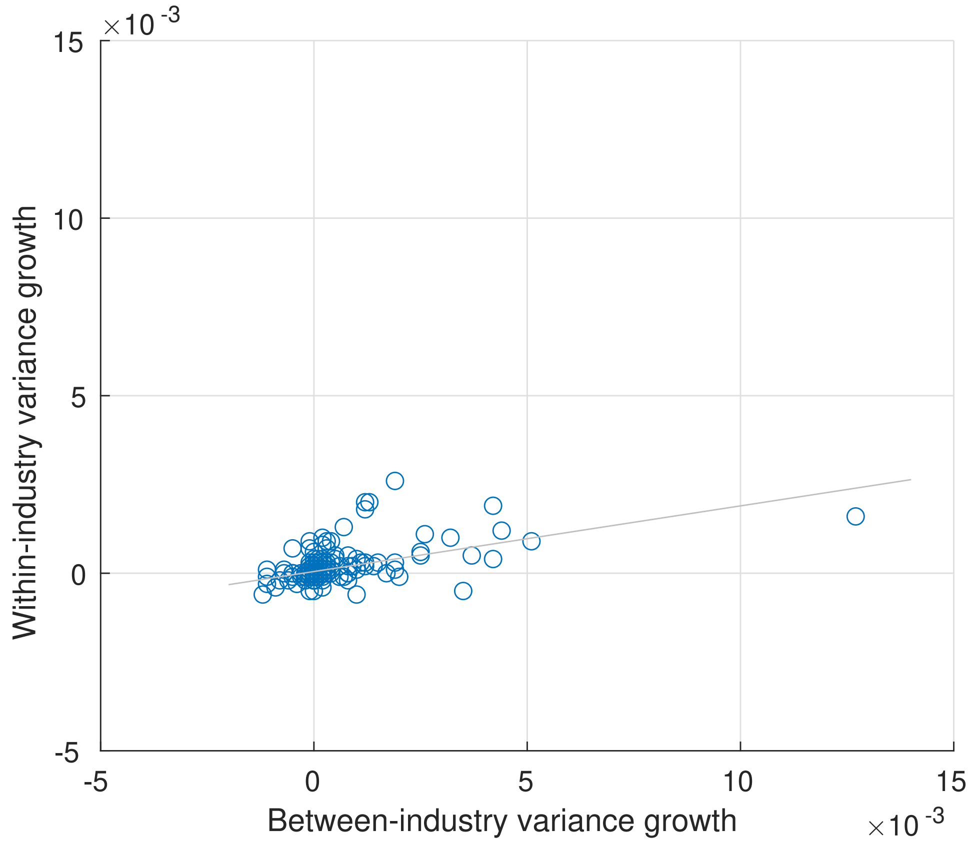
This appendix includes discussion of the between-firm, within-industry component of rising earnings inequality. Tables C4 and C5 show patterns of the top ten industries for the between-firm, within-industry contribution. The top ten industries alone contribute 59% to the between-firm, within-industry component while accounting for only 16% of employment. Four of the top ten industries in Table C4 are also among the top ten industries (for the between-industry component) in Table A5. These industries include Computer Systems Design (5415), Other Information Services (5191), Restaurants and Other Eating Places (7225), and Individual and Family Services (6241). For the six non-overlapping 4-digit industries, five overlap at the 3-digit or 2-digit level.

The overlap in the ranking of industries in terms of the between-industry component and between-firm, within-industry component is far from perfect. A good example of this is Grocery Stores which is in the bottom three for the between-firm, within-industry component (contributing negatively) and in the top ten for the between-industry component. This is a low-earnings industry that has exhibited a substantial decrease in average earnings (see Table 3), with an accompanying decrease in the firm premium (Table A5). However, within the industry, there has been a modest compression of earnings across firms within the industry. Most of this is due to decrease in sorting across firms within the industry.

While there is a strong relationship between the magnitude of the between-firm, within-industry components and the between-industry components, the between-industry components are much smaller in magnitude. This translates into a slope coefficient in Figure C1 of 0.18.

Tables C1 and C2 illustrate that the within-industry, between-firm component is also concentrated

Figure C1: Industry contributions to between-industry variance growth and within-industry variance growth



Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees.

in a relatively small fraction of industries. The top 36 industries with a contribution in excess of 1% account for more than 100% of the overall within-industry, between-firm contribution. 24 of the top 36 are high-paying industries, and similar to the between-industry, high-paying industry results, earnings changes are relatively more important than employment changes in accounting for their contribution. In contrast, for the 12 low-paying industries in the top 36, employment changes are relatively more important than earnings changes.

Table C1: Industry contributions to within-industry variance growth by variance contribution and average earnings

Industry share of within-industry variance growth	Number of industries	Total employment share	Total contribution to within-industry variance growth	Total share of within-industry variance growth
$> 5\%$	6 industries	13.9%	0.012	42.6%
1% to 5%	30 industries	24.2%	0.019	68.0%
0.05% to 1%	84 industries	25.2%	0.009	31.7%
-0.05% to 0.05%	73 industries	6.0%	-0.000	-0.5%
$< -0.05\%$	108 industries	30.7%	-0.012	-41.8%
Overall	301 industries	100.0%	0.028	100.0%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings $>$ \$3770 in EINs with 20 or more employees. Employment shares are calculated as the average of 1996-2002 and 2012-2018 employment shares.

Table C2: Industry contributions to within-industry variance growth by variance contribution and average earnings

Industry relative pay	Number of industries	Emp. share	Within-ind. var. growth	Within-ind var. cont.	Shift-share: employment	Shift-share: earnings
			36 industries with variance contribution $> 1\%$			
High-paying	24 industries	21.7%	0.019	66.9%	44.0%	56.0%
Low-paying	12 industries	16.4%	0.012	43.7%	60.8%	39.2%
			265 industries with variance contribution $\leq 1\%$			
High-paying	141 industries	34.3%	-0.005	-18.1%		
Low-paying	124 industries	27.6%	0.002	7.5%		
Overall	301 industries	100.0%	0.028	100.0%	86.4%	13.6%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings $>$ \$3770 in EINs with 20 or more employees. Employment shares are calculated as the average of 1996-2002 and 2012-2018 employment shares. Shift-share results are summed across industries and normalized by the total contribution so that the two components sum to 100%. The two rows for the 265 industries with variance contribution $\leq 1\%$ have missing cells because the denominator for the shift-share decomposition is close to zero.

Table C3: Sources of within- industry variance growth, by top 36 industries

Industry relative earnings	Number of industries	Total contribution to within-industry variance growth	Share of contribution explained by within-industry: segregation	pay premium	sorting
High-Paying	24 industries	66.9%	47.5%	13.3%	39.2%
Low-Paying	12 industries	43.7%	46.0%	15.7%	38.3%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings $>$ \$3770 in EINs with 20 or more employees.

Table C4: Industry contributions to within-industry variance growth, top 10 and bottom 3 industries

4-digit NAICS	Industry title	Relative earnings: average	Relative earnings: change	Employment share: average	Employment share: change	Within-ind. var. growth	Within-ind. var. contrib.
5613	Employment Services	-0.685	0.017	3.9%	0.6%	0.003	9.4%
5416	Management & Consulting	0.381	0.069	0.9%	0.6%	0.002	7.1%
6211	Offices of Physicians	0.254	0.098	1.7%	0.5%	0.002	7.0%
5415	Computer Systems Design	0.663	0.012	1.7%	0.9%	0.002	6.9%
6216	Home Health Care Services	-0.525	-0.016	0.8%	0.4%	0.002	6.5%
7225	Restaurants & Other Eating Places	-0.739	-0.027	4.9%	2.0%	0.002	5.8%
6214	Outpatient Care Centers	0.167	0.250	0.6%	0.5%	0.001	4.6%
5191	Other Information Services	0.798	0.699	0.2%	0.3%	0.001	4.3%
6241	Individual and Family Services	-0.490	-0.155	0.8%	0.6%	0.001	4.0%
4541	Electronic Shopping & Mail-Order	0.064	0.446	0.3%	0.1%	0.001	3.7%
4451	Grocery Stores	-0.378	-0.194	2.4%	0.0%	-0.001	-1.9%
3344	Semiconductor Manufacturing	0.556	0.299	0.8%	-0.5%	-0.001	-2.0%
3341	Computer Manufacturing	0.911	0.191	0.5%	-0.5%	-0.001	-2.3%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees. Average log earnings for industry *k* are relative to the economy average.

Table C5: Industry contributions to within-industry variance growth, top 10 and bottom 3 industries

4-digit NAICS	Industry title	Within-ind. var. share	Segregation	Pay premia	Sorting	Shift share: earnings	Shift share: employment
5613	Employment Services	9.4%	53.6%	11.0%	35.4%	54.5%	45.5%
5416	Management & Consulting	7.1%	51.5%	12.0%	36.5%	9.7%	90.3%
6211	Offices of Physicians	7.0%	50.8%	10.8%	38.5%	52.4%	47.6%
5415	Computer Systems Design	6.9%	55.7%	12.0%	32.3%	17.2%	82.8%
6216	Home Health Care Services	6.5%	38.1%	14.9%	47.0%	38.0%	62.0%
7225	Restaurants & Other Eating Places	5.8%	55.9%	13.0%	31.1%	25.9%	74.1%
6214	Outpatient Care Centers	4.6%	40.0%	14.6%	45.4%	48.0%	52.0%
5191	Other Information Services	4.3%	26.4%	29.8%	43.8%	28.3%	71.7%
6241	Individual and Family Services	4.0%	31.5%	27.9%	40.5%	27.8%	72.2%
4541	Electronic Shopping & Mail-Order	3.7%	43.7%	11.7%	44.7%	69.3%	30.7%
4451	Grocery Stores	-1.9%	27.8%	9.3%	63.0%	101.1%	-1.1%
3344	Semiconductor Manufacturing	-2.0%	16.1%	28.6%	55.4%	-128.1%	228.1%
3341	Computer Manufacturing	-2.3%	43.8%	20.3%	35.9%	19.3%	80.7%

Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees.

D CPS-LEHD Matching and Analysis

The Census Bureau has attached Protected Identification Keys (PIKs) to the CPS-ASEC for survey years since 1996. PIKs are the Census Bureau's unique individual identifier. Knowing the PIK and the earnings reference year allows us to link the CPS-ASEC to the annualized version of the LEHD. Not every record in the CPS-ASEC has a PIK attached.

Since not every CPS-ASEC record has a PIK, we adjust the the CPS-ASEC weights with a propensity score adjustment. We have done so, estimating year-specific logistic regressions where the dependent variable is equal to 1 if the CPS-ASEC record has a PIK, and 0 otherwise. The explanatory variables are dummy variables for CPS state, age, gender, race, Hispanic origin, foreign born, marital status, and education. We only keep observations in the merged data where an individual is in both the CPS-ASEC and in the LEHD. We estimate another set of year-specific logistic regressions where the dependent variable is "1 if PIKed CPS-ASEC matches to the LEHD, 0 otherwise." The explanatory variables are dummy variables for CPS age, gender, race, Hispanic origin, foreign born, marital status, and education. Dummy variables for state are not included in this propensity score model since the CPS-ASEC is national but the LEHD data we are using is restricted to 18 states. We further adjust the CPS-ASEC weights in the matched sample by dividing by the predicted value. All statistics from the linked CPS-LEHD data use these twice-adjusted propensity score weights

For the CPS-LEHD integrated data, we don't impose the 20+ size restriction for the EIN-based firms in the LBD. This helps show our results are robust to this restriction as we compare the full LEHD results with that restriction to the CPS-LEHD results without that restriction.

We estimate the human capital earnings equation used by Hoffman, Lee and Lemiuex (2020) with the CPS-LEHD data as follows:

$$y_i = AgeEduc_i\beta_1 + Occupation_i\beta_2 + Industry_i\beta_3 + \varepsilon_i. \tag{D1}$$

$AgeEduc_i$ is a vector of dummy variables that are equal to one if worker i has a particular combination of age and education, and are equal to zero otherwise.[4] The marginal effects of these demographic categories on earnings is given by β_1. $Occupation_i$ is a vector of dummy variables equal to one if worker i is employed in one of nine occupations, and are zero otherwise. The marginal effect of each

[4]Specifically, we allow for an effect for each of eight five-year age ranges {26-30, 31-35,..., 61-65} interacted with five education dummies {less than high school, high school graduate, some college, college graduate, post-graduate}.

Table D1: Variance decomposition using an earnings equation: AKM vs. human capital

Data Source Specification	LEHD AKM	CPS-LEHD linked AKM	CPS-LEHD linked human capital
Between-industry:	61.9%	66.2%	66.2%
Segregation	25.2%	26.9%	15.3%
Pay premia	8.7%	9.6%	21.4%
Covariance	28.0%	29.6%	29.5%

Notes: The first column is taken from Table 7. The second and third columns are from the linked CPS-LEHD dataset, common-coded sample, see Equation (D2) for definitions. In all columns, the three components add up to the total between-industry contribution.

occupation category is given by the vector β_2. Analogously, $Industry_i$ is a vector of dummy variables for each of 299 4-digit NAICS industries, with marginal effects given by the vector β_3. We use log earnings from LEHD as the dependent variable y_i.

