Increasing Earnings Inequality: Reconciling Evidence from Survey and Administrative Data*

John Haltiwanger,† Henry R. Hyatt,‡ and James R. Spletzer§

September 8, 2023

Abstract

Analyses of survey data highlight observable person characteristics such as education and occupation as critical factors for rising earnings inequality, while industry has an offsetting effect. In contrast, analysis of administrative data highlights that rising between-firm earnings dispersion and, in turn, between-industry earnings dispersion dominates the rise in earnings inequality. We construct a novel integrated dataset based upon CPS microdata linked with LEHD administrative records. We find that most of the rise in earnings inequality is accounted for by rising between-industry inequality. This finding reflects a substantial contribution of increased sorting and segregation of observable person characteristics between industries.

JEL Codes: J31, J21

Keywords: inequality, industry, wage differentials, sorting, segregation, pay premium

Forthcoming, Special Issue of Journal of Labor Economics

* Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. We thank Keith Bailey, Nick Bloom, Jonathan Fisher, Fatih Guevenen, Thomas Lemieux, Christina Patterson, Kristin Sandusky, and participants in the U.S. Census Bureau research lunch, the 2022 Society of Labor Economists conference, the 2022 NBER Summer Institute, the Fall 2022 NBER Wage Dynamics in the 21st Century conference, and the 2023 Allied Social Science Associations conference for helpful comments and suggestions. All results have been reviewed to ensure that no confidential information is disclosed (CBDRB-FY22-159, CBDRB-FY22-311, CBDRB-FY22-356, and CBDRB-FY23-0245). John Haltiwanger was also a Schedule A employee of the U.S. Census Bureau at the time of the writing of this paper.

† *Corresponding author* University of Maryland. Email: halt@umd.edu

‡ Center for Economic Studies, U.S. Census Bureau. Email: henry.r.hyatt@gmail.com

§ Retired. Email: james.r.spletzer@gmail.com

1 Introduction

What drives increasing earnings inequality? Recent analyses of employer-employee matched administrative data (hereafter often referred to as administrative data) for the US shows that differences across employers drive recent increases in inequality. Song et al. (2019) demonstrate that increasing earnings inequality in the U.S. is attributable to rising between-firm dispersion. Haltiwanger, Hyatt, and Spletzer (2022), hereafter abbreviated as HHS, show that most of the rise in firm level inequality is accounted for by rising between-industry inequality from about ten percent of 4-digit NAICS industries. Using Longitudinal Employer-Household Dynamics (LEHD) data, HHS demonstrate that rising between-industry inequality accounts for more than 60 percent of increasing earnings inequality over the last several decades.

A much larger literature on US inequality uses public-use microdata from the Current Population Survey (CPS). The CPS allows researchers to examine time trends in earnings for more than half a century, and as such is a popular reference point for studying increasing inequality. Studies using the CPS have traditionally focused on the role of individual characteristics such as age, education, and gender in accounting for increasing inequality. Recent studies such as Acemoglu and Autor (2011) and Hoffmann, Lee, and Lemieux (2020, hereafter HLL) have explored how industries and occupations contribute to rising inequality. These recent studies find a supporting role for rising between-occupation differences in earnings with a modest or even a negative contribution of industry.

These two strands of the inequality literature do not provide a consistent answer regarding the contribution of workers, firms, occupations, and industries to rising inequality. Whereas studies using administrative data emphasize the importance of the firm in earnings determination and increasing inequality, the CPS is largely silent on the role of the firm (other than employer characteristics such as firm size or industry). In seeming contrast to HHS, recent studies that use CPS microdata find that industry-level differences offset rather than contribute to increasing inequality. For example, the abstract of HHS begins with, "Most of the rise in overall earnings inequality is accounted for by rising between-industry inequality," whereas HLL state (page 67) "the between-group variance component linked to industry has been declining over time," and Stansbury and Summers (2020) write (page 4) "since the 1980s there has been... a decline of about one third in the dispersion of industry wage premia."

To address these questions, we have assembled a new and unique data source. We have linked re-

spondents in the CPS Annual Social and Economic Supplement (ASEC) with administrative records from the U.S. Census Bureau's LEHD data. Using this unique integrated CPS-LEHD data, we find that we can match the finding from the administrative data that about two-thirds of the rise in earnings inequality is due to the contribution of rising between-industry inequality. This finding stems from several factors. First, studies that utilize the CPS tend to identify the marginal effect of employer characteristics such as industry conditional on already controlling for other factors such as age, education, and occupation. In other words, the conventional CPS interpretation of industry's effect does not emphasize the role of the covariances between industry and observable characteristics such as education and occupation. Indeed, we show that even when using the public-use CPS data alone (i.e., before integrating with the LEHD), industry accounts for about 23% of the increase in inequality once these covariance relationships are accounted for.

Second, we find that accurate measurement of industry is critical. The CPS industry codes are coarse and suffer from significant measurement limitations. With the linked CPS-LEHD data, we are able to replace the CPS codes with detailed (4-digit NAICS) high-quality codes from the matched employer-employee data. We find that using the CPS for all variables other than industry, rising between-industry dispersion accounts for 66% of rising inequality of earnings in the CPS when using LEHD 4-digit NAICS codes. The dominant role of rising between-industry dispersion thus holds in the CPS once methodological and measurement (especially industry coding) issues are accounted for.

Third, we build upon the recent literature to decompose the contribution of between-industry effects into sorting, segregation, and industry premia components. We use a human capital equation to distinguish these between-industry components.[1] In our setting, the variance of industry premia exclude the contributions of observable characteristics such as age, education, and occupation. Sorting captures the contribution of characteristics yielding high earnings such as education becoming increasingly concentrated in industries with positive earnings premia. Segregation captures the contribution of workers with characteristics yielding high earnings increasingly working with each other (regardless of industry premia). Using CPS earnings, we find that, of the 66% from rising between-industry dispersion, 35% is from segregation and 34% from sorting. Using administrative data earnings, the analogous contributions of sorting and segregation are 15% and 30%. While the details of

[1]See HHS for a related decomposition of the (1) within-firm, (2) between-firm, within-industry, and (3) between-industry contributions using the Abowd, Kramarz, and Margolis (1999, AKM) earnings equation instead of the human capital equation. HHS build on the decomposition of within- and between-firm proposed by Song et al. (2019). Card, Rothstein, and Yi (2022) explore the relationship between education and industry-level sorting in an AKM framework.

these components differ quantitatively, they are broadly consistent (and later in the text we discuss potential sources of these differences).

In reconciling the insights from the literature that focuses on the CPS with the recent literature using administrative data, we also identify several sources of differences between the household and administrative data. The most significant are the aforementioned differences in industry codes. We also find that the administrative data exhibits greater dispersion in earnings across individuals in the cross section. We find, like Bollinger et al. (2019), that there is "trouble in the tails" in the cross-sectional earnings distribution of the CPS. In addition, we show that the increase in earnings dispersion is about twice as large using administrative data. This finding raises questions about inferences about increasing inequality from household survey data alone.

The paper proceeds as follows. In Section 2, we build on the work of HLL to present some tabulations using public-use CPS microdata alone that allow us to begin answering several of the methodology and measurement issues involved. We show that methodology matters. In Section 3, we adjust for differences in the sample composition of how analysts often use the CPS and various administrative records. In Section 4, we describe linking the CPS and the LEHD and we present some interesting descriptive statistics on measurement differences in the two datasets. An analysis of increasing inequality in our linked CPS-LEHD dataset is presented in Section 5. Our analysis shows some unusual geographic issues in the CPS data, and we explore this in Section 6. Concluding remarks are in Section 7.

2 Industry and increasing inequality in the CPS

The analysis of inequality in the literature using the CPS focuses on observable worker characteristics along with key characteristics of the job including industry, occupation, and location. HLL conduct their own independent analysis using the CPS to help summarize the large literature using the CPS.[2] This independent analysis is an excellent synthesis of the existing literature and serves as the starting point for our own analysis.[3]

HLL find (see their Figure 3, page 63) that most of the inequality growth is due to the sum of four

[2]See HLL for the citations to the seminal contributions to the inequality literature using the CPS.

[3]We will use the CPS-ASEC data posted by HLL to the *Journal of Economic Perspectives* website to replicate and extend their results. We offer our thanks to HLL for making their data and replication code available. Unless otherwise stated, all references in this paper to the "CPS" refer to the CPS-ASEC.

variance components: (i) rising within-group dispersion for high-school-educated workers, (ii) rising within-group dispersion for college-educated workers that is greater than the increase in growth in dispersion for high-school-educated workers, (iii) rising between group dispersion for education, and (iv) composition effects linked to the shift from high-school-educated to college-educated workers. In interpreting these findings, it is important to emphasize the important role of rising within-group inequality.