Let $Z = [AgeEduc_i \; Occupation_i]$ concatenate the vectors of age by education dummies and occupation dummies. Analogously, let β_Z concatenate the marginal effects vectors β_1 and β_2. Letting $\overline{Z}_k$ denote a vector of averages of each dummy variable, we can 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:

$$\begin{aligned}\underbrace{\mathrm{var}(y_{i,k})}_{\substack{\text{earnings}\\\text{variance}}} &= \underbrace{\mathrm{var}(Z_{i,k}\beta_Z - \overline{Z}_k\beta_Z)}_{\substack{\text{within-industry dispersion}\\\text{from age, education,}\\\text{and occupation}}} + \underbrace{\mathrm{var}(\overline{Z}_k\beta_Z)}_{\substack{\text{between-industry}\\\text{segregation}}} + \\ &\underbrace{\mathrm{var}(Industry_{i,k}\beta_3)}_{\substack{\text{between-industry}\\\text{pay premia}}} + \underbrace{2\mathrm{cov}(\overline{Z}_k\beta_Z, Industry_{i,k}\beta_3)}_{\text{between-industry covariance}} + \underbrace{\mathrm{var}(\varepsilon_{i,k})}_{\substack{\text{residual dispersion}\\\text{(within-industry)}}}\end{aligned} \quad \text{(D2)}$$

The terms on the right hand side correspond to an alternative but analogous decomposition to that using the AKM decomposition of earnings. In the main text, we combine the covariance and segregation terms into a combined sorting component.

Table D1 is an enhanced version of Table 8 that breaks out the covariance and segregation components of the decomposition using the CPS-LEHD linked data.

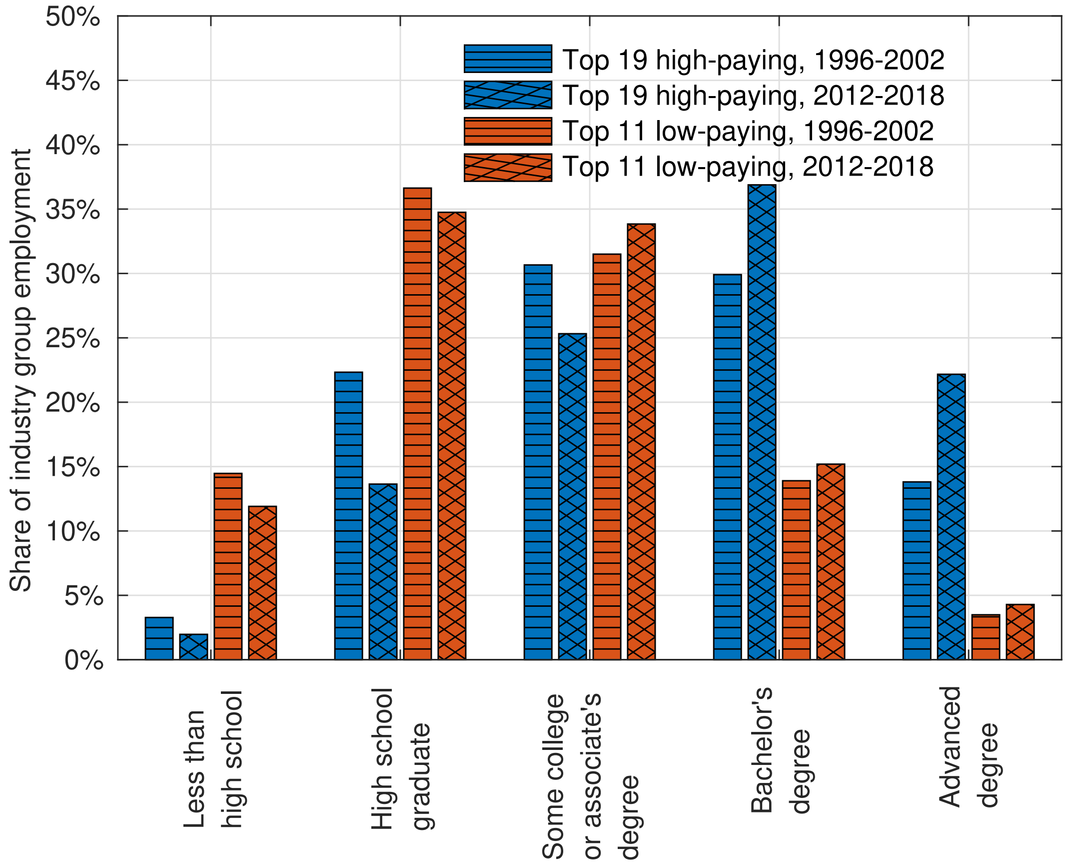
We also use a decomposition of the change in the mean of log earnings at the industry level given by:

$$\underbrace{\Delta(\bar{y}^k - \bar{y})}_{\substack{\text{change in industry} \\ \text{average relative earnings}}} = \underbrace{\Delta\overline{AgeEduc}_k\beta_1}_{\substack{\text{change attributable} \\ \text{to age and education}}} + \underbrace{\Delta\overline{Occupation}_k\beta_2}_{\substack{\text{change attributable} \\ \text{to occupation effects}}} + \underbrace{\Delta Industry_k\beta_3}_{\substack{\text{change attributable} \\ \text{to industry effects}}}. \tag{D3}$$

This decomposition is used in Figure 6.

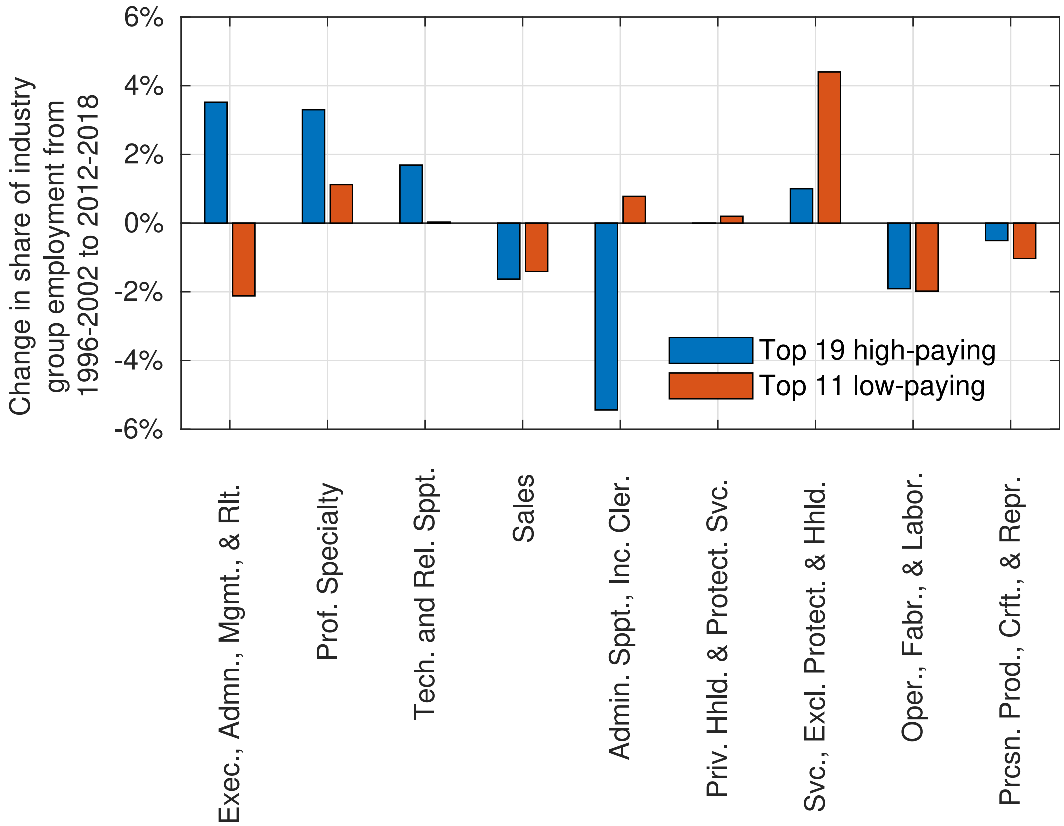
The matched CPS-LEHD integrated data is also used in Haltiwanger, Hyatt and Spletzer (2023). This complementary paper focuses on distinct issues relative to the current paper. Much of the literature on rising earnings inequality uses household survey data and specifically the CPS. The complementary paper focuses on the differences between the survey and administrative data that yield differences in inferences on changes in inequality. We identify two major measurement issues that raise questions about the use of the CPS for the study of earnings inequality. First, the CPS has "trouble in the tails" of the earnings distribution in both the cross section and in terms of changes over time. That is, the dispersion of earnings for the same individuals is substantially smaller in the CPS compared to the administrative (LEHD) data. Moreover, this compression of earnings dispersion carries over to changes in earnings dispersion. For the same individuals, we find that earnings inequality increases by only half as much in the survey data than in the administrative data. Second, the CPS has significant issues with industry mismeasurement. We find that even at the broad sector level, for the same persons the industry classification in the CPS differs from that in the administrative data in about 40% of the cases. These two key measurement-oriented contributions are unique to this complementary paper. one of the main conclusions in the complementary paper is that the measurement limitations of the CPS raise questions about using it as the primary basis for making inferences about inequality.

Figure D1: Employment share by education group class, by industry group



Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > $3770. Changes in the employment shares are expressed in terms of percentage points. The denominator is total employment in the respective industry group.

Figure D2: Change in employment share by occupation, by industry group



Notes: Authors' tabulations of linked CPS-LEHD microdata. Tabulations include workers with annual real earnings > \$3770. Changes in the employment shares are expressed in terms of percentage points. The denominator is total employment in the respective industry group.

Table D2: Correlation matrix, all years pooled, linked CPS-ASEC and LEHD data correlations

	Std. Deviation	LEHD earnings	CPS person	CPS ind $Z\beta$	CPS ind$\beta 3$	LEHD person	LEHD firm $,Xb$	LEHD firm ψ	LEHD ind $\theta, X\beta$	LEHD ind ψ	CPS residual	LEHD residual
LEHD earnings	0.896	1.000	0.431	0.381	0.398	0.659	0.348	0.296	0.480	0.467	0.756	0.346
CPS person	0.387		1.000	0.000	0.000	0.481	0.237	0.115	0.000	0.000	0.000	0.008
CPS ind $Z\beta$	0.182			1.000	0.559	0.000	0.000	-0.003	0.783	0.703	0.000	-0.004
CPS ind $\beta 3$	0.343				1.000	-0.005	0.001	-0.001	0.746	0.841	0.000	-0.005
LEHD person	0.607					1.000	-0.007	-0.004	-0.005	-0.011	0.598	-0.029
LEHD firm $\theta, X\beta$	0.282						1.000	0.172	-0.001	-0.004	0.323	-0.008
LEHD firm ψ	0.222							1.000	-0.005	-0.004	0.326	-0.001
LEHD ind $\theta, X\beta$	0.263								1.000	0.885	0.000	-0.006
LEHD ind ψ	0.197									1.000	0.000	-0.005
CPS residual	0.677										1.000	0.454
LEHD residual	0.333											1.000

Notes: "CPS person" is defined as $Z_{ik}\beta - \overline{Z_k\beta}$, where Z is $Age * Educ$ and occupation "CPS ind $Z\beta$" is defined as $\overline{Z_k\beta}$ "CPS ind $\beta 3$" is defined as Industry$_{ik}\beta_3$ "LEHD person" is defined as $\theta^i - \overline{\theta}^{j,k} + X_t^i\beta - \overline{X}^{j,k}\beta$ "LEHD firm $\theta, X\beta$" is defined as $\overline{\theta}^k) + \overline{X}^k\beta$ "LEHD firm ψ" is defined as $\psi^{j,k} - \overline{\psi}^k)$ "LEHD industry $\theta, X\beta$" is defined as $\overline{\theta}^k) + \overline{X}^k\beta)$ "LEHD industry ψ"$isdefinedas$var$(\overline{\psi}^k)$ "CPS residual" is defined as $Y_{ik} - Z_{ik}\beta - (Industry)_{ik}\beta_3$, corresponding to column 5 in Table 7. "LEHD residual" is defined as $y_{i,j,t} - \theta_i - \psi_j - X_{i,t}\beta$, corresponding to the right half of Table 8. Statistics in blue have a p-value less than 1%

Table D3: Decomposition of variance growth in the top 30 industries using a human capital equation

			Between-	Sorting			Segregation				
Industry group	Interval	Emp. share	industry variance	Total	Age by educ.	Occu-pation	Total	Age by educ.	Occu-pation	Covar-iance	Industry premia
	1996-2002	18.9%	0.032	0.014	0.009	0.005	0.006	0.003	0.001	0.003	0.012
	2012-2018	21.6%	0.072	0.032	0.020	0.012	0.013	0.005	0.002	0.006	0.027
Top 19	Change	2.8%	0.039	0.019	0.011	0.008	0.007	0.002	0.001	0.003	0.014
high-paying	Share of total bet.-ind. var. change		59.7%								
	Share of top 19 bet.-ind. var. change		100%	47.0%	27.9%	19.0%	16.8%	6.1%	2.8%	7.9%	36.5%
	1996-2012	14.0%	0.045	0.018	0.012	0.006	0.005	0.002	0.001	0.002	0.023
	2012-2018	18.9%	0.075	0.032	0.022	0.010	0.011	0.006	0.001	0.004	0.033
Bottom	Change	4.8%	0.030	0.014	0.010	0.004	0.006	0.003	0.001	0.002	0.010
11 low-paying	Share of total bet.-ind. var. change		45.8%								
	Share of bottom 11 bet.-ind. var. change		100%	46.4%	32.1%	14.2%	19.5%	10.3%	1.7%	7.6%	33.8%

Notes: Authors' tabulations of linked CPS-LEHD microdata.