Firms do not have a direct role in the household survey-based analysis but are captured indirectly via industry and location effects. Occupation effects, which have become increasingly analyzed (see, e.g., Acemoglu and Autor (2011)), capture some combination of unobserved worker characteristics and firm effects. HLL quantify the marginal contribution of occupation, industry, and location effects over and above their baseline analysis of worker characteristics. Figure 4 (page 68) of HLL shows that increasing occupation wage differentials play an important but supporting role compared to the baseline contribution from worker characteristics only. The marginal contribution of inter-industry wage differentials is actually negative after controlling for the baseline worker and occupation effects.[4] Location effects contribute little to rising dispersion.

As we discuss below, HLL's approach limits the role of occupation, industry, and location since all of the covariances with baseline worker characteristics are attributed to the latter (HLL acknowledge this in their paper).[5] We build on HLL's approach using the CPS by quantifying such covariance effects directly.

2.1 Replicate CPS results from Hoffmann, Lee, and Lemieux (2020)

We estimate the following human capital earnings equation:

$$y_i = AgeEduc_i\beta_1 + Occupation_i\beta_2 + Industry_i\beta_3 + \varepsilon_i, \tag{1}$$

where y is log real annual earnings and i is individual. $AgeEduc_i$ is a vector of dummy variables that are equal to one if worker i has that combination of age and education, and are equal to zero

[4] Stansbury and Summers (2020) also find a negative contribution of industry after controlling for person and occupation effects.

[5] HLL write on page 67: "Our objective here is to assess how much of of the rise in income dispersion can be explained by these factors, above and beyond what is already being explained by education... We note that this calculation may understate the full contribution of changing demand by occupation, industry, and location, because it does not capture the part of the contribution that is being mediated through education."

otherwise. Specifically, we allow for a separate 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}. The marginal effects of these demographic categories on earnings is given by β_1. $Occupation_i$ is a vector of dummy variables that are equal to one if worker i is employed in that occupation, and are equal to zero otherwise. The marginal effect of each of the nine occupation categories is given by the vector β_2. Analogously, $Industry_i$ is a vector of dummy variables for each of (initially) twelve SIC industries, with marginal effects given by the vector β_3.

Initially to replicate HLL, for each five-year interval {1975-1979, 1980-1984,..., 2015-2018}, we estimate the human capital earnings equation in three steps:

$$y_i = AgeEduc_i\beta_1 + \varepsilon_i \tag{2}$$

$$y_i = AgeEduc_i\beta_1 + Occupation_i\beta_2 + \varepsilon_i \tag{3}$$

$$y_i = AgeEduc_i\beta_1 + Occupation_i\beta_2 + Industry_i\beta_3 + \varepsilon_i \tag{4}$$

Equation (2) is used as the baseline equation, and measures the percentage of variance explained by age and education. We denote the marginal contribution of occupation as the additional percentage of the variance explained by equation (3) relative to equation (2). The marginal contribution of industry is obtained last by subtracting the percentage of variance explained by equation (3) from that of equation (4).

Table 1 replicates key findings from HLL (Figure 4). To facilitate comparisons with results derived from administrative data, we implement five differences: (1) we use labor income instead of total income, (2) we pool males and females instead of presenting gender specific results, (3) we use 7-year intervals instead of 5-year intervals (so we match the time intervals used by HHS), (4) we use both SIC and NAICS measures of industry, and (5) we delete the year 2000 from the data.[6] The SIC industry has 12 categories, and the NAICS measure has 18 categories. Coding of the CPS industry variable into NAICS follows Table C-5 of Pollard (2019). We find that in each 7-year interval, the contribution of age by education effects are large and growing over time. Importantly, most of the variation in earnings dispersion in levels and changes is unexplained in these CPS human capital

[6]We delete the year 2000 from our dataset because we cannot link the CPS with administrative records in this year, and so is done to ensure consistency with our later results. Appendix Table A1 replicates HLL Figure 4 more closely by using similar 5 year intervals, but similarly uses total labor income and pools males and females. These estimates are very similar to what we obtain from a direct replication of HLL, which we show in Appendix Table A2.

Table 1: Estimation of the human capital earnings equation using CPS-ASEC data

	1975-1981	1982-1988	1989-1995	1996-2002	2004-2010	2012-2018	Growth 1975-81 to 2012-18	Growth 1996-02 to 2012-18
Earnings variance	0.283	0.310	0.333	0.360	0.380	0.397	0.113	0.037
Using 12 SIC industries								
Age and education	15.5%	18.1%	21.3%	23.1%	24.2%	23.8%	44.4%	30.5%
Occupation	8.1%	7.1%	6.5%	6.4%	7.0%	6.7%	3.2%	10.0%
Industry	6.0%	5.3%	4.4%	3.6%	2.8%	2.7%	-5.6%	-5.9%
Residual	70.4%	69.5%	67.8%	67.0%	66.0%	66.9%	58.2%	65.7%
Using 18 NAICS industries								
Age and education	15.5%	18.1%	21.3%	23.1%	24.2%	23.8%	44.4%	30.5%
Occupation	8.1%	7.1%	6.5%	6.4%	7.0%	6.7%	3.2%	10.0%
Industry	7.8%	7.3%	6.1%	4.9%	4.4%	4.5%	-3.7%	0.8%
Residual	68.5%	67.5%	66.2%	65.6%	64.5%	65.0%	56.2%	58.7%

Notes: Authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. Our earnings variable is the natural log of real annual labor earnings. Our regression specification is based on HLL Figure 4, except we use labor earnings instead of total income, we pool male and females, the year 2000 is deleted, and we use different measures of industry in the top and bottom panels. Our SIC 12 follows HLL using 12 categories of the Standard Industrial Classification. NAICS 18 refers to 18 categories of the North American Industrial Classification System. Coding of CPS industry data (indly) into NAICS industries follows Table C-5 of Pollard (2019). "Age and education" is the fraction of the variance of labor earnings explained by Equation (2). "Occupation" is the marginal contribution of including occupation, obtained by subtracting the percentage of the variance explained by equation (2) from that of equation (3). "Industry" is the marginal contribution of industry, obtained by subtracting the percentage of variance explained by equation (3) from that of equation (4). "Residual" is the fraction of the variance that is unexplained when estimating equation (4).

earnings equations.

In each seven-year interval, the marginal contribution of SIC industry is positive: 6.0% of 0.283 in 1975-1981, 3.6% of 0.360 in 1996-2002, and 2.7% of 0.397 in 2012-2018. In each seven-year interval, the marginal contribution of NAICS sectors is larger than the marginal contribution of SIC industry. The marginal contribution of SIC industry is falling over time: -5.6% of 0.113 over the 1975-1981 to 2012-2018 intervals, and -5.9% of 0.037 over the 1996-2002 to 2012-2018 intervals. The marginal contribution of NAICS sectors is falling over the longer time interval (-3.7% of 0.113 over the 1975-1979 to 2012-2018 intervals), and is slightly positive (0.8% of 0.037) over the 1996-2002 to 2012-2018 intervals.

These results in Table 1 tell us that switching from SIC to NAICS switches the sign but has a relatively small effect on the marginal contribution of industry to variance growth in the CPS. Furthermore, these industry results are relatively insensitive to different time periods.

2.2 Between- vs. within-industry earnings variance decomposition

To take a step in the direction of the full variance decomposition we consider below, we explore a variant of the HLL method, changing the order in which observable characteristics are used in estimation. We first estimate industry, then add age and education, and finally add occupation. Specifically, for each seven-year interval {1975-1981, 1982-1988,..., 2012-2018}, we estimate the human capital earnings equation in three steps:

$$y_i = Industry_i\beta_3 + \varepsilon_i \tag{5}$$

$$y_i = Industry_i\beta_3 + AgeEduc_i\beta_1 + \varepsilon_i \tag{6}$$

$$y_i = Industry_i\beta_3 + AgeEduc_i\beta_1 + Occupation_i\beta_2 + \varepsilon_i \tag{7}$$

Equation (5) is used as the baseline equation, and measures the percentage of variance explained by industry. We denote the marginal contribution of age and education as the additional percentage of the variance explained by equation (6) relative to equation (5). The marginal contribution of occupation is obtained last by subtracting the percentage of variance explained by equation (6) from that of equation (7).

In putting industry first in estimating equation (5), we obtain by construction a within vs between-

industry decomposition of earnings given by:

$$\underbrace{\text{var}(y_{i,k}-\bar{y})}_{\substack{\text{earnings}\\ \text{variance}}} = \underbrace{\text{var}(y_{i,k}-\overline{y_k})}_{\substack{\text{within-industry}\\ \text{dispersion}}} + \underbrace{\text{var}(\overline{y_k}-\bar{y})}_{\substack{\text{between-industry}\\ \text{dispersion}}} \tag{8}$$

The first term on the right hand side of the above equation 8 is the within-industry component of variance, and the second term is the between-industry variance. Note that because we are keeping track of the relevant industry-level average for worker i, we add a subscript for industry k and so express earnings as $y_{i,k}$ to capture the earnings of worker i employed in industry k. The second term is equivalent to the R-squared from estimating equation (5).

We find (see Table 2) that with industry first and using NAICS industries that the contribution of between-industry is 23.1% to rising earnings inequality. This finding highlights that an important component of the contribution of observables such as age, education, and occupation in Table 1 reflects how these observables are allocated across industries. The marginal effects of age, education, and occupation reported in Table 2 reflect the within industry contributions of these variables using the sequential regression approach of equations (5) through (7). Observe that that by construction that the residuals reported in Table 2 are identical with those reported in Table 1.

Table 2 tells us that the contribution of industry to variance growth is positive (23.1%), and is substantially larger than the corresponding marginal contribution of industry in Table 1 (0.8%). This increase, from 0.8% to 23.1%, is about one-third of the difference between HLL's and HHS's industry effects. Thus, the decomposition methodology matters, and the variance decomposition in the next subsection will show us that this large difference originates from covariances that are implicit in the within and between estimate but excluded from the marginal contribution of industry.

2.3 Variance decomposition of the human capital earnings equation

To build on the results shown thus far, we re-write the human capital earnings equation (1) used by HLL as

$$y_{i,k} = Z_{i,k}\beta_Z + Industry_{i,k}\beta_3 + \varepsilon_{i,k}, \tag{9}$$

where Z concatenates the $AgeEduc_i$ and $Occupation_i$ vectors, and β_Z concatenates the marginal effects vectors β_1 and β_2. Define $\overline{Z_k\beta_Z}$ as the industry mean of $Z_{i,k}\beta_Z$. Taking variances of both sides of

Table 2: Estimation of the human capital earnings equation using CPS-ASEC data (industry first)

	1975-1981	1982-1988	1989-1995	1996-2002	2004-2010	2012-2018	Growth 1975-81 to 2012-18	Growth 1996-02 to 2012-18
Earnings variance	0.283	0.310	0.333	0.360	0.380	0.397	0.113	0.037
				Using 12 SIC industries				
Between-industry:	4.8%	4.8%	4.6%	4.2%	4.7%	5.1%	6.0%	13.8%
Within-industry:	95.2%	95.2%	95.4%	95.8%	95.3%	94.9%	94.0%	86.2%
Age and education	18.3%	20.0%	21.6%	22.9%	23.1%	22.1%	31.5%	14.0%
Occupation	6.6%	5.8%	6.0%	5.9%	6.1%	5.9%	4.3%	6.5%
Residual	70.4%	69.5%	67.8%	67.0%	66.0%	66.9%	58.2%	65.7%
				Using 18 NAICS industries				
Between-industry:	6.8%	6.7%	6.2%	5.9%	6.4%	7.5%	9.2%	23.1%
Within-industry:	93.2%	93.3%	93.8%	94.1%	93.6%	92.5%	90.8%	76.9%
Age and education	18.0%	19.9%	21.6%	22.8%	23.3%	21.9%	31.9%	13.8%
Occupation	6.7%	6.0%	6.0%	5.7%	5.9%	5.6%	2.7%	4.4%
Residual	68.5%	67.5%	66.2%	65.6%	64.5%	65.0%	56.2%	58.7%

Notes: Authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. Our earnings variable is the natural log of real annual labor earnings. Our regression specification is based on HLL Figure 4, except we use labor earnings instead of total income, we pool male and females, the year 2000 is deleted, and we use different measures of industry in the top and bottom panels. Our SIC 12 follows HLL using 12 categories of the Standard Industrial Classification. NAICS 18 refers to 18 categories of the North American Industrial Classification System. Coding of CPS industry data (indly) into NAICS industries follows Table C-5 of Pollard (2019). "Industry" is the fraction of the variance of labor earnings explained by Equation (5). "Age and education" is the marginal contribution of including age and education, obtained by subtracting the percentage of the variance explained by equation (5) from that of equation (6). "Occupation" is the marginal contribution of occupation, obtained by subtracting the percentage of variance explained by equation (6) from that of equation (7). "Residual" is the fraction of the variance that is unexplained when estimating equation (7).

the human capital earnings equation results in:

$$\underbrace{\text{var}(y_{i,k})}_{\substack{\text{earnings}\\ \text{variance}}} = \underbrace{\text{var}(Z_{i,k}\beta_Z - \overline{Z_k\beta_Z})}_{\substack{\text{within-industry dispersion}\\ \text{from age, education,}\\ \text{and occupation}}} + \underbrace{\text{var}(\overline{Z_k\beta_Z})}_{\substack{\text{between-industry}\\ \text{segregation}}} + \\ \underbrace{\text{var}(Industry_{i,k}\beta_3)}_{\substack{\text{between-industry}\\ \text{pay premia}}} + \underbrace{2\text{cov}(\overline{Z_k\beta_Z}, Industry_{i,k}\beta_3)}_{\text{between-industry sorting}} + \underbrace{\text{var}(\varepsilon_{i,k})}_{\substack{\text{residual dispersion}\\ \text{(within-industry)}}} \tag{10}$$

In equation (10), subtracting $\overline{Z_k\beta_Z}$ from $Z_{i,k}\beta_Z$ in the first term on the right hand side and then adding it back later ensures that the within- and between-industry terms in equation (10) replicate the within- and between-industry terms in equation (8). Each of the terms on the right hand side of this variance decomposition can be labeled as follows: $\text{var}(Z_{i,k}\beta_Z - \overline{Z_k\beta_Z})$ is the within-industry effect of observable person characteristics, $\text{var}(\overline{Z_k\beta_Z})$ is industry segregation, defined as how persons with similar observables (as given by $Z_{i,k}\beta_Z$) concentrate within industries, $\text{var}(Industry_{i,k}\beta_3)$ is dispersion attributable to the industry pay premia, and $2\text{cov}(\overline{Z_k\beta_Z}, Industry_{i,k}\beta_3)$ is industry sorting, defined as how frequently high-paid workers in terms of observable characteristics $\overline{Z_k\beta_Z}$ work for high-paying industries, and how frequently low-paid workers in terms of observable characteristics work for low-paying industries. This terminology is analogous to that used by Song et al. (2019) and HHS but applied here to observable CPS characteristics rather than person and firm effects.

Table 3 presents the results of this variance decomposition for 7-year intervals using NAICS industries.[7] By construction, the between-industry overall contribution is equal to the sum of industry-level segregation, industry-level pay premia, and industry-level sorting. Looking at variance growth from 1996-2002 to 2012-2018, industry segregation and industry sorting are positive (14.8% and 7.3% respectively), with the industry pay premia very small (1.0%). This pattern is similar to HHS, although the magnitudes here are smaller than in HHS: (14.8%, 1.0%, 7.3%) here, versus (25.2%, 8.7%, 28.0%) in HHS. Looking at the cross-sectional regressions, the residual accounts for 65.0% of CPS earnings variance in the 2012-2018 time period when using NAICS. This is very different than the 13.7% in HHS, who use an AKM earnings equation to decompose earnings. Looking at variance growth from 1996-2002 to 2012-2018, 58.7% of variance growth is unexplained when using NAICS. This is very different than the -3.9% in HHS.

[7] Appendix Table A3 provides a comparison of results by SIC and NAICS industries.

Tables 1 and 3 tells us that methodology matters. Moving from the marginal contribution of industry to a full variance decomposition of the human capital earnings regression increases the industry effect from 0.8% to 23.1% (roughly one-third of the difference between HLL and HHS). Covariances matter – this is evident in the segregation and sorting terms, which measure how labor composition varies across industries. Industry earnings differentials, conditional on segregation and sorting, are very small (1.0%) in our preferred specification.

These methodological issues can be interpreted in terms of the difference in the way person characteristics are treated in Tables 1 and 3. In the former, age by education plus occupation effects account for 40.5% of the increase in earnings inequality. In Table 3, within-industry person characteristics inclusive of age by education and occupation account for 18.2% of rising dispersion for the 1996-02 to 2012-18 periods. This difference is because, as noted above, in Table 3, person effects that are associated with sorting and segregation across industries have been separated into distinct terms. Adding the 18.2% with the sorting and segregation effects yields 40.3%, which is very similar to the 40.5% in Table 1. While it is not an identity that the marginal contribution of industry in Table 1 is equal to the industry pay premia in Table 3 (the covariance structures underlying the two tables are different), they are similar in magnitude.

We also provide guidance about the relative contribution of age by education vs occupation effects in Table 3. Age by education accounts for two-thirds (11.8% of 18.2%) of the within-industry person characteristics contribution to variance growth (far right column of Table 3), with the remainder accounted for by occupation and the covariance between age by education with occupation. For between-industry segregation, more than half (8.4% of 14.8%) is accounted for by the covariance between age by education and occupation, with the remainder accounted for by age by education and occupation. For between-industry sorting, almost three-quarters (5.3% of 7.3%) is accounted for by the covariance between industry and occupation, with the other one-quarter accounted for by the covariance between industry and age by education. In short, both age by education and occupation are important contributing factors for the contribution of within-industry person characteristics, between-industry segregation, and between-industry sorting. Given the attention given to education in the literature, it is instructive that over half of the contribution of age by education is associated with sorting and segregation effects.