E Comparison to Song et al. (2019)

It is useful to consider the original decomposition by Song et al. (2019) that did not extend the decomposition to the within- and between-industry components. Using the Equation (4), denote the firm-level average worker effect of firm j during interval p (hereafter suppressing the superscript for interval p) as $\bar{\theta}^{j,k}$, and similarly denote the average observable characteristics as $\bar{X}^{j,k}$. The variance of earnings can be written as

$$\begin{aligned}
\operatorname{var}(y_t^{i,j,k}) = & \underbrace{\operatorname{var}(\theta^i - \bar{\theta}^{j,k}) + \operatorname{var}(X_t^i\beta - \bar{X}^{j,k}\beta) + 2\operatorname{cov}(\theta^i - \bar{\theta}^{j,k}, X_t^i\beta - \bar{X}^{j,k}\beta)}_{\text{within-firm person effects and observables}} + \\
& \underbrace{\operatorname{var}(\bar{\theta}^{j,k}) + \operatorname{var}(\bar{X}^{j,k}\beta) + 2\operatorname{cov}(\bar{\theta}^{j,k}, \bar{X}^{j,k}\beta)}_{\text{total segregation}} + \underbrace{\operatorname{var}(\psi^{j,k})}_{\text{total pay premia}} + \\
& \underbrace{2\operatorname{cov}(\bar{\theta}^{j,k}, \psi^{j,k}) + 2\operatorname{cov}(\bar{X}^{j,k}\beta, \psi^{j,k})}_{\text{total covariance}} + \underbrace{\operatorname{var}(\varepsilon_t^{i,j,k})}_{\text{residual (within-firm)}} .
\end{aligned} \tag{E1}$$

Exploring the worker- and firm-level contributions involves collecting terms from this basic decomposition.

Between-firm dispersion can be expressed through the contributions of the covariance contribution (what Song et al. (2019) denoted as sorting), segregation, and firm premia.[5] The covariance term reflects the covariance between worker and firm effects, given by $2\operatorname{cov}(\bar{\theta}^{j,k}, \psi^{j,k}) + 2\operatorname{cov}(\bar{X}^{j,k}\beta, \psi^{j,k})$. In other words, the covariance term reflects the extent to which highly-paid workers work for high-paying firms. Segregation reflects the concentration within firms of workers of the same type (captured by person effects), given by $\operatorname{var}(\bar{\theta}^{j,k}) + \operatorname{var}(\bar{X}^{j,k}\beta) + 2\operatorname{cov}(\bar{\theta}^{j,k}, \bar{X}^{j,k}\beta)$. The remaining contributor to between-firm dispersion is reflected in the firm premia term $\operatorname{var}(\psi^{j,k})$.

The remaining dispersion is within-firm dispersion. Worker-level effects are given by $\operatorname{var}(\theta^i - \bar{\theta}^{j,k}) + \operatorname{var}(X_t^i\beta - \bar{X}^{j,k}\beta) + 2\operatorname{cov}(\theta^i - \bar{\theta}^{j,k}, X_t^i\beta - \bar{X}^{j,k}\beta)$. Residual dispersion $\operatorname{var}(\varepsilon_t^{i,j,k})$ occurs within firms.[6]

Before comparing our estimates with Song et al. (2019), it is worth noting that Card, Heining and Kline (2013) considered an alternative simpler decomposition of the variance of earnings that focuses

[5]There is a covariance component of segregation that reflects the covariance across different worker effects. This covariance is quite distinct from the covariance reflecting the covariance between industry premia and worker effects.

[6]The estimated residual from Equation (4) is by construction orthogonal to worker effects, as well as the effects of worker characteristics. But the estimated residual can be correlated with the deviation of worker effects and the effects of observable characteristics from their respective firm-level averages because they are not explicitly controlled for in Equation (4).

Table E1: Comparison to Song et al. (2019), males only

	Song et al. (2019) growth 1994-2000 2007-2013	Our estimates of LEHD growth 1996-2002 2012-2018
Total variance increase	0.096	0.126
Within-firm share	13.5%	15.5%
Between-firm share	86.5%	84.5%
Segregation	35.5%	37.4%
Pay premia	14.6%	11.8%
Covariance	37.5%	35.3%

Notes: Song et al. (2019) estimates taken from their Table V (page 36).

on the variance of the person effects (unobservable worker effects + observables), the variance of the firm effects, twice the covariance of the firm and person and the variance of the residual. The primary difference is that the Song et al. decomposition permits decomposing the variance of the person effects into within and between firm components.

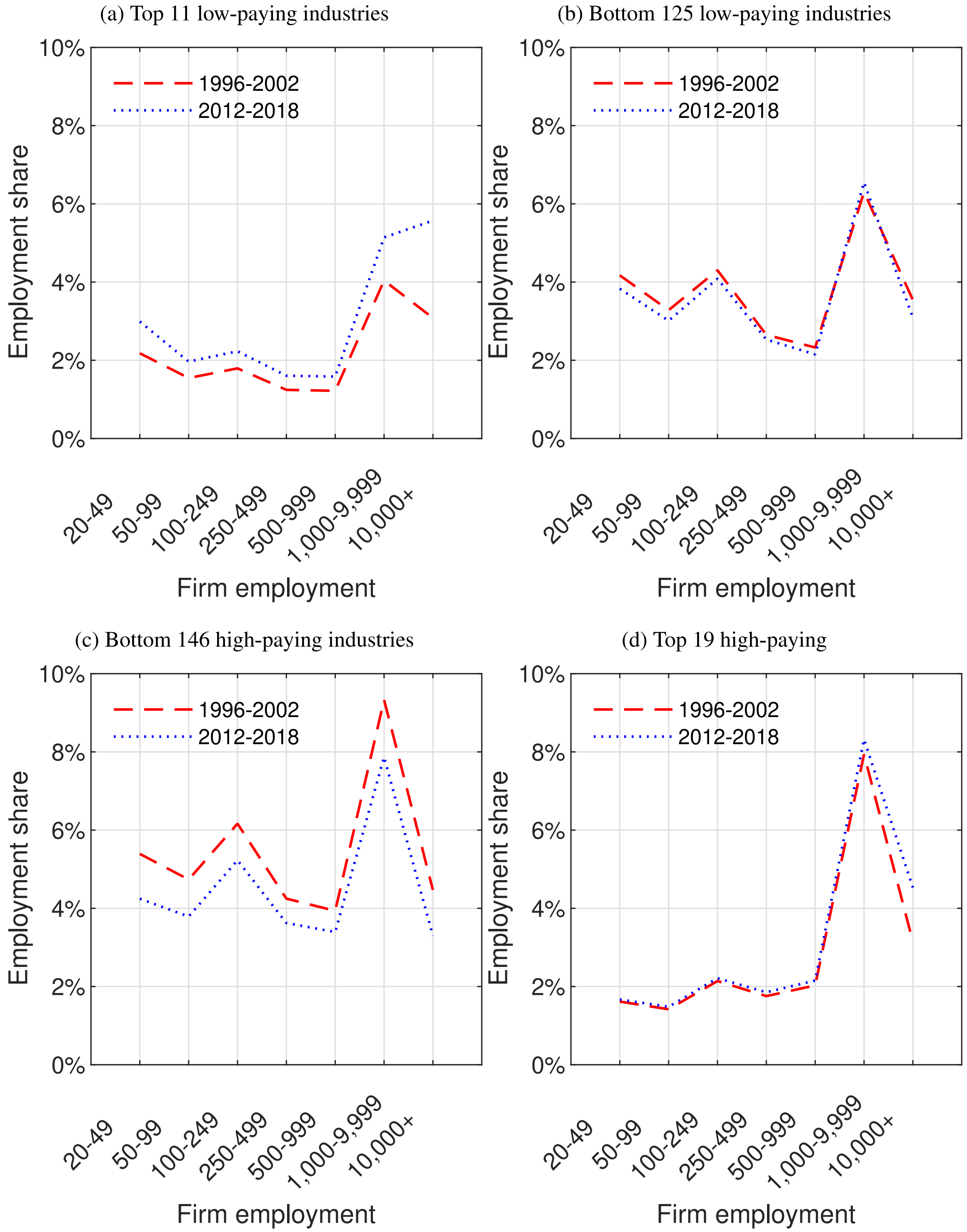
Our industry-enhanced decomposition can be easily collapsed into the Song et al. (2019). Our results in Table A4 show that firm-level segregation accounts for 36.8% (=11.6%+25.2%) of inequality growth over the 1996-2002 to 2012-2018 intervals, firm-level covariance (reminder Song et al. (2019) denote this as *sorting*) accounts for 36.6% (=8.6%+28.0%), and the rising firm premia accounts for 11.6% (=2.9%+8.7%). We have estimated our results separately for males and females which we summarize here to compare to Song et al. (2019) (details available upon request). Our results for males are similar: segregation 37.4%, covariance 35.3%, and pay premia 11.8%, see Table E1. These segregation, covariance, and firm premia results for males are similar to those of males in Song et al. (2019) when looking at variance growth over their 1994-2000 to 2007-2013 intervals: segregation 35.5%, sorting 37.5%, and pay premia 14.6%. These contributions are broadly similar to those in the longer time interval (1980-1986 to 2007-2013) reported in Song et al. (2019) with one notable exception: there is a smaller role for firm premia in the longer time interval (-1.4%). We also find a close correspondence of results for females to those reported in Song et al. (2019) over similar time periods. These results imply that our findings indicate that the between-firm contribution to increasing inequality reported by Song et al. (2019) is largely – but not entirely – determined at the industry level.

The tight relationship of the role of between-firm effects comparing our results with Song et al. (2019) mitigates concerns about our use of an 18-state LEHD sample. Any such concerns are further mitigated by the analysis we report from the LBD in Appendix Table H1. In our 18-state LEHD results, 73% of the rising between firm dispersion is accounted for by rising between-industry dispersion. In the 18-state and 50-state LBD, the analogous statistics are 73% and 69%, respectively. Taken together, the comparisons with Song et al. (2019) and calculations from the LBD imply that the patterns we report here from administrative data are robust to using an 18-state or 50-state sample.

F Mega firms

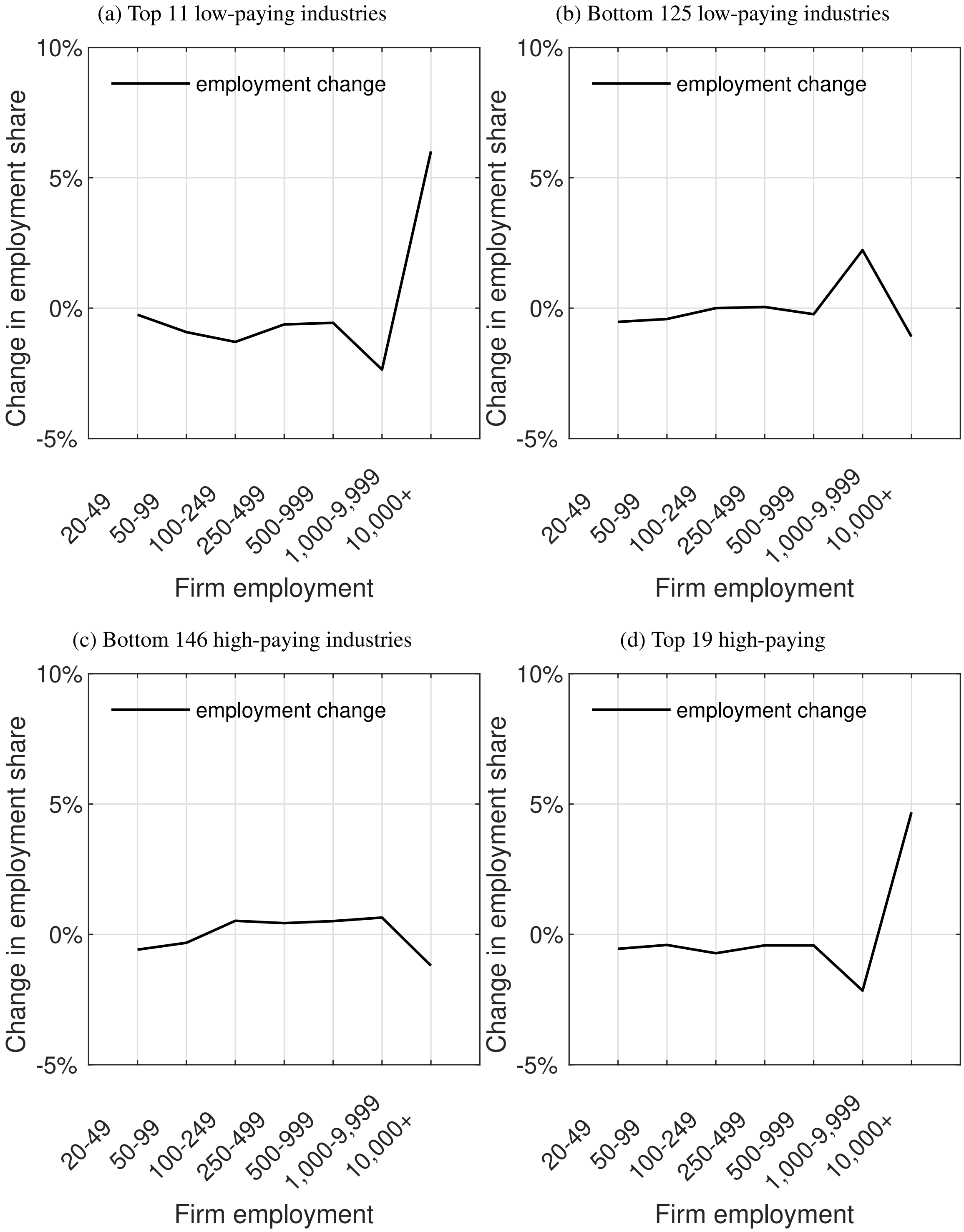
This appendix includes supplementary tables and figures for our analysis of mega firms.