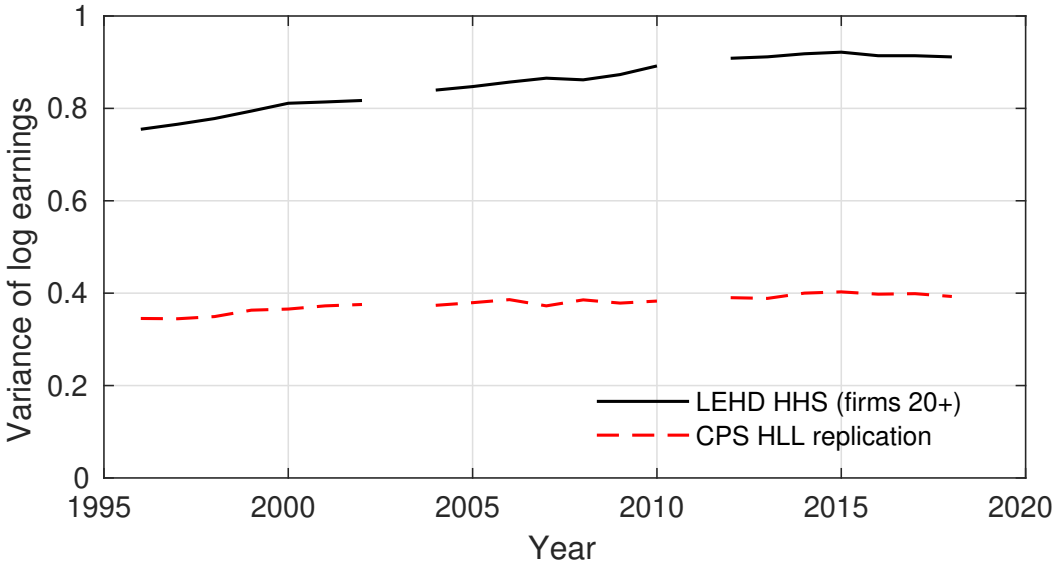
It is also worth emphasizing that the contribution of occupation is mostly via sorting and segregation effects between industries. Within-industry occupation effects contribute 6.3% ($= 4.1\% + 2.2\%$)

Table 3: Variance decomposition of the human capital earnings equation, CPS-ASEC data

	1975-1981	1982-1988	1989-1995	1996-2002	2004-2010	2012-2018	Growth 1975-81 to 2012-18	Growth 1996-02 to 2012-18
Earnings variance	0.283	0.310	0.333	0.360	0.380	0.397	0.113	0.037
Using 18 NAICS industries								
Within-industry:	93.2%	93.3%	93.8%	94.1%	93.6%	92.5%	90.8%	76.9%
Age, education & occupation:	24.7%	25.9%	27.6%	28.5%	29.1%	27.5%	34.6%	18.2%
Age and education	10.7%	11.5%	12.1%	13.0%	13.5%	12.9%	18.4%	11.8%
Occupation	7.8%	7.3%	7.5%	7.2%	7.3%	6.9%	4.7%	4.1%
Covariance: age+educ. & occ.	6.2%	7.1%	8.0%	8.2%	8.3%	7.7%	11.4%	2.2%
Residual	68.5%	67.5%	66.2%	65.6%	64.5%	65.0%	56.2%	58.7%
Between-industry:	6.8%	6.7%	6.2%	5.9%	6.4%	7.5%	9.2%	23.1%
Segregation:	2.9%	2.9%	3.2%	3.3%	3.8%	4.4%	8.1%	14.8%
Age and education	1.8%	1.7%	1.6%	1.8%	1.9%	2.0%	2.4%	3.8%
Occupation	0.6%	0.4%	0.5%	0.5%	0.5%	0.7%	0.9%	2.5%
Covariance: age+educ. & occ.	0.5%	0.7%	1.0%	1.1%	1.3%	1.8%	4.9%	8.4%
Pay premia	10.2%	9.0%	7.3%	6.0%	5.4%	5.5%	-6.2%	1.0%
Sorting:	-6.3%	-5.2%	-4.3%	-3.4%	-2.8%	-2.4%	7.3%	7.3%
Covariance: age+educ. & ind.	-4.1%	-3.4%	-2.5%	-2.2%	-2.0%	-1.8%	3.9%	2.0%
Covariance: industry & occ.	-2.3%	-1.8%	-1.8%	-1.2%	-0.8%	-0.6%	3.5%	5.3%

Notes: Authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. Pooled males and females. The year 2000 is deleted. Earnings is natural log of real annual labor earnings. The 18 NAICS aggregate industries are defined following the North American Industrial Classification System. Coding of CPS industry data (indly) into NAICS industries follows Table C-5 of Pollard (2019). Definitions follow equation (10).

Figure 1: Variance of HLL CPS-ASEC and HHS LEHD earnings, by year



Notes: "CPS HLL replication" shows authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. "LEHD HHS (firms 20+)" shows authors' tabulations of LEHD administrative records as utilized in HHS, which considers only people who work at firms that employ at least twenty people. See Table 4 for further definitions applied to select the sample in each dataset.

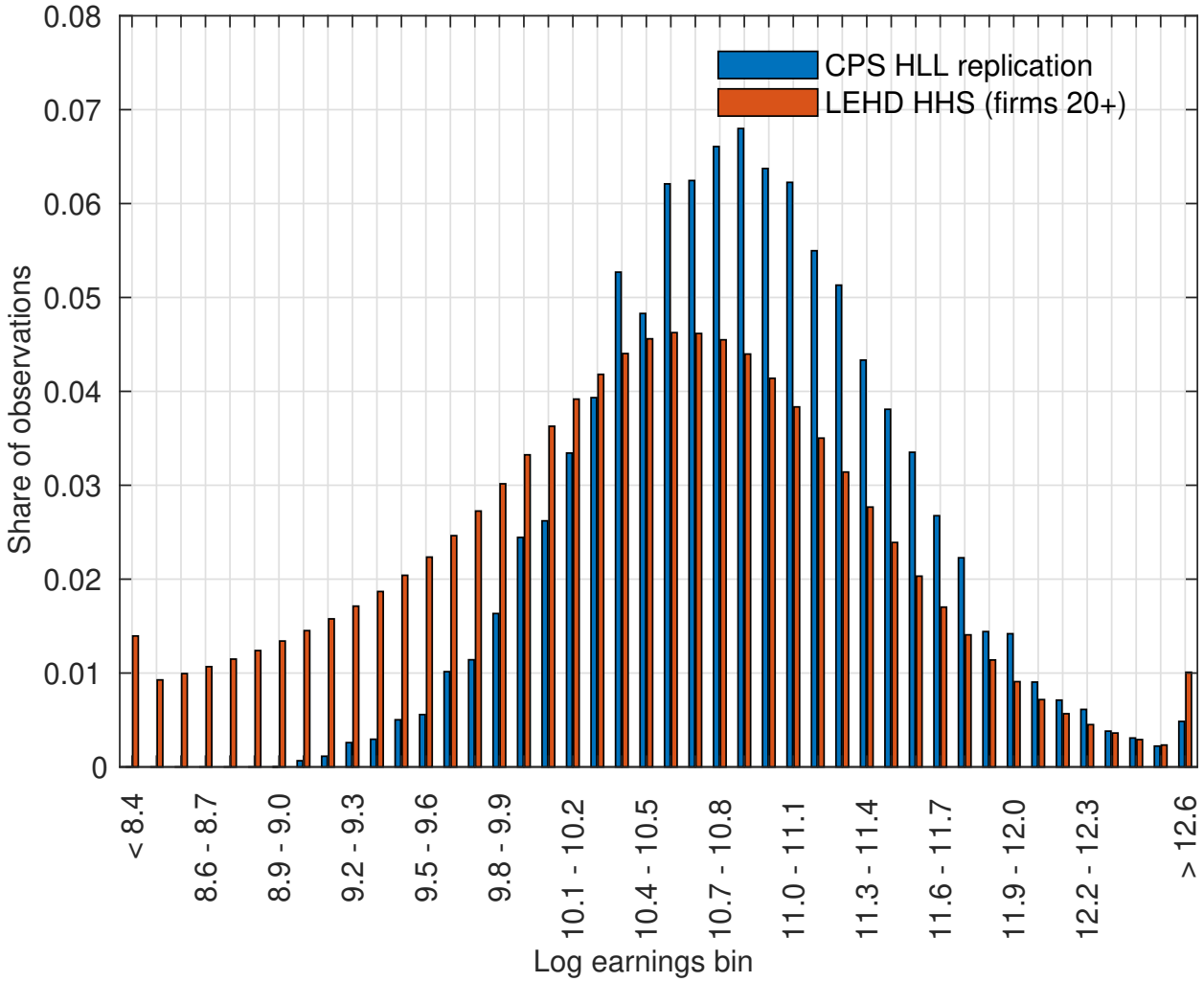
of rising dispersion including both the direct and covariance effects. Between-industry segregation effects from occupation contribute 10.9% ($=8.4\%+2.5\%$) including both direct and covariance effects. Between-industry sorting effects from occupation account for 5.3%. In total, the occupational sorting and segregation across industries accounts for 16.2% (=$10.9\%+5.3\%$) of the rise in inequality. In sum, occupations account for 22.5% (=$16.2\%+6.3\%$) of the increase in inequality.