Figure F1: Employment share by size class and industry group



Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > $3770 in EINs with 20 or more employees.

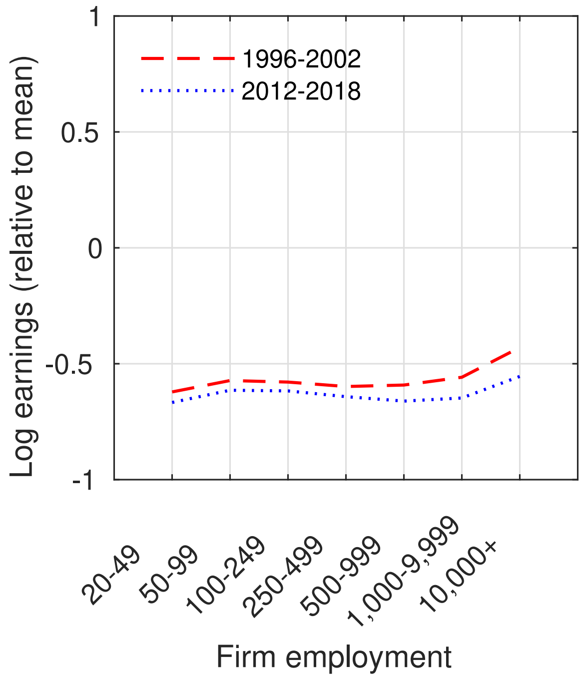
Figure F2: Change in employment share by size class and industry group



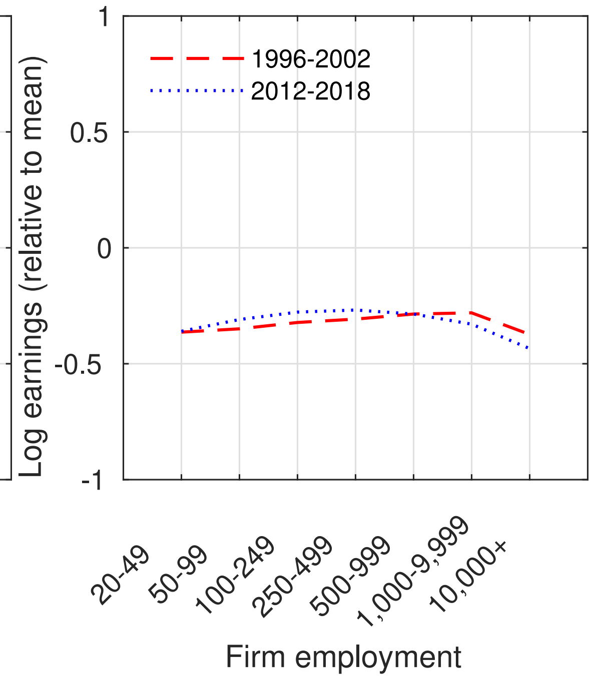
Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > $3770 in EINs with 20 or more employees.

Figure F3: Earnings per worker by size class and industry group

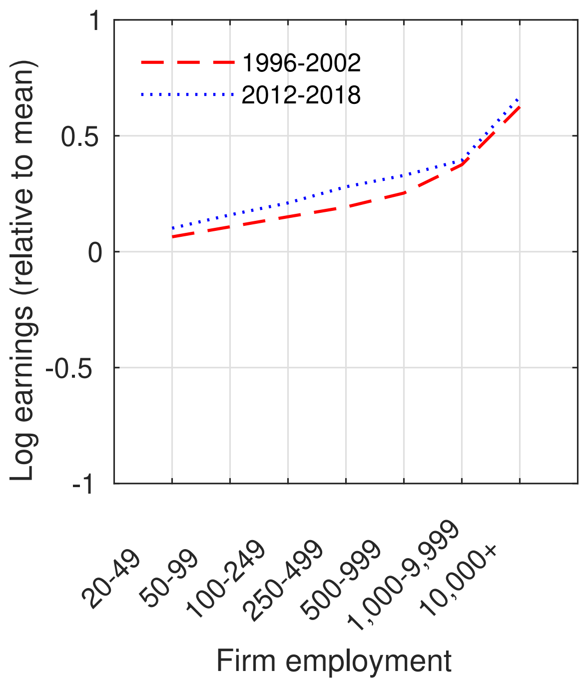
(a) Top 11 low-paying industries



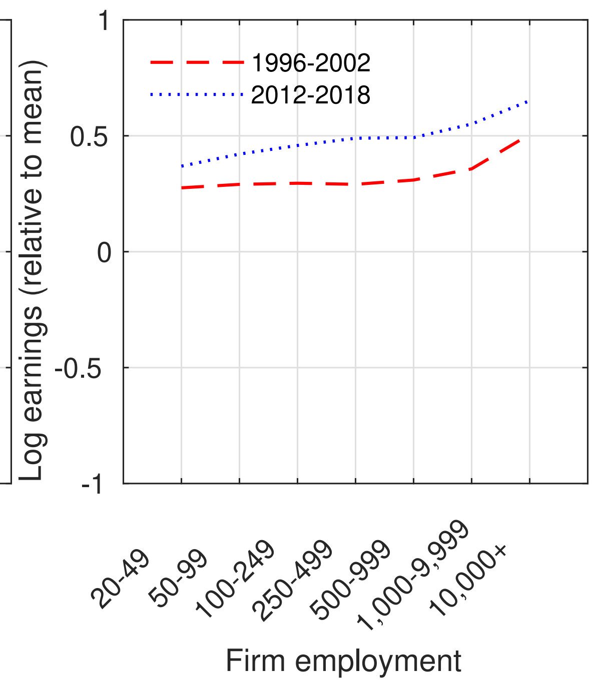
(b) Bottom 125 low-paying industries



(c) Bottom 146 high-paying industries

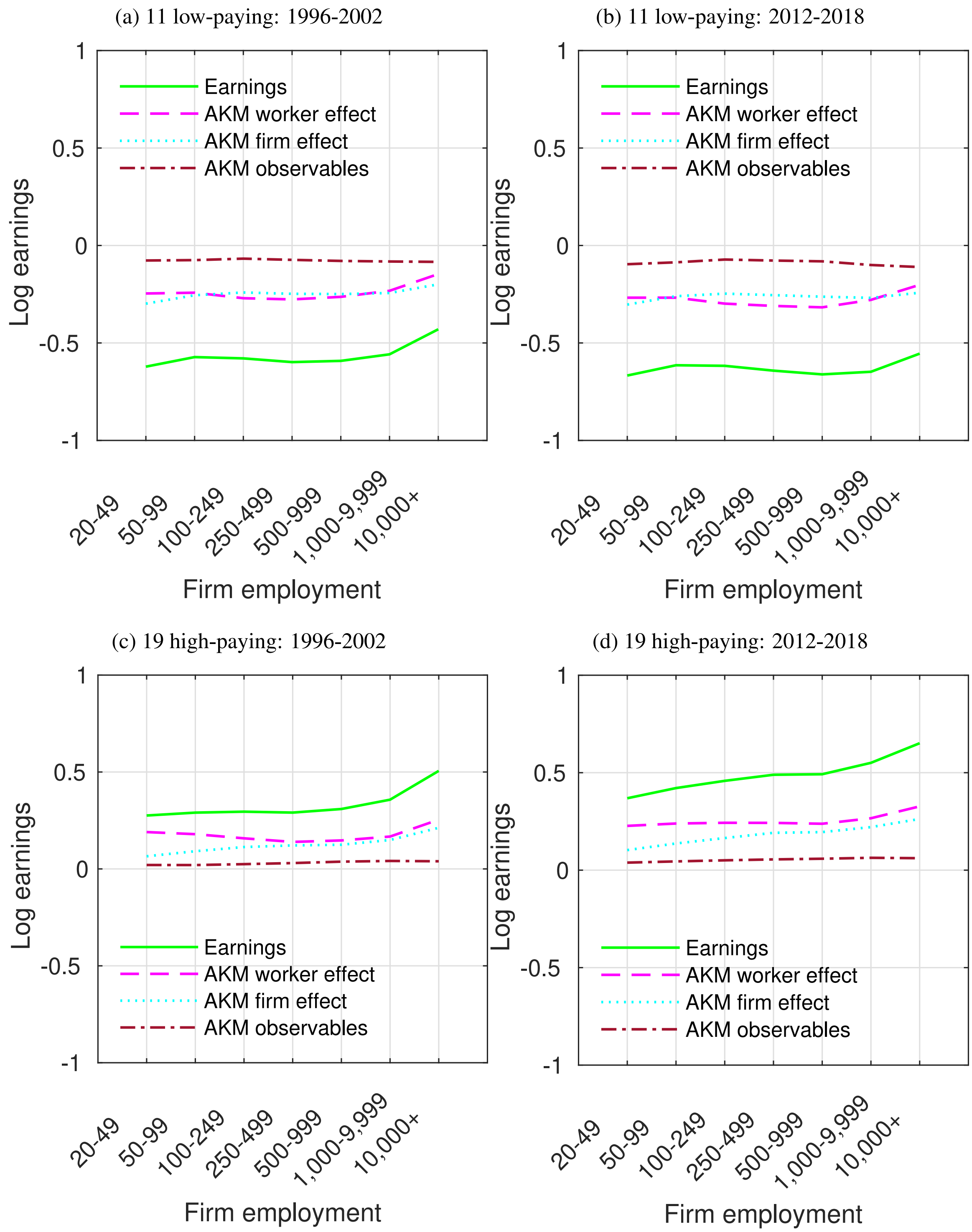


(d) Top 19 high-paying



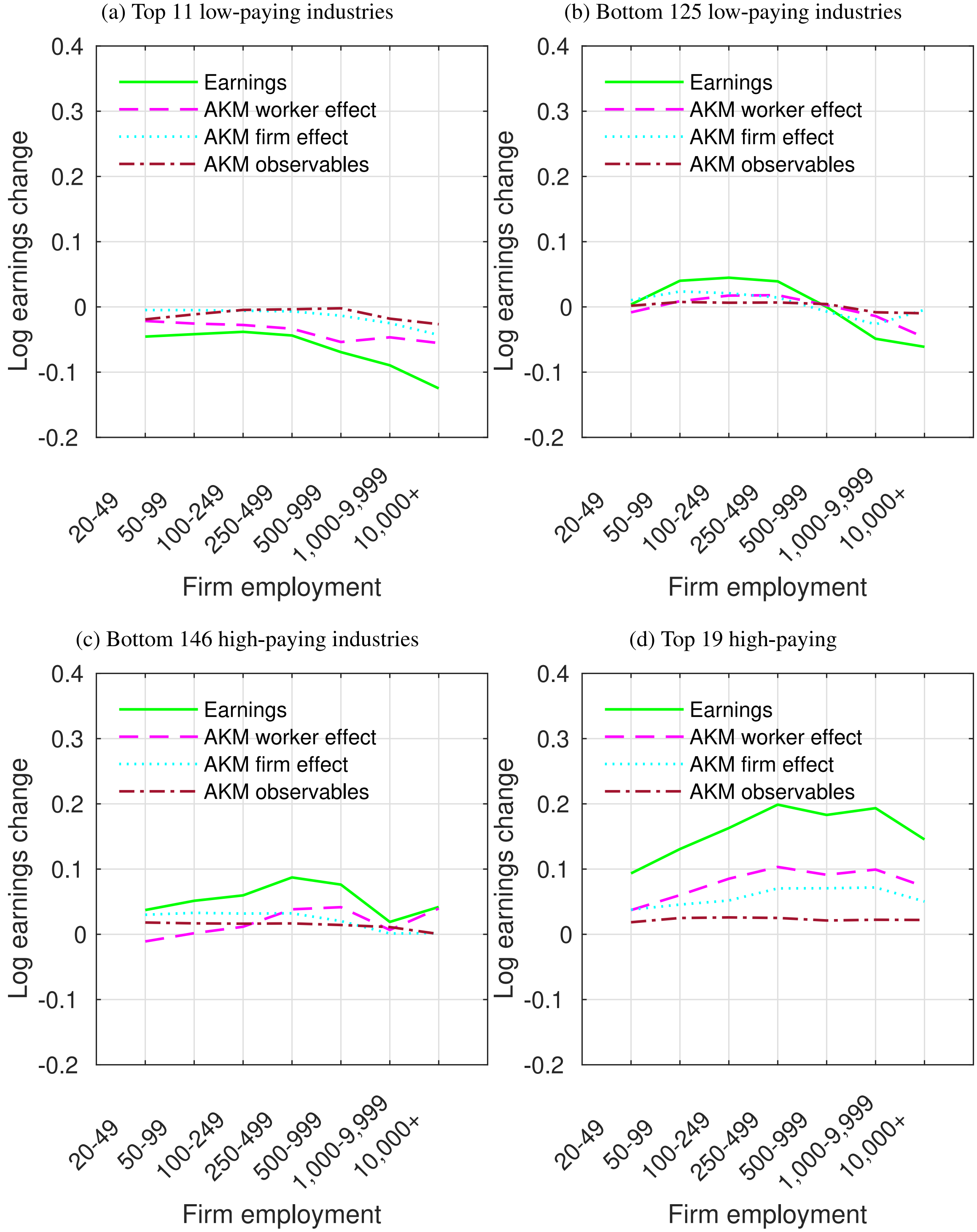
Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings $>$ \$3770 in EINs with 20 or more employees.