3 Adjusting for sample selection differences used in the analysis of survey and administrative data

We seek to understand what underlies the different industry effects between the CPS data as estimated by HLL and the LEHD data as estimated by HHS. The first step is to ensure the CPS and the LEHD samples are similar, and the second step (described in Section 4) is to create a linked dataset that will allow us to examine the effects of differences in how earnings and industry are measured for a given individual.

Figure 1 shows that the earnings variance trends in the CPS used by HLL and the LEHD used by HHS are very different in levels. The variance of HLL CPS earnings is 0.393 in 2018 while the

Figure 2: Earnings distributions of HLL CPS-ASEC and HHS LEHD



Notes: Results pool all years. "CPS HLL replication" shows authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. "LEHD HHS (firms 20+)" shows authors' tabulations of LEHD administrative records as utilized in HHS, which considers only people who work at firms that employ at least twenty people. See Table 4 for further definitions applied to select the sample in each dataset.

variance of HHS LEHD earnings is 0.911 in 2018. Figure 2 shows the distributions of the HLL CPS earnings and the HHS LEHD earnings are very different. The HHS LEHD has a much larger left tail than the HLL CPS. The HLL CPS data is bottom coded at annual earnings of \$7840 (weeks worked $>$ 49, usual hours $>$ 40, and a real hourly wage $>$ \$4 using a 2018 CPI deflator), whereas the HLL LEHD data is bottom coded at annual earnings of \$3770 (weeks worked $>$ 13 and a real hourly wage $>$ \$7.25 using a 2013 PCE deflator). Figures 1 and 2 clearly show that HLL and HHS are not analyzing increasing inequality using the same annual earnings distributions.

Table 4 shows the eight identifiable differences in the HLL CPS and the HHS LEHD data. The

Table 4: Common coding of HLL CPS-ASEC and HHS LEHD data

Criterion	HLL CPS-ASEC	Common coding	HHS LEHD
Earnings	Wage & salary + self employment + farm	Wage & salary	Wage & salary
Age	26-65	20-60	20-60
Top coding	Truncate top 1% each year (by gender)	Mean of top 0.001% pooled all years	Mean of top 0.001% pooled all years
Bottom coding	Weeks worked $>$ 49 & usual hours $>$ 40 & real hourly wage $>$ \$4 & annual real earnings $>$ \$7840	Annual real earnings $>$ \$3770	Annual real earnings $>$ \$3770
Government jobs	Include all government jobs	Exclude longest job last year that is government	Exclude all government jobs
Deflator	CPI (2018=100)	PCE (2013=100)	PCE (2013=100)
Firm size	Any	Any	Firm size $\geq$ 20
Years	1975-2018	1996-2002, 2004-2010, & 2012-2018	1996-2002, 2004-2010, & 2012-2018

Notes: The columns labeled "HLL CPS-ASEC" and "HHS LEHD" are authors' summaries of the sample selection methods in the HLL and HHS papers, respectively. "Common coding" is defined by the authors, selecting criteria from either paper.

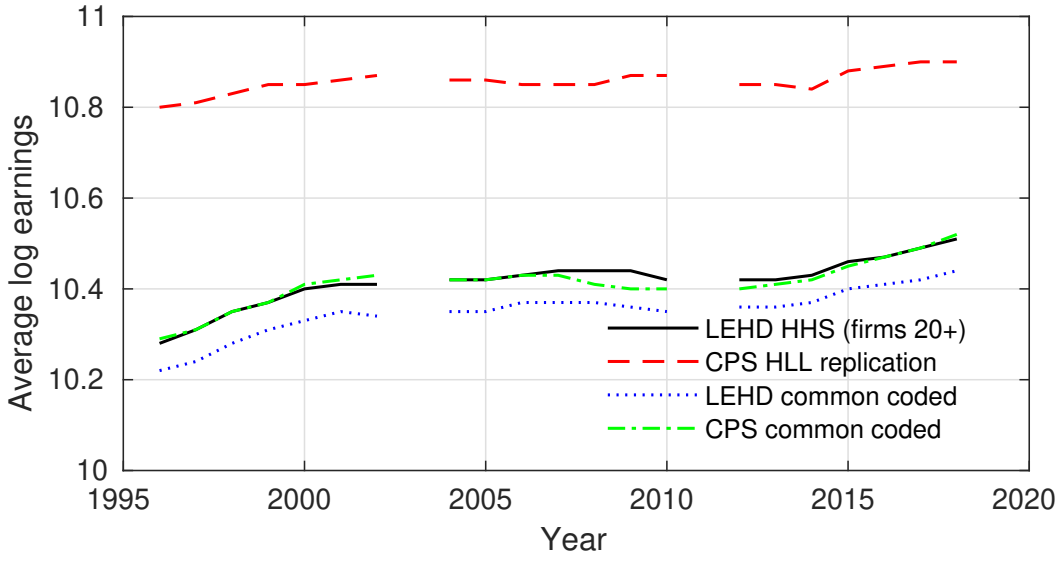
middle column of Table 4 shows how we reconcile these differences and create a "common-coded" sample for both the CPS and the LEHD. Common coding is based on: (i) earnings: exclude self-employment and farm earnings from the CPS; (ii) age: change 26-65 to 20-60 in the CPS; (iii) top coding: change 1% annual truncation to 0.001% pooled censoring in the CPS; (iv) bottom coding: change \$7840 bottom code to \$3770 in the CPS; (v) government jobs: exclude, when identified, government jobs in the CPS; (vi) deflator: change the 2018-indexed CPI to a 2013-indexed PCE; (vii) firm size restrictions: relax firm size $>$ 20 in the LEHD; (viii) years: change HLL's 1975-2018 to HHS's 1996-2002, 2004-2010, and 2012-2018.[8]

Figure 3 shows the effect of common coding on the CPS and the LEHD. Common coding decreases CPS mean earnings and increases CPS earnings variance from HLL levels. The largest con-

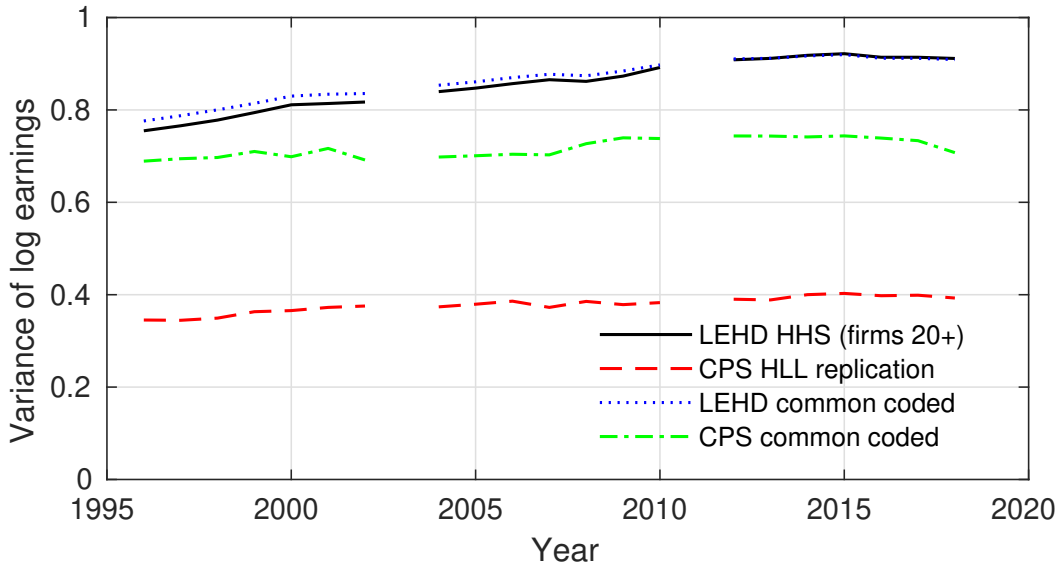
[8]Note that we use the IPUMS CPS microdata as compiled by Flood et al. (2021) to supplement the replication data that HLL posted to the *Journal of Economic Perspectives* website as not all the variables necessary for common coding were in HLL.

Figure 3: Means and variances, before and after common coding

(a) Average log earnings

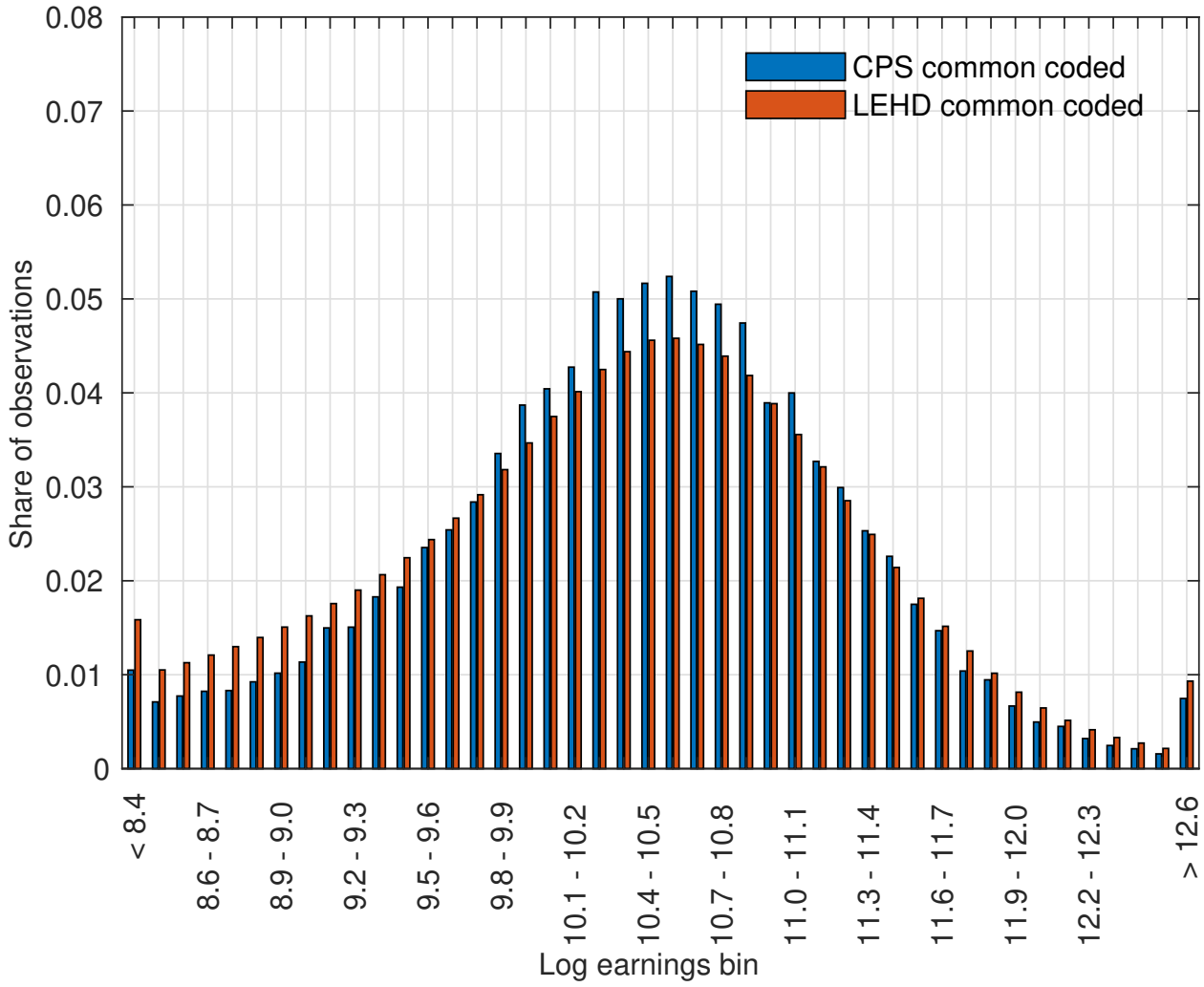


(b) Variance of log earnings



Notes: "CPS HLL replication" shows authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. "LEHD HHS (firms 20+)" shows authors' tabulations of LEHD administrative records as utilized in HHS, which considers only people who work at firms that employ at least twenty people. "Common coding" applies a consistent set of sample selection criteria to the HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website and the HHS LEHD administrative records dataset. See Table 4 for further definitions applied to select the sample in each dataset.

Figure 4: Earnings distributions of common-coded CPS-ASEC and LEHD



Notes: Results pool all years. "Common coding" applies a consistent set of sample selection criteria to the HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website ("CPS common coded") and the HHS LEHD administrative records dataset ("LEHD common coded"). See Table 4 for further definitions applied to select the sample in each dataset.

tributor to these changes is the bottom coding, where we add many lower earnings individuals to the HLL data.[9] Common coding has a small decrease on LEHD mean earnings and little if any effect on LEHD variance from HHS levels.

Figure 4 shows the distributions of the common-coded CPS earnings and the common-coded LEHD earnings. The distributions are now more similar. The common-coded LEHD has a slightly wider left tail than the common-coded CPS, which suggests that the common-coded LEHD measures more low earnings persons than does the common-coded CPS. This is consistent with the Abraham et

[9]Appendix Figure A1 shows the effects, one-by-one, of the common coding on the HLL data.

al. (2013) finding that low earnings is one characteristic predicting having an LEHD earnings record and not being measured as employed in the CPS.

Figures 1 and 2 show us that HLL and HHS analyzed different earnings distributions. Figure 4 shows us that the common-coded CPS and the common-coded LEHD have similar earnings distributions. But as we show in Section 5, common coding does not substantially change the between-industry contribution to variance growth. The between-industry contribution to variance growth in the common-coded CPS is still dramatically below that estimated by HHS. To try to further understand this, we need to link the common-coded CPS and the common-coded LEHD individual-level microdata.

4 A linked CPS-LEHD dataset

4.1 Merging the CPS and the LEHD

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.

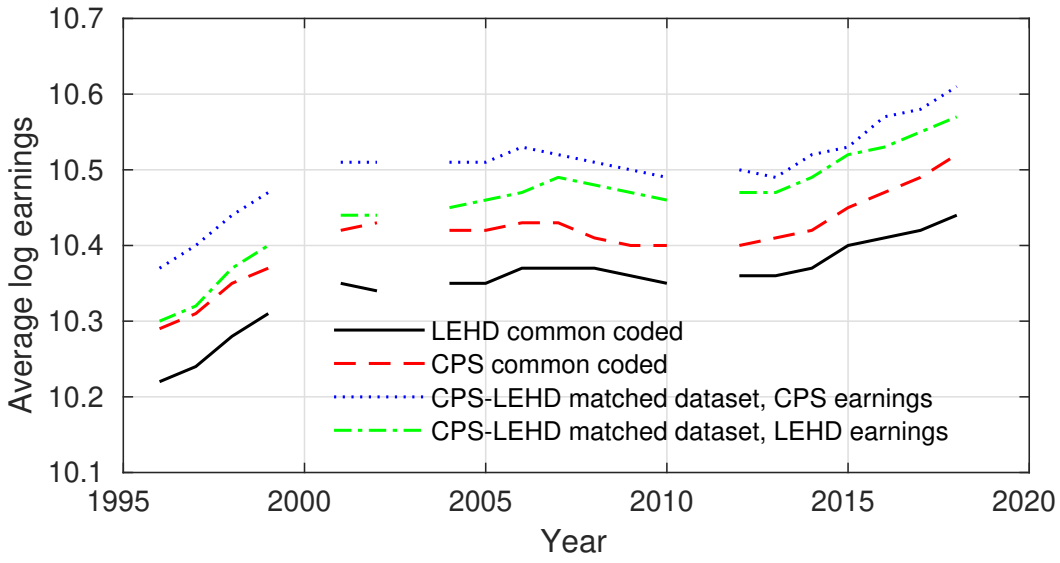
Not every record in the CPS-ASEC has a PIK attached. As noted by Bollinger et al. (2019), the Census Bureau changed its consent protocol to link respondents to administrative data beginning with the survey year 2006 CPS-ASEC. Similar to Bollinger et al. (2019), we find that the PIK rate for our common-coded CPS in the 1996 to 2004 reference years is between 60 and 80 percent, with the exception of 2000, and is then between 88 and 92 percent for reference years 2005 to 2018. The PIKs are poor quality for earnings reference year 2000, and we do not link the CPS-ASEC with the LEHD for this year.[10]

The fact that not every CPS-ASEC record has a PIK highlights the need to adjust the CPS ASEC weights with a propensity score adjustment. We have done so, running 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 output the predicted values from these regressions for each person-year observation, and then adjust the CPS-ASEC weights in the matched sample by

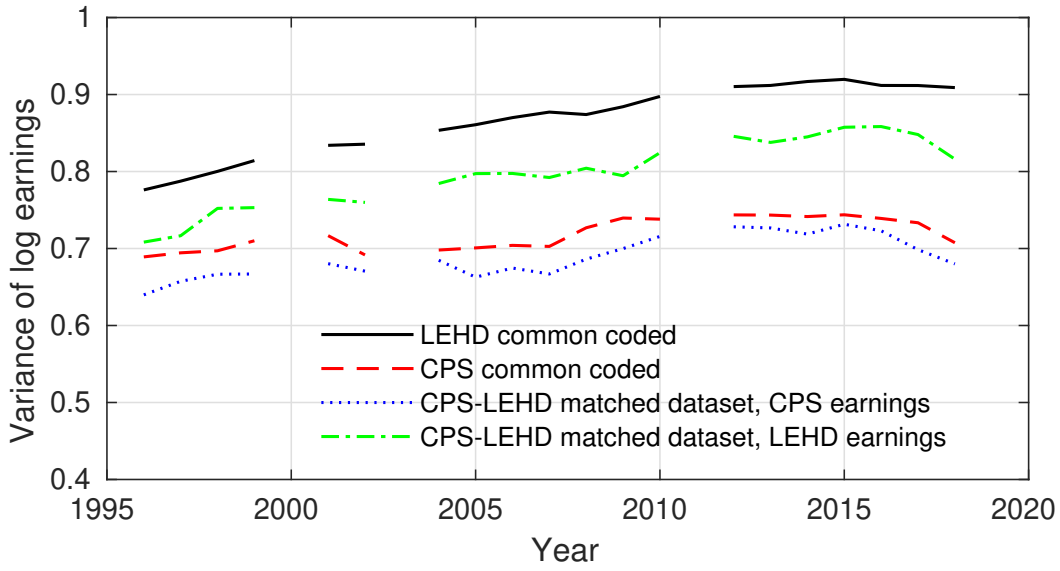
[10]The year-specific PIK rates for the common-coded CPS-ASEC are given in Appendix Figure A3.