Figure F4: Earnings levels by size class for select industries



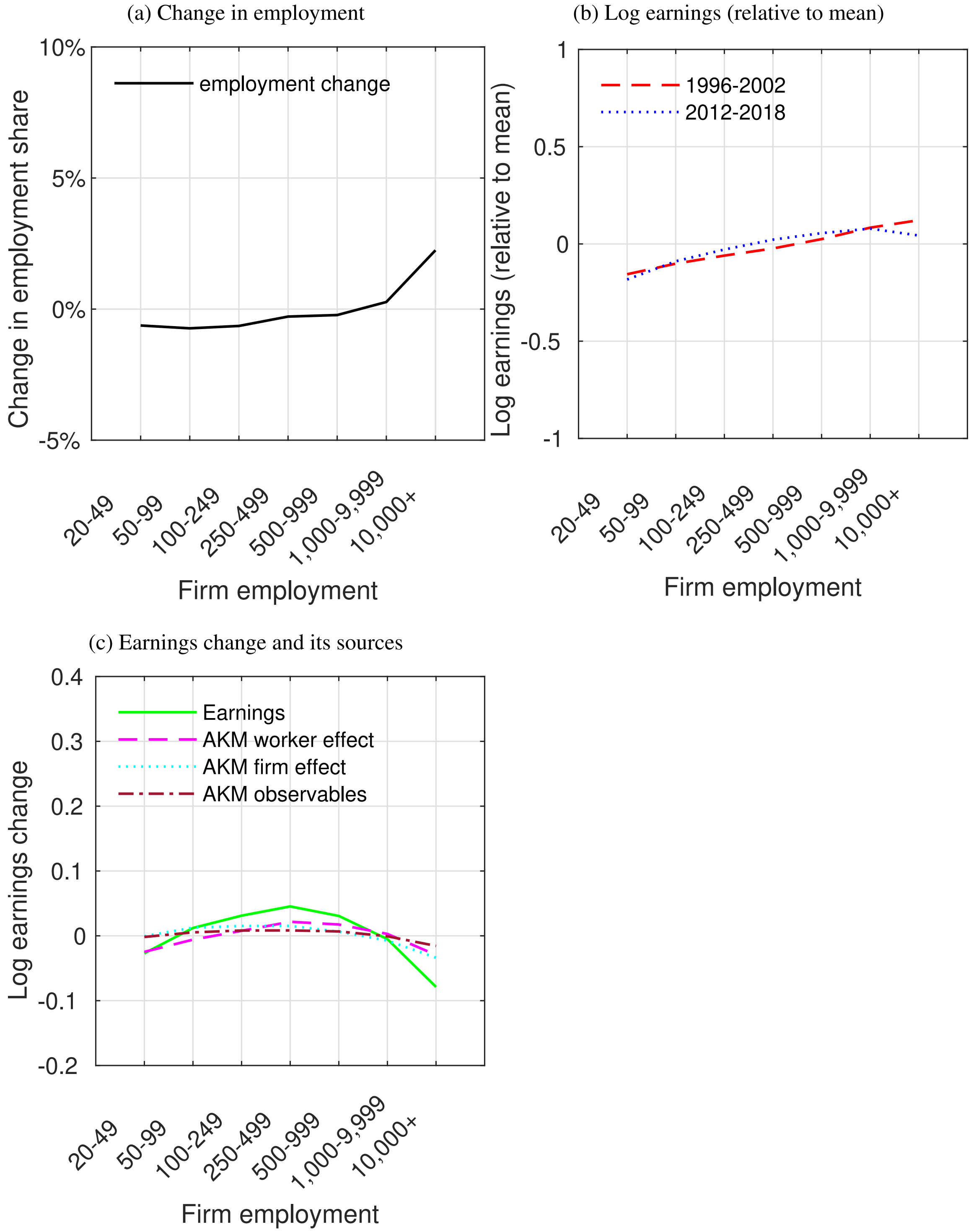
Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > $3770 in EINs with 20 or more employees.

Figure F5: Earnings change by size class for select industries, by industry group



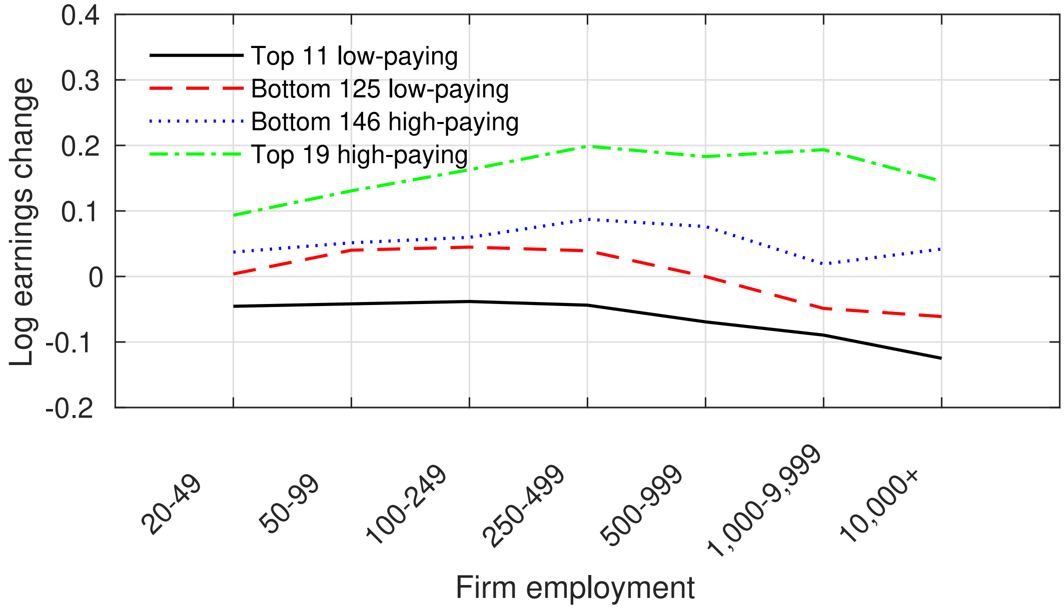
Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees. Size class is in terms of employment. See Equation (4) for definitions.

Figure F6: Employment and earnings by size class, national



Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > \$3770 in EINs with 20 or more employees. Size class is in terms of employment.

Figure F7: Change in log earnings by size class, by industry group



Notes: Authors' tabulations of LEHD microdata. Tabulations include workers with annual real earnings > $3770 in EINs with 20 or more employees.

G Occupation by industry dataset and analysis

We downloaded the May OEWS national industry specific data from https://www.bls.gov/oes/tables.htm. We start our analysis in 2002, as the 1996-2001 are published on a SIC basis. The unit of observation in each year-specific file is 4-digit industry and occupation. We restrict the data in each year to keep only industry totals and major (2 digit) occupations. We end our analysis in 2016 because 2017 and 2018 have a different level of industry aggregation.[7] There are 22 major occupations.[8]

We then create a balanced panel of 284 industries for each year 2002-2016.[9] There are nine industries that are published by OEWS for some years 2002-2016 but not all.[10] We had to adjust our analysis dataset because the OEWS switched to NAICS 2012 in 2012. One of the biggest changes with the NAICS 2012 was replacing NAICS 7221 "Full Service Restaurants" and 7222 "Limited Service Eating Places" with adding NAICS 7225 "Restaurants and Other Eating Places."[11]

We use OEWS data to derive a balanced panel of 284 industries for each year 2002-2016. For reasons stated above, we do not include 2017 and 2018 in creating the balanced panel. The unit of observation is an industry-year. 284 industries by 15 years yields a total of 4260 observations.

We now add one more variable into the OEWS balanced panel: an indicator of whether the 4-digit NAICS industry is in HHS's nineteen high-paying industries (H19), is in HHS's eleven low-paying industries (L11), is in HHS's 146 other high-paying industries (H146), or is in HHS's 125 other low-paying industries (L125).

Merging the balanced panel OEWS (284 industries) with the 301 industries in HHS is not a one-to-one merge. There are 20 industries in HHS that are not in the OEWS, which are mostly in agricul-

[7] For example, published OEWS industry 3250A1 in 2017 aggregates industries 3251, 3253, 3253, and 3259, which are published in detail in earlier years.

[8] These are: Management (11), Business and Financial Operations (13), Computer and Mathematical Science (15), Architecture and Engineering (17), Life, Physical, and Social Science (19), Community and Social Services (21), Legal (23), Education, Training, and Library (25), Arts, Design, Entertainment, Sports, and Media (27), Healthcare Practitioner and Technical (29), Healthcare Support (31), Protective Service (33), Food Preparation and Serving Related (35), Building and Grounds Cleaning and Maintenance (37), Personal Care and Service (39), Sales and Related (41), Office and Administrative Support (43), Farming, Fishing, and Forestry (45), Construction and Extraction (47), Installation, Maintenance, and Repair (49), Production (51), Transportation and Material Moving (53).

[9] We do not include 2017 and 2018, as including these two years would result in a balanced panel of only 225 industries.

[10] These are Postal Service (4911), Internet Publishing (5161), Telecommunications Carriers (5173), Satellite Telecommunications (5174), Cable Distribution (5175) Internet Service Providers (5181), Monetary Auth Central Bank (5211), Insurance Benefit Funds (5251), and Other Investment Funds (5259).

[11] In 2011, NAICS 7221 and 7222 had 4.6 and 4.1 million employees, respectively, and 91% of employment in each of these industries was in occupation 35 "Food Preparation and Serving." In 2012, NAICS 7225 had 8.9 million employees, with 91% in occupation 35. We have recoded all occurrences of NAICS 7221 and 7222 in 2002-2011 to NAICS 7225.

ture.[12] There are 3 industries in the OEWS that are not in HHS.[13] The OEWS–HHS linked data is therefore a balanced panel of 281 industries for each year 2002-2016. The unit of observation is an industry-year (281 industries * 15 years = 4215 observations).

All nineteen of the high-paying industries identified using LEHD data are in the OEWS balanced panel of 281 industries, and all 11 of the industries in the HHS "11 Low-Paying Industries" are in the OEWS balanced panel of 281 industries. 141 of the 146 "Other High-Paying Industries" and 110 of the 125 "Other Low-Paying Industries" are in the OEWS balanced panel of 281 industries. In Table 9, we focus on the decomposition in Equation (G2) for the the 19 high-paying and the 11 low-paying industries.

Aggregating the OEWS–HHS data over NAICS industries and HHS groups {19 high-paying, 11 low-paying, 146 high-paying, 125 low-paying} allows for analysis of occupational distribution in our 4 groups of industries, as shown in Appendix Figures G1 and G2, as well as Appendix Tables G1 and G2.

To explore the concentration of occupations in industries, we conduct additional exercises. First, we use the data underlying our 4-digit industry by 2-digit occupation analysis for 2002, 2003, 2015, 2016. An attractive feature of this data is that missingness is not a substantial problem and industries and occupations are harmonized across time.