Figure 5: Means and variances, before and after linkage

(a) Average log earnings



(b) Variance of log earnings



Notes: "Common coding" applies a consistent set of sample selection criteria to the HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website ("CPS common coded") and the HHS LEHD administrative records dataset ("LEHD common coded"). "CPS-LEHD matched dataset, LEHD earnings" reports LEHD earnings from the CPS-LEHD matched dataset. "CPS-LEHD matched dataset, CPS earnings" reports CPS earnings for the CPS-LEHD matched dataset. See Table 4 for further definitions applied to select the sample in each dataset. Further details about the construction of the CPS-LEHD matched dataset are provided in Section 4.1.

dividing the original weight by the predicted value.

We then merge the PIKed common-coded CPS-ASEC data with the common-coded LEHD data. We only keep observations where an individual is in both the CPS-ASEC and in the LEHD.[11] We run 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 common-coded LEHD is 18 states.[12] We output the predicted values from these regressions for each person-year observation, and then adjust the (already adjusted) CPS ASEC weights in the matched sample by dividing by the predicted value. All statistics from the linked CPS-LEHD data will use these twice-adjusted propensity score weights.

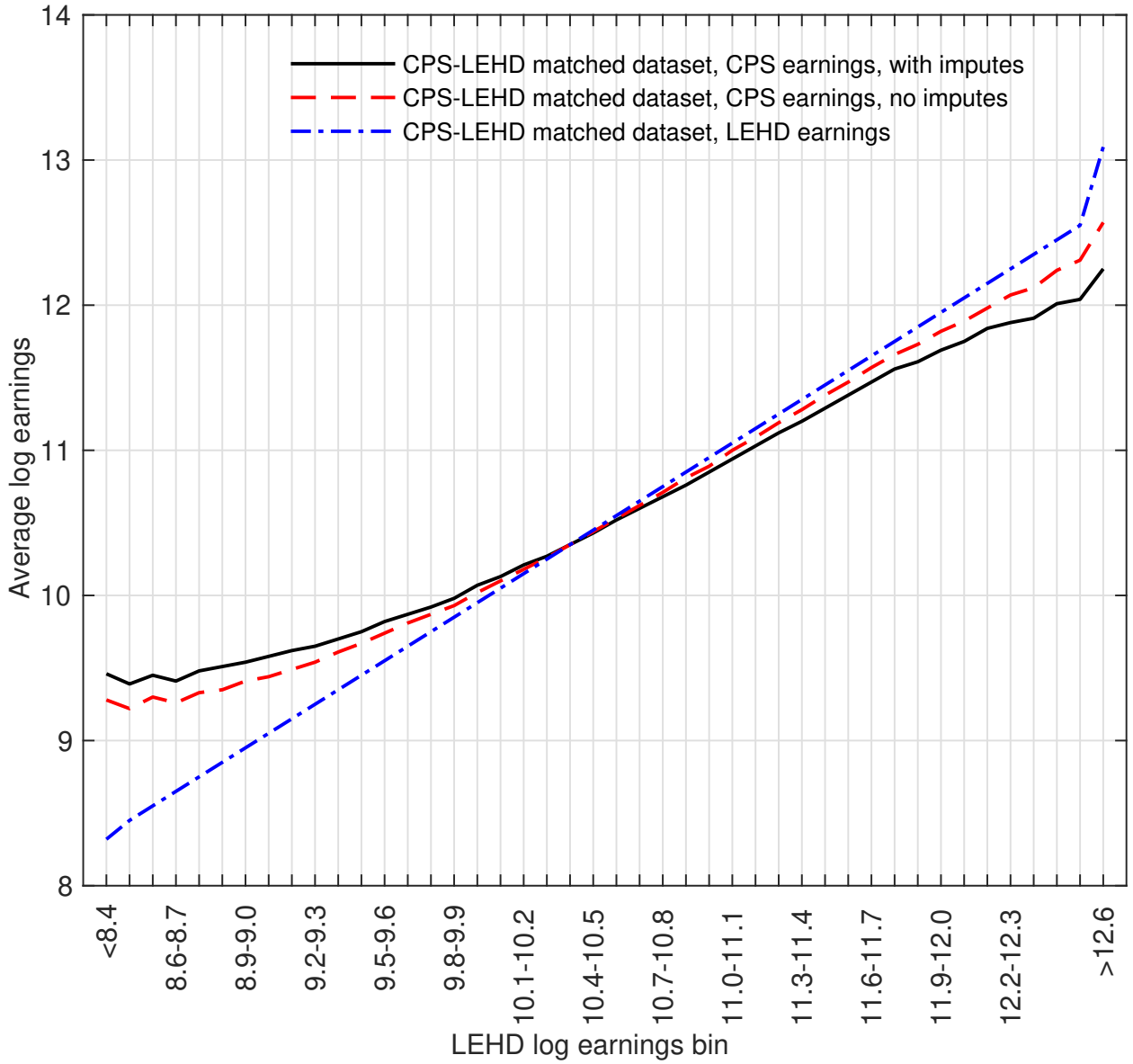
The linked CPS-LEHD data has a different earnings distribution than the two source datasets we linked. This is evident in Figure 5. Both the CPS and the LEHD in the linked CPS-LEHD have higher mean earnings and lower earnings variance than they do in the common-coded data.[13] Figure 6 provides a related perspective on differences in the tails. As in Bollinger et al. (2019) we find "trouble in the tails." Earnings per worker are low in the CPS relative to the administrative data for high earnings individuals and high in the CPS relative to the administrative data for low earnings individuals. Bollinger et al. (2019) emphasized non-response bias which we find plays some role as can be seen in Figure 6. However, even after removing imputed cases the pattern in the tails remains. These earnings differences in the tails help explain why the variance of earnings in the common-coded linked CPS are lower than the variance in the common-coded linked LEHD. Because the literature on inequality using the CPS does not exclude imputed cases, we keep them in our analysis.

[11] We do not analyze the off-diagonal cells where an individual who has an earnings record in the common-coded LEHD is not employed in the common-coded CPS, nor where an individual who is employed in the common-coded CPS has no corresponding earnings record in the common-coded LEHD. It is not correct to interpret the off-diagonals of the CPS-LEHD matching exercise as reflecting differences in employment. Our LEHD dataset contains administrative data from 18 states, so there are many individuals in the national CPS with no earnings record in the 18-state LEHD. The LEHD is a universe while the CPS is a survey, so there are many individuals in the LEHD who are not sampled in the CPS.

[12] These 18 states are: CA, CO, CT, HI, ID, IL, KS, LA, MD, MN, MT, NC, NJ, OR, RI, TX, WA, and WI.

[13] To illustrate the similarities and differences in earnings in the CPS and LEHD in the CPS-LEHD linked data, Appendix Figure A4 shows the equivalent of Figure 4 from the linked data. The two earnings distributions are roughly similar, but Appendix Figure A5 illustrates substantial differences when computing the pdf of CPS minus LEHD earnings at the individual level. In Appendix Figure A5, there is substantial mass near zero, but there are clear differences in the tails.