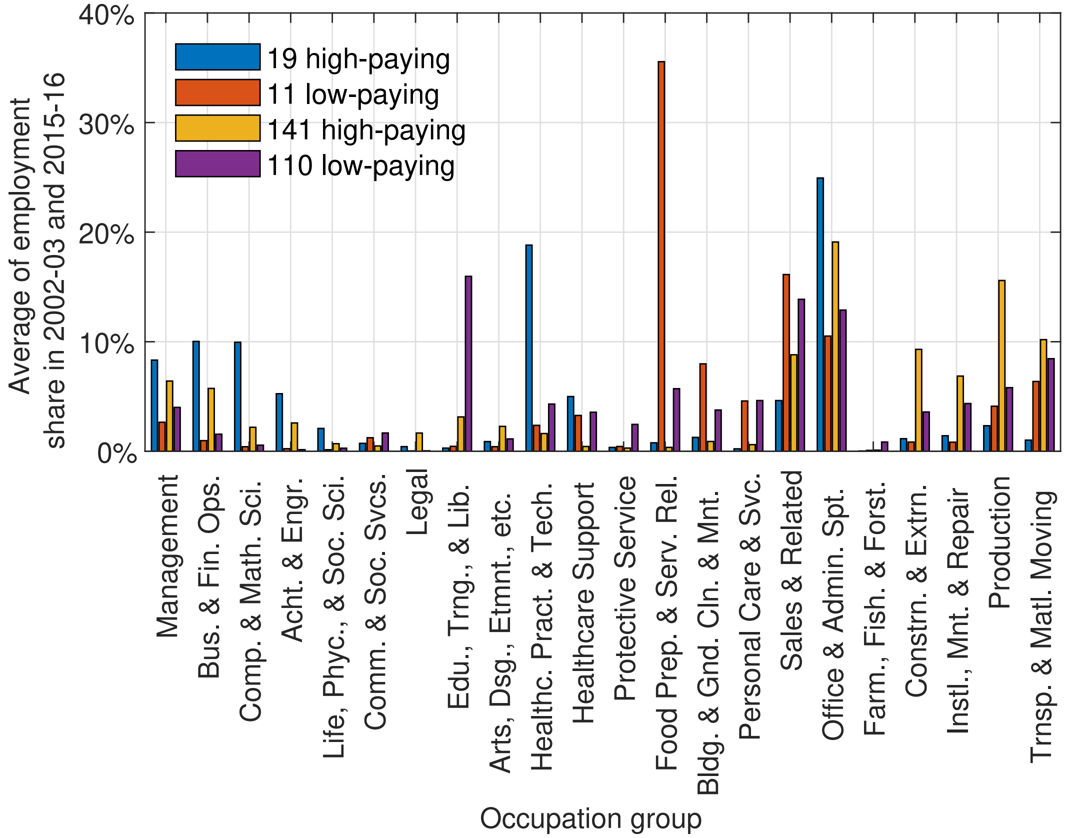
Figure G3 uses our core data to quantify the share of employment in each occupation in the top 20 4-digit industries. Recall in interpreting these findings that there 281 industries in this dataset. For the median occupation, the concentration ratio is 83%. It is lowest in occupations such as Management, Office and Administrative Support, Installation and Maintenance, Production, and Transportation and Materials Moving.

Concentration of occupations is, not surprisingly, even more pronounced for more detailed occupations. For this purpose, we use the OEWS in 2015. We only use one year to avoid the concordance issues (we only constructed a harmonized dataset from 2002-2016 at the 2-digit SOC level). There are issues with missingness but this will tend to be in cells with smaller number of firms and employment.

[12] Specifically, these are Oilseed and Grain Farming (1111), Vegetable and Melon Farming (1112), Fruit and Tree Nut Farming (1113), Greenhouse Nursery (1114), Other Crop Farming (1119), Cattle Ranching and Farming (1121), Hog and Pig Farming (1122), Poultry and Egg Production (1123), Aquaculture (1125), Other Animal Production (1129), Timber Tract Operations (1131), Forest Nurseries (1132), Fishing (1141), Support Activities Forestry (1153), Postal Service (4911), Satellite Telecommunications (5174), Monetary Auth. Central Bank (5211), Insurance Benefit Funds (5251), Other Investment Funds (5259), and Private Households (8141).

[13] These are Federal Executive Branch and US Postal Service (9991), State Government (9992), and Local Government (9993).

Figure G1: Average employment by occupation group



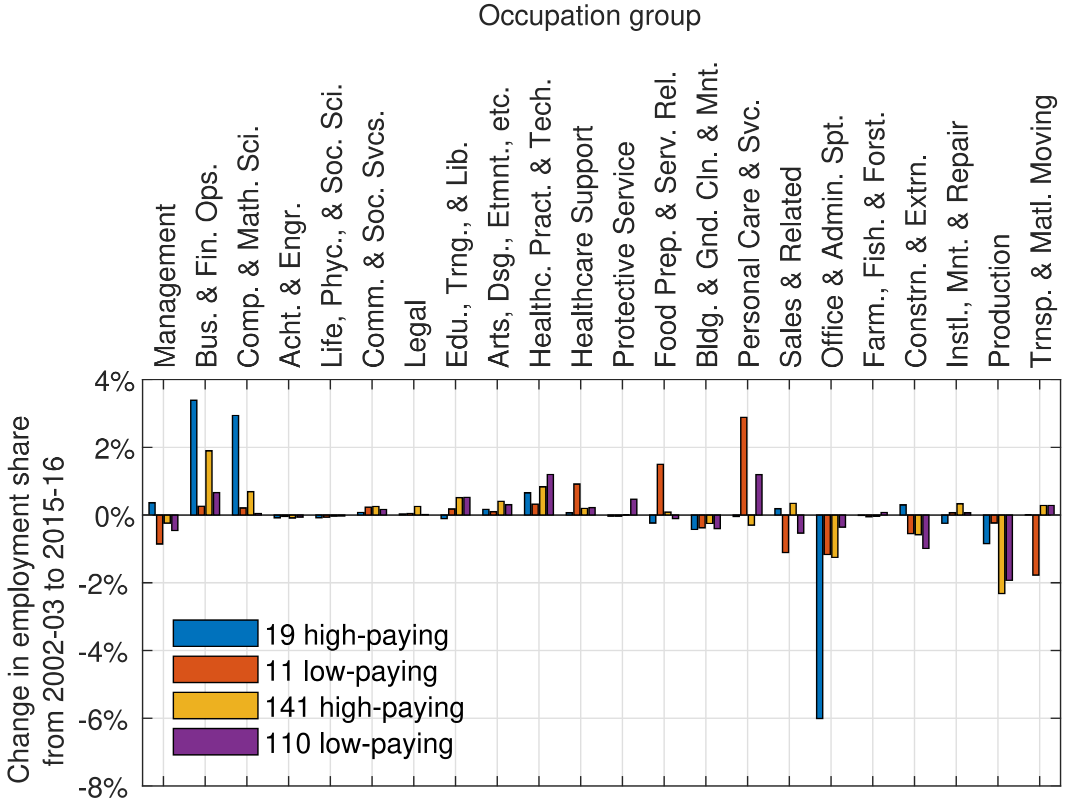
Notes: Authors' calculations of published OEWS aggregates.

Given that the next chart examines concentration for 3-digit and 6-digit occupations, only medians are reported.

The results of this exercise are shown in Figure G4. At the 3-digit occupation level, using the top 20 industries the median is 93% and the top 4 industries the median is 59%. At the 6-digit occupation level, the top 20 industries the median is 1 and for the top 4 the median is 85%. Even for truck drivers there is substantial concentration. At the 6-digit level, 81% of tractor trailer drivers are concentrated in the top 20 4-digit industries and 59% of tractor trailer drivers are concentrated in the top 4 4-digit industries: Cement and Concrete Product Manufacturing (3273), Grocery and Related Product Merchant Wholesalers (4244), General Freight Trucking (4841), and Specialized Freight Trucking (4842).

Using the OEWS, we don't have person-level information but we can estimate a related decomposition using the industry by occupation by time interval data. Specifically, we estimate the following equation for the sub-intervals 2002-03 and 2015-16 (for occupations indexed by j and industries in-

Figure G2: Change in employment by occupation group



Notes: Authors' calculations of published OEWS aggregates.

dexed by k, estimated on employment-weighted basis):

$$y_{j,k} = Occupation_{j,k}\beta_2 + Industry_{j,k}\beta_3 + \varepsilon_{j,k}. \tag{G1}$$

Taking (employment-weighted) variances of both sides of Equation (G1) yields: [14]

$$\begin{aligned}\underbrace{\mathrm{var}(y_{j,k})}_{\substack{\text{earnings}\\ \text{variance}}} = &\underbrace{\mathrm{var}(Occupation_{j,k}\beta_2 - \overline{Occupation}_k\beta_2)}_{\substack{\text{within-industry dispersion}\\ \text{from occupation}}} + \underbrace{\mathrm{var}(\overline{Occupation}_k\beta_2)}_{\substack{\text{between-industry}\\ \text{segregation}}} + \\ &\underbrace{\mathrm{var}(Industry_{j,k}\beta_3)}_{\substack{\text{between-industry}\\ \text{pay premia}}} + \underbrace{2\mathrm{cov}(\overline{Occupation}_k\beta_2, Industry_{j,k}\beta_3)}_{\text{between-industry covariance}} + \underbrace{\mathrm{var}(\varepsilon_{j,k})}_{\substack{\text{residual dispersion}\\ \text{(within-industry)}}}\end{aligned} \tag{G2}$$

[14]The notation convention for Equation (G2) is broadly similar to that for Equation (5) where for example $\overline{Occupation}_k\beta_2$ is the industry mean of $Occupation_{j,k}\beta_2$. In the current context, the employment-weighting plays a critical role. The employment-weighted estimation of Equation (G1) yields vectors of occupation and industry effects. The employment-weighting in the implementation of Equation (G2) is what yields the variation within and across industries in the contribution of occupation effects.

Table G1: Average employment by occupation and industry

	19 high-paying	11 low-paying	Other high-paying	Other low-paying
Management	8.3%	2.7%	6.4%	4.0%
Business and Financial Operations	10.0%	1.0%	5.7%	1.6%
Computer and Mathematical Science	9.9%	0.4%	2.2%	0.6%
Architecture and Engineering	5.3%	0.2%	2.6%	0.1%
Life, Physical, and Social Science	2.1%	0.1%	0.7%	0.3%
Community and Social Services	0.7%	1.2%	0.5%	1.7%
Legal	0.4%	0.0%	1.7%	0.0%
Education, Training, and Library	0.3%	0.5%	3.1%	16.0%
Arts, Design, Entertainment, Sports, and Media	0.9%	0.4%	2.3%	1.1%
Healthcare Practitioner and Technical	18.8%	2.4%	1.6%	4.3%
Healthcare Support	5.0%	3.3%	0.4%	3.6%
Protective Service	0.4%	0.4%	0.3%	2.5%
Food Preparation and Serving Related	0.8%	35.6%	0.4%	5.7%
Building and Grounds Cleaning and Maintenance	1.3%	8.0%	0.9%	3.8%
Personal Care and Service	0.2%	4.6%	0.6%	4.6%
Sales and Related	4.6%	16.1%	8.8%	13.9%
Office and Administrative Support	24.9%	10.5%	19.1%	12.9%
Farming, Fishing, and Forestry	0.0%	0.1%	0.1%	0.8%
Construction and Extraction	1.2%	0.8%	9.3%	3.6%
Installation, Maintenance, and Repair	1.4%	0.8%	6.9%	4.4%
Production	2.3%	4.1%	15.6%	5.8%
Transportation and Material Moving	1.0%	6.4%	10.2%	8.5%
OEWS suppressed employment	0.1%	0.3%	0.6%	0.4%

Notes: Authors' calculations of OEWS data.

Equation (G2) isolates the contributions of industry and occupation in overall earnings inequality. Earnings dispersion at the occupation level is separated into within- and between-industry components. Industry pay premia are common across occupations. In the main text, we combine the covariance and segregation terms into a combined sorting component.

Table G3 underlies column 4 of 8 including the break out of the covariance and segregation terms. In comparing the results for all industries and the top 30 industries using the OEWS, we obtain a larger share of the increase in between industry dispersion from industry premia using the top 30 industries. From table G3, 3.2% of between industry dispersion ($100 * 2.8/87.8$) is accounted for increasing dispersion in between industry premia. However, in table 9, rising dispersion in between industry premia accounts for roughly 20% of the rise in between industry dispersion for the top 30 industries. This difference is driven by the fact that for the OEWS for the industries outside the top

Table G2: Change in employment by occupation and industry

	19 high-paying	11 low-paying	Other high-paying	Other low-paying
Management	0.4%	-0.9%	-0.2%	-0.5%
Business and Financial Operations	3.4%	0.3%	1.9%	0.7%
Computer and Mathematical Science	2.9%	0.2%	0.7%	0.0%
Architecture and Engineering	-0.1%	0.0%	-0.1%	-0.1%
Life, Physical, and Social Science	-0.1%	-0.1%	0.0%	0.0%
Community and Social Services	0.1%	0.2%	0.3%	0.2%
Legal	0.0%	0.0%	0.3%	0.0%
Education, Training, and Library	-0.1%	0.2%	0.5%	0.5%
Arts, Design, Entertainment, Sports, and Media	0.2%	0.1%	0.4%	0.3%
Healthcare Practitioner and Technical	0.7%	0.3%	0.8%	1.2%
Healthcare Support	0.1%	0.9%	0.2%	0.2%
Protective Service	0.0%	0.0%	0.0%	0.5%
Food Preparation and Serving Related	-0.2%	1.5%	0.1%	-0.1%
Building and Grounds Cleaning and Maintenance	-0.4%	-0.4%	-0.2%	-0.4%
Personal Care and Service	0.0%	2.9%	-0.3%	1.2%
Sales and Related	0.2%	-1.1%	0.3%	-0.5%
Office and Administrative Support	-6.0%	-1.2%	-1.2%	-0.4%
Farming, Fishing, and Forestry	0.0%	0.0%	0.0%	0.1%
Construction and Extraction	0.3%	-0.5%	-0.6%	-1.0%
Installation, Maintenance, and Repair	-0.2%	0.1%	0.3%	0.1%
Production	-0.8%	-0.2%	-2.3%	-1.9%
Transportation and Material Moving	0.0%	-1.8%	0.3%	0.3%
OEWS suppressed employment	-0.1%	-0.5%	-1.0%	-0.4%

Notes: Authors' calculations of OEWS data.

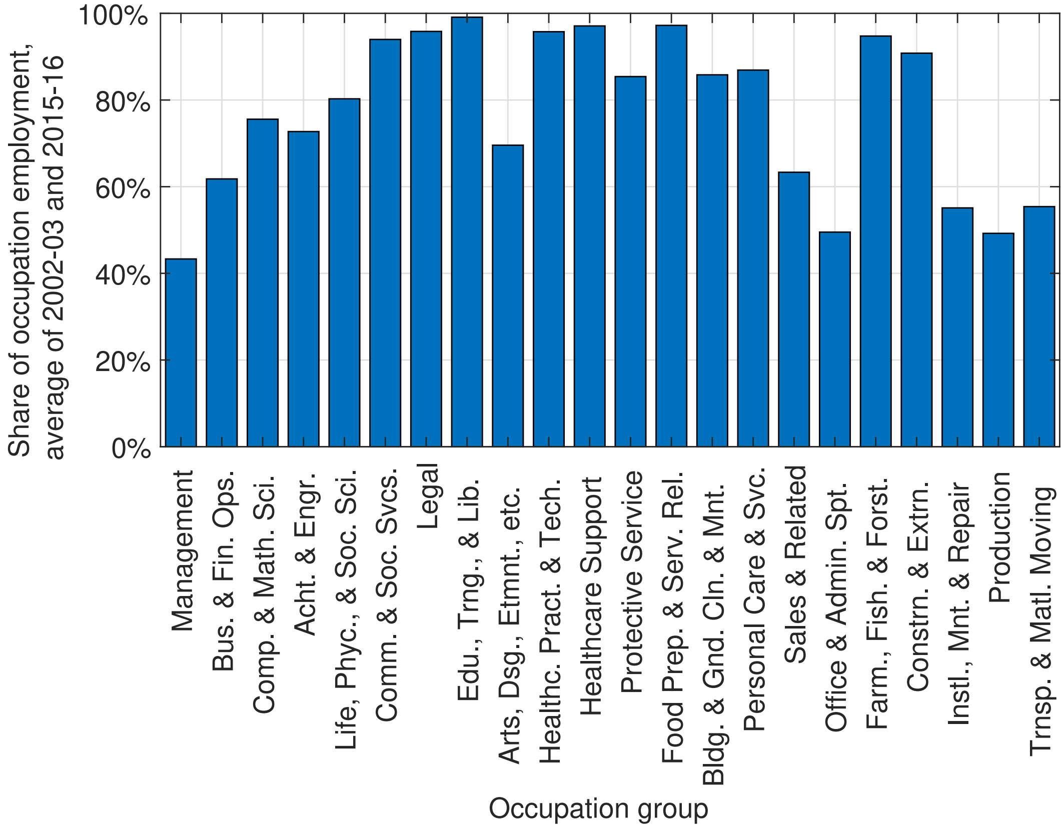
30 industries there is a modest but notable decline in dispersion in between industry premia.

We also use the decomposition of the change in the mean of industry earnings given by:

$$\underbrace{\Delta(\bar{y}^k - \bar{y})}_{\substack{\text{change in industry}\\\text{average relative earnings}}} = \underbrace{\Delta\overline{Occupation}_k \beta_2}_{\substack{\text{change attributable}\\\text{to occupation effects}}} + \underbrace{\Delta Industry_{j,k} \beta_3}_{\substack{\text{change attributable}\\\text{to industry effects}}}. \tag{G3}$$

This expression is used in Figure 6.

Figure G3: 2-digit occupational employment share, top 20 4-digit industries average (2002-2003 2015-2016)



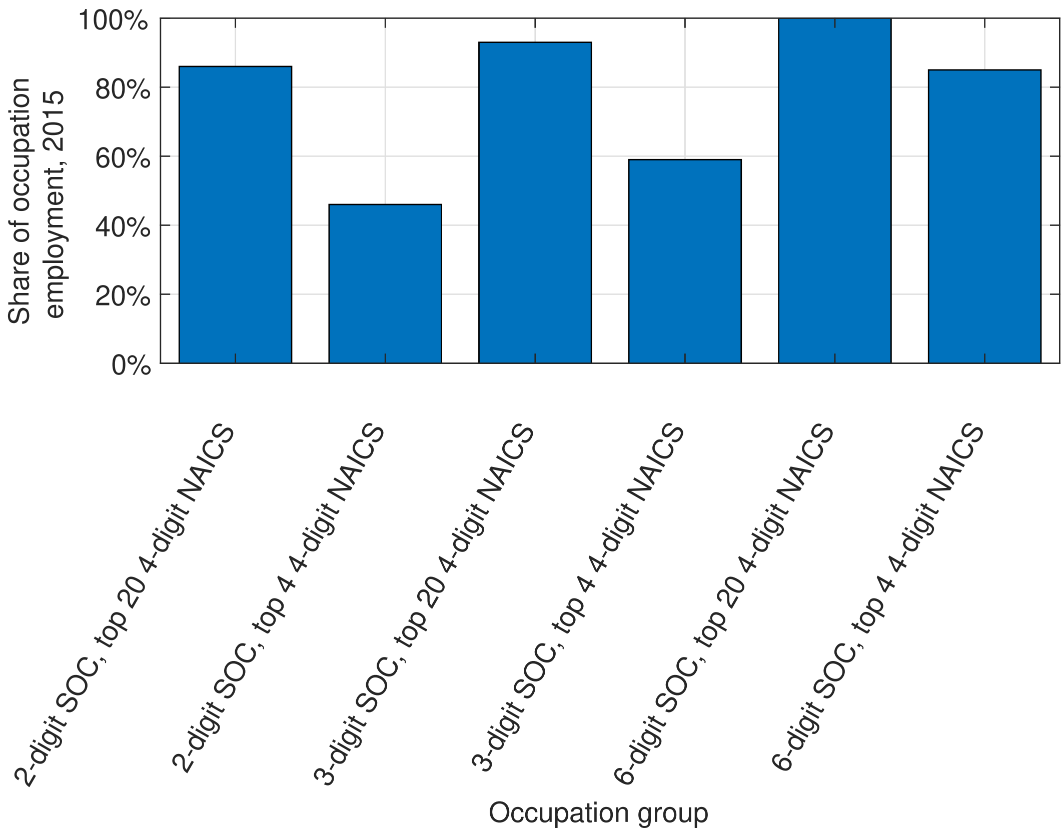
Notes: Authors' calculations of published OEWS aggregates.

Table G3: Variance decomposition using OEWS data on occupation and industry

	2002-2003	2015-2016	Change 2002-03 to 2015-16
Earnings variance	0.213	0.242	0.028
Within-industry:	49.0%	44.7%	12.2%
Occupation	41.4%	38.5%	16.0%
Residual	7.6%	6.2%	-3.8%
Between-industry:	51.0%	55.3%	87.8%
Segregation	10.9%	13.3%	31.3%
Pay premia	26.9%	24.1%	2.8%
Covariance	13.1%	17.9%	53.7%

Notes: OEWS sample for 281 4-digit industries and 22 occupations. See Equation (G2) for definitions.

Figure G4: Occupational concentration in industries, median occupation, 2015



Notes: Authors' calculations of published OEWS aggregates.

H Estimates from the Longitudinal Business Database

In this Appendix, we explore robustness and sensitivity issues using the LBD. This permits the analysis of early time periods in section H.2 but also other sensitivity issues (e.g., the sensitivity of results to using 18 states as in the LEHD analysis vs 50 states). The LBD data have the advantage of covering a longer period of time and all states.

There are some challenges in making direct comparisons between the LBD and LEHD data patterns. First, since the data are not person-level, we can compute dispersion between firms and between industries. In our analysis of the LEHD (and CPS-LEHD) data, the analysis commences with measuring log real earnings per worker in a given year. For the LBD, the analysis commences with measuring log real earnings per establishment in a given year. Then aggregation proceeds from there depending on exercise. This implies that comparisons involve differences between log of the mean vs. mean of the log where the latter is preferred. These differences will be exascerbated by within establishment dispersion in earnings. Second, our LBD based measures are based on first quarter real earnings per worker since the LBD does not have quarterly employment for quarters 2-4 prior to 2002. Third, the LBD industry codes while of similar high quality to those from the QCEW frame use a longitudinally consistent methodology described in Chow et al. (2022). In spite of these limitations, the broad patterns between the LEHD and LBD analysis are similar. In this Appendix, we take advantage of these similarities to explore robustness analyses. We first present these robustness analyses and then include some further discussion of the differences between the LBD and LEHD.

H.1 Robustness: 18 vs. 50 states

Our first robustness exercise is to use the LBD to examine the sensitivity of the between-firm and between-industry patterns to restricting to 18 states, to the 20+ firm size threshold and to using EIN or establishment as the unit of observation for the business. For establishment based calculations, we compute total between establishment variances and then between 4-digit NAICS. We do this for establishments in all 50 states and then for a restricted sample of 18 states. For the EIN level analysis selection of state and industry is based on the dominant state and industry. Note to make results more comparable that the 18 state establishment sample is the same as the 18 state EIN sample (that is states selected if they are the EIN dominant state). Also, for the EIN 20+ employment samples, this is based on national employment for the EIN. Results are shown in Table H1. As is evident, the patterns are

Table H1: Between-employer and between-industry variance

Interval	Between-employer	Between-industry	Industry share
Establishment-level, 50 state			
1996-2002	0.600	0.278	46.2%
2012-2018	0.701	0.344	49.0%
Growth	0.101	0.066	65.4%
Establishment-level, 18 state			
1996-2002	0.600	0.280	46.6%
2012-2018	0.727	0.368	50.6%
Growth	0.126	0.088	69.7%
EIN-level, 50 State			
1996-2002	0.531	0.267	50.3%
2012-2018	0.622	0.331	53.1%
Growth	0.091	0.064	69.8%
EIN-level, 18 State			
1996-2002	0.528	0.269	51.0%
2012-2018	0.642	0.353	55.1%
Growth	0.114	0.084	73.9%
EIN-level, size 20+, 50 State			
1996-2002	0.495	0.282	57.0%
2012-2018	0.600	0.355	59.1%
Growth	0.105	0.073	69.2%
EIN-level, size 20+, 18 State			
1996-2002	0.492	0.285	57.9%
2012-2018	0.625	0.382	61.2%
Growth	0.134	0.098	73.0%

Notes: Authors' calculations of LBD data.

robust to these alternatives. In all cases, rising between-industry inequality accounts for about 70% of the increase in between firm inequality.

A second robustness check is to compare the changes in the percent of employment at firms with more than 10K employees from 1996-02 to 2012-18 from the LEHD 18-state database and the Business Dynamic Statistics (BDS) which are tabulations directly from the LBD. The LEHD mega firm definition are EINs with more than 10K employees in the 18-state database. The BDS mega firm definition are Census enterprises (based on operational control) with more than 10K employees nationwide. In spite of these differences in definitions and scope, the patterns in Table H2 are remark-

Table H2: Comparison of LEHD and BDS Changes in employment shares at mega firms

Percent change from 1996-02 to 2012-18		
Top 19 high-paying	1.4	1.2
Top 11 low-paying	2.5	2.5
Bottom 125 low-paying	-0.45	0.0
Bottom 145 high-paying	-1.2	-1.8

Notes: Authors' calculations of LEHD and BDS data. Industries classified using LEHD ranking of contribution of industries.

ably similar. Our finding that the top 30 industries account for the rise in employment at mega firms is robust to using the economy-wide BDS.

The patterns for specific industries are also quite similar between the LBD and the LEHD. For this purpose, we focus on EIN-based definitions of firms and restrict EIN-based firms to having 20 or more employees. We use all 50 states. Reassuringly, the broad patterns between our results using the LEHD and LBD data are similar. Rising between-industry dispersion between 1996-2002 and 2012-2018 is 0.075 using the LEHD and 0.073 using the LBD. Moreover, this reflects 73% of the increase in between-firm dispersion using the LEHD and 69% using the LBD.

We have also used the public domain QCEW data available at the 4-digit level for additional robustness. An advantage of using the QCEW is that it is based on the same sample frame as the LEHD and the earnings measure is annual earnings per worker. The disadvantage is that the data are available in the public domain at only the 4-digit level. This implies the log of the mean vs. mean of the log aggregation issues are even more severe with the QCEW. We find that contributions of 4-digit industries to changes in between-industry dispersion in the 1996-2002 to 2012-2018 are highly correlated with both the LEHD and LBD patterns. Specifically, the correlation between LEHD and QCEW between-industry contributions is 0.83 and between LBD and QCEW is 0.80. Both of these correlations are higher than the correlation between LEHD and LBD of about 0.70. It is not surprising that the QCEW patterns have a higher correlation with the LEHD and LBD patterns than the latter two have with each other. The QCEW shares some properties with the LEHD as noted but shares the aggregation issues of using the LBD.

H.2 Between-industry inequality in earlier years

A limitation of the LEHD data is that states only provided data starting in the the 1990s. In this section, we use the LBD to examine the changing pattern of between-industry dispersion in earlier years. For

Table H3: Contribution to the change in between-industry variance (times 100), by industry group

	1980-1986 vs. 1996-2002			1996-2002 vs. 2012-2018		
NAICS Supersector	Negative	Positive	Total	Negative	Positive	Total
Natural Res. & Mining	-0.19	0.00	-0.19	-0.01	0.29	0.28
Construction	-0.10	0.02	-0.08	0.00	0.22	0.22
Manufacturing	-1.92	0.45	-1.47	-1.04	0.32	-0.72
Trade, Transp., & Util.	-0.33	1.56	1.23	-0.54	1.88	1.34
Information	-0.01	0.73	0.72	-0.28	0.78	0.50
Financial Activities	-0.02	1.95	1.93	-0.32	1.96	1.64
Prof. & Bus. Services	-0.05	1.81	1.76	-0.41	1.24	0.84
Educ. & Hlth. Services	-1.08	0.72	-0.35	-0.09	2.05	1.95
Leisure & Hospitality	-0.38	1.18	0.80	-0.08	1.25	1.17
Other Services	-0.10	0.07	-0.03	-0.09	0.14	-0.05
Total, all industries	-4.18	8.49	4.31	-2.86	10.14	7.28

Notes: Authors' calculations of LBD data.

the LBD, we can't compute the comprehensive decomposition of overall earnings inequality that is the focus of our LEHD and CPS-LEHD analyses above since the LBD does not include measures of person-level earnings. Instead, we measure earnings per worker at the establishment level and aggregate to the firm level. Accordingly, some caution is required in directly comparing the LBD and LEHD between-industry data patterns. Even with these issues, the patterns in the LBD largely mimic those from the LEHD for the 1996-2002 to 2012-2018 period. This finding provides confidence for proceeding with the analysis in this section exploring earlier years using the LBD.

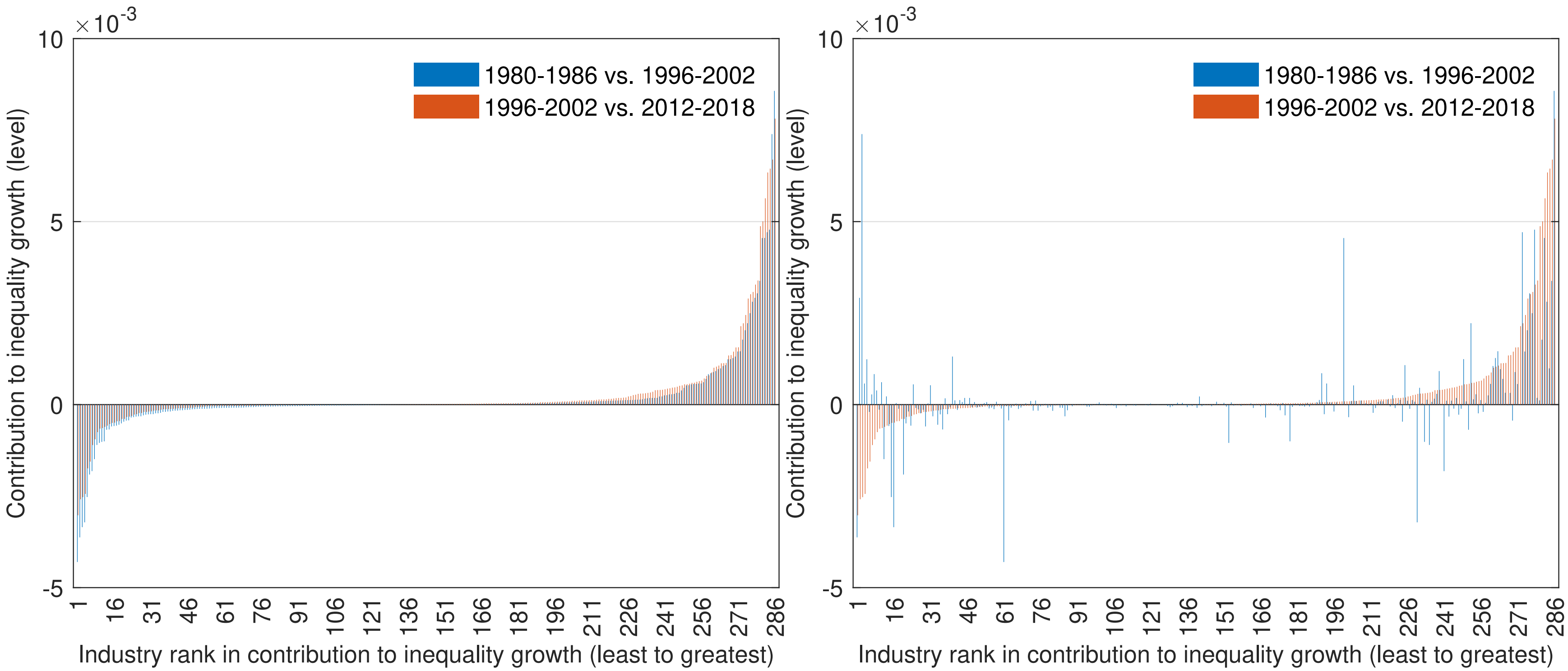
We start with Table H3, which shows the contribution to rising inequality by NAICS Supersector (multiplied by 100 to avoid excessive decimal places). The overall growth in between-industry inequality was larger (7.28 vs. 4.31) in the later than the earlier period. The lower growth in between-industry inequality reflects both less of a positive contribution (10.14 vs. 8.49) as well as more of an offsetting contribution (4.18 vs. 2.86). Manufacturing (31-33) had a mostly offsetting effect throughout the time series, but this was greater in magnitude in the earlier period. Of all Education and Health Services (61-62) changed the most, moving from the second-largest offsetting effect in the earlier period to the largest contribution in the later period.

We compare the differences in changes in between-industry earnings dispersion from 1980-1986 to 1996-2002 with those from 1996-2002 to 2012-2018 using the LBD. We find that the shape in the distributions are broadly similar with both periods exhibiting right skewness, see Figure H1(a). Both periods also have a substantial number of industries that have an offsetting (negative) impact on

Figure H1: Contribution to inequality by rank (levels) in the Longitudinal Business Database

(a) Ranking each period independently

(b) Using 1996-2002 vs. 2012-2018 rankings



Notes: Authors' calculations of LBD data. See Equation (2) for definitions.

between-industry inequality.

While the broad patterns are the same across time periods, there are some notable differences. As shown in Figure H1(b), there is considerable churning of the industries in the earlier and later periods. The correlation across these periods in the between-industry contributions is about 0.5.

Underlying this positive correlation is overlap for distinct industries. For example, Restaurants and Other Eating Places (7225) is the top contributor in both the earlier and later periods. Moreover, 21 of the top 30 positive contributors in the earlier period are also top 30 contributors as identified in the LEHD data in Section 4. Software Publishers (5112) and Computer Systems Design and Related Services (5415) consistently have among the highest contributions to rising inequality, reflecting the increasing use of computers and related devices from the 1980s until today. In contrast, Other Information Services (5191), which includes search engines and streaming services, had a limited contribution to rising inequality in the earlier period when such technologies did not exist or had minimal availability.

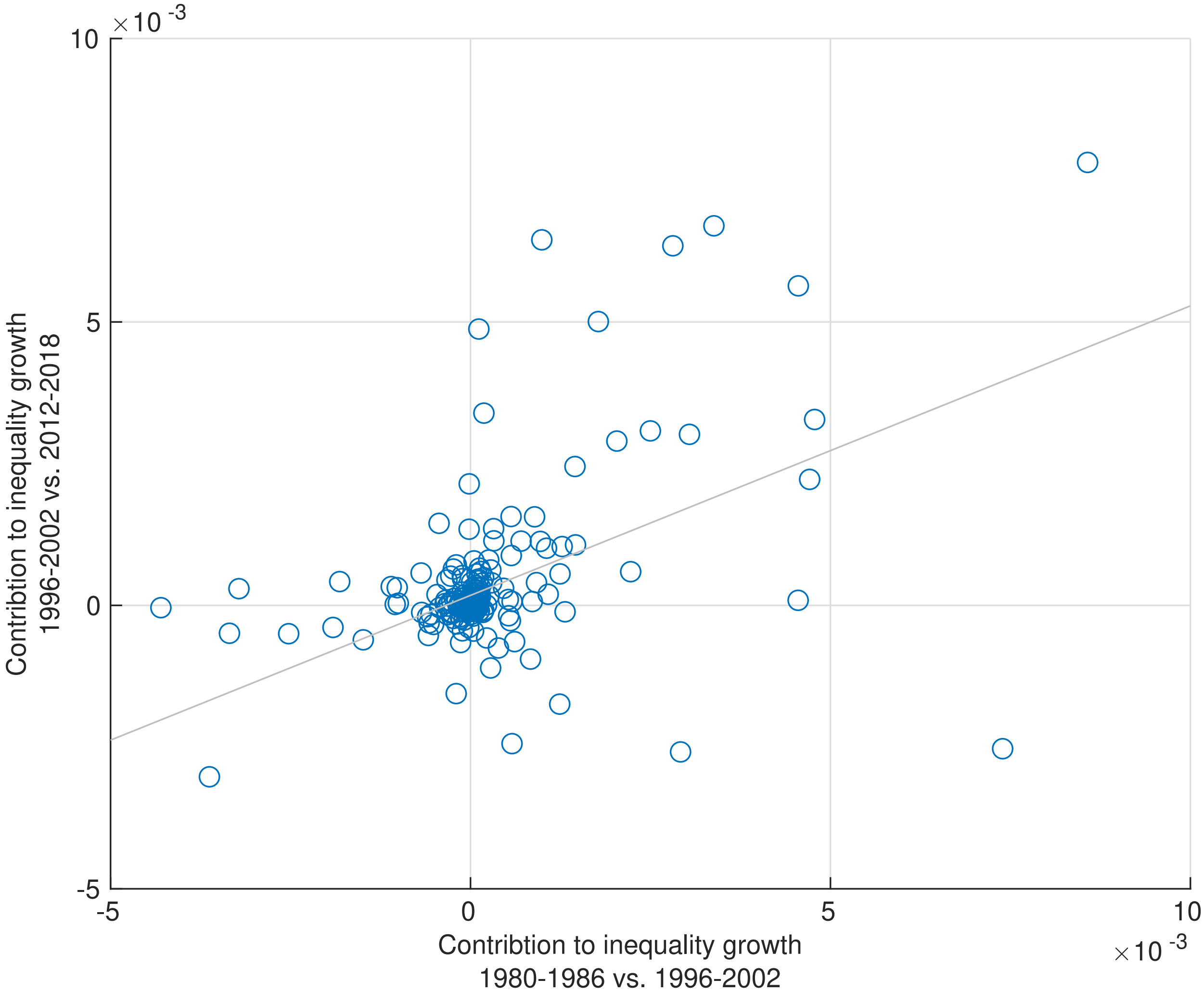
The distribution of industry contributions in Figure H1(a) are different in the earlier and later periods. From the 1980s to the 1990s, industries had more of an offsetting effect on inequality (0.042 vs. 0.023), and less of a contributing effect (0.085 vs. 0.101).[15] Thus, on net, industries contributed less to rising inequality in the earlier than the later period (0.043 vs. 0.073).

An especially important difference between the earlier and later periods is the role of manufacturing. As discussed above, manufacturing industries have positive industry earnings differentials but exhibit declining employment. In the later period, the LBD reports that manufacturing industries collectively offset about one-tenth of rising inequality. These factors were even more important in the earlier period where manufacturing industries collectively offset about one-third of rising inequality.

Focusing on the industries with the largest drag (in excess of one percent), there are eight such industries in the later period (using the LBD) and two of them in manufacturing. In the earlier period, there are twenty industries with eleven of them are in manufacturing. This strong role for manufacturing likely reflects the increasing globalization of production and associated import competition during the 1990s. For example, Motor Vehicle Parts Manufacturing (3363) and Aerospace Product and Parts Manufacturing (3364) each offset more than five percent of rising inequality. Iron and Steel Mills and

[15] Numbers in this paragraph sum the columns of Figure H1(a) as reported in Appendix Table H3. While the overall change in between-industry inequality is similar in the LBD and the LEHD (0.073 and 0.075, respectively), the LBD reports a higher magnitude of contributions and offsets to inequality.

Figure H2: Contribution to inequality by rank (levels) in the Longitudinal Business Database (scatter)



Notes: Authors' calculations of LBD data. See Equation (2) for definitions. The solid line reflects the least squares fit between the earlier and later contributions to inequality growth.

Ferroalloy Manufacturing (3311) and Cut and Sew Apparel Manufacturing (3152) each offset more than three percent. Health care also had an offsetting effect on inequality in the earlier period.

An alternative visual summary of the degree of overlap and differences in the contribution of industries is presented in a scatter plot in Figure H2. The large number of industries with little contribution in both period are evident as well as the right skewness with a small number of industries making very large positive contributions. There are more industries making a notable negative contribution in the earlier period.

References

[1] Chow, Melissa, Teresa Fort, Nathan Goldschlag, James Lawrence, Elisabeth Perlman, Martha Stinson, T. Kirk White. 2021. "Redesigning the LBD Database." CES Working Paper CES-WP-21-08.