Figure 6: "Trouble in the tails" in the CPS earnings distribution



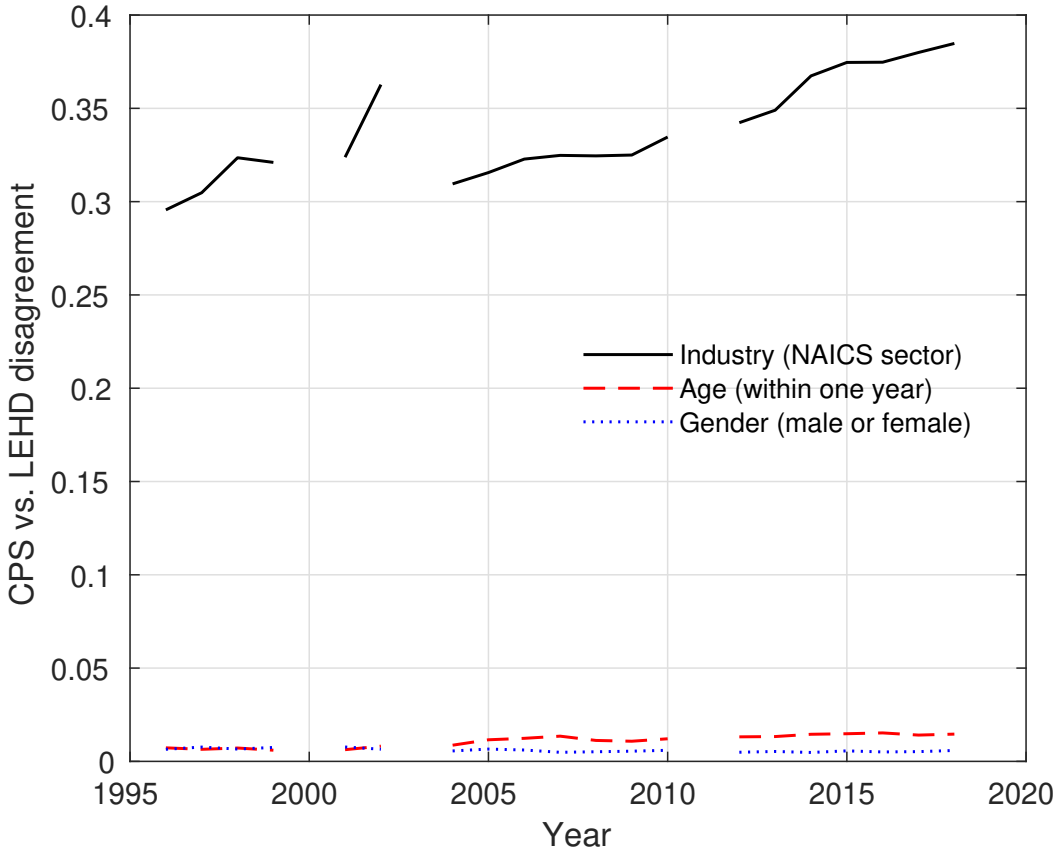
Notes: "CPS-LEHD matched dataset, LEHD earnings" reports LEHD earnings from the CPS-LEHD matched dataset. "CPS-LEHD matched dataset, CPS earnings" reports CPS earnings for the CPS-LEHD matched dataset, "with imputes" including all CPS observations, and "no imputes" excluding those CPS observations with imputed earnings. See Table 4 for further definitions applied to select a "common coded" sample. Further details about the construction of the CPS-LEHD matched dataset are provided in Section 4.1.

4.2 Measurement differences in the linked CPS-LEHD Data

We take a brief divergence here and ask about measurement differences in the linked CPS-LEHD data. Figure 7 shows the disagreement between age and gender in the CPS and the LEHD.[14] Disagreement

[14]Disagreement on age is defined as CPS age $<$ LEHD age or if CPS age $>$ LEHD age +1, since CPS age may be asked in February, March, or April of the following year.

Figure 7: Linked CPS-ASEC and LEHD differences



Notes: See Table 4 for the criteria applied to select a "common coded" sample. Further details about the construction of the CPS-LEHD matched dataset are provided in Section 4.1. Age disagreement = 0 if CPS age = LEHD age or if CPS age = LEHD age + 1, age disagreement = 1 otherwise. Industry disagreement is defined using a NAICS measure of industry with 18 categories.

on gender is always less than 1%. Disagreement on age is less than 2%, with evidence of an upward trend from 0.7% in the 1996-2002 time period to 1.4% in the 2012-2018 time period. More striking is industry disagreement in the CPS and the LEHD, where industry is measured as 18 sectors following Pollard (2018). 30%-40% of persons disagree on industry sector, with an upward trend.

This large disagreement on NAICS industry sector in the linked CPS-LEHD data suggests that differences in the industry variable could be one source for why the CPS and LEHD have different between-industry contributions to variance level and growth. The LEHD industry measures are of high quality from the establishment-level programs at BLS and Census. These agencies have a strong incentive to track industry carefully as their detailed industry statistics are critical for the NIPAs and productivity statistics. CPS industry is based on self-reported descriptions by the respondent that are coded into sectors. Limitations of the CPS industry codes are well-known (e.g., Mellow and Sider

(1983) and Dey et al. (2010)).[15]

5 Increasing inequality in the linked CPS-LEHD dataset

5.1 Within and between-industry variance decompositions

We begin our analysis of inequality in the linked CPS-LEHD data by focusing on within and between-industry contributions. We start here given the large differences discussed above in industry contributions from the CPS and LEHD datasets. Column 1 of Table 5 presents the variance decomposition using our minor modifications of HLL's CPS-ASEC sample (this is the same sample used in Tables 1 to 3). Column 2 presents the variance decomposition from the common-coded CPS. Comparing columns 1 and 2, the between-industry variance level increases, from 7.5% to 13.9% in 2012-2018, and the between-industry variance growth increases, from 23.1% to 29.3%. Column 3 presents the variance decomposition from the linked CPS-LEHD data. Compared to column 2 (the common-coded CPS), the between-industry variance level increases only slightly, from 13.9% to 14.7% in 2012-2018, but the between-industry variance growth increases substantially, from 29.3% to 46.0%. One key observation from column 3 is that using the linked CPS-LEHD data increases the growth in earnings inequality substantially in the direction of the greater increase in dispersion in the administrative data (from 0.035 in the common-coded CPS to 0.050 in the linked data). We discuss these issues further below.

Column 4 uses the same CPS-LEHD linked sample as column 3, but uses a measure of NAICS sector from the LEHD rather than from the CPS. This increases the between-industry variance growth from 46.0% to 52.2%. Column 5 uses a four-digit NAICS measure with 299 categories from the LEHD rather than the sector level with 18 categories. This has large effects on the between-industry variance level (from 11.4% to 20.9% in 2012-2018) and also has large effects on the between-industry variance growth, from 52.2% to 65.5%. Column 6 changes the earnings measure from the CPS to the LEHD. The between-industry variance level increases (from 20.9% to 26.9% in 2012-2018), but the between-industry variance growth is unaffected.

[15] Appendix Table A4 presents evidence on the extent of disagreement by specific sectors. Four sectors stand out as having especially low agreement. Wholesale Trade, Information, Educational Services, and Other Services all have less than a 50 percent agreement between CPS and LEHD industry codes. As we will see below, the Information sector is especially important in accounting for discrepancies between CPS tabulations using an 18 vs 50 state sample.

Table 5: Within and between-industry variance decompositions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Data	CPS	CPS	Linked CPS-LEHD	Linked CPS-LEHD	Linked CPS-LEHD	Linked CPS-LEHD	LEHD
Sample	HLL JEP	Common coded	Common coded	Common coded	Common coded	Common coded	Common coded
Earnings measure	CPS	CPS	CPS	CPS	CPS	LEHD	LEHD
Industry measure	CPS 18	CPS 18	CPS 18	LEHD 18	LEHD 299	LEHD 299	LEHD 299
				Variance level 1996-2002			
Earnings variance	0.360	0.703	0.667	0.667	0.667	0.746	0.811
Within-industry	94.1%	86.9%	87.6%	91.7%	82.5%	78.3%	79.6%
Between-industry	5.9%	13.1%	12.4%	8.3%	17.5%	21.7%	20.4%
				Variance level 2012-2018			
Earnings variance	0.397	0.738	0.717	0.717	0.717	0.845	0.914
Within-industry	92.5%	86.1%	85.3%	88.6%	79.1%	73.1%	74.6%
Between-industry	7.5%	13.9%	14.7%	11.4%	20.9%	26.9%	25.4%
				Change from 1996-02 to 2012-18			
Variance growth	0.037	0.035	0.050	0.050	0.050	0.100	0.103
Within-industry	76.9%	70.7%	54.0%	47.8%	34.5%	33.8%	35.5%
Between-industry	23.1%	29.3%	46.0%	52.2%	65.5%	66.2%	64.5%

Notes: "HLL JEP" shows authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. "Common coding" applies a consistent set of sample selection criteria to the HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website and the HHS LEHD administrative records dataset. See Table 4 for further definitions applied to select the sample in each dataset. Further details about the construction of the CPS-LEHD matched dataset are provided in Section 4.1. "Earnings measure" indicates whether CPS or LEHD earnings is used in the decomposition. "Industry measure:" "CPS 18" refers to 18 NAICS sectors from the CPS-ASEC (recoding CPS-ASEC variable indly following Table C-5 of Pollard 2019), "LEHD 18" refers to NAICS sectors from the LEHD, and "LEHD 299" refers to 299 4-digit NAICS industries from the LEHD. Definitions follow equation (8).

An important finding is the large increase in variance growth between columns 5 and 6, from 0.050 when using the CPS earnings measure to 0.100 when using the LEHD earnings measure. As we discuss above, there is trouble in the tails of the CPS earnings distribution which acts to suppress both the level and growth in the variance of earnings. These are reasons to prefer the results using LEHD earnings. In any event, we have much more confidence whether using CPS or LEHD earnings in using LEHD based industries. Finally, in column 7, we show the variance decomposition using the full common-coded LEHD with 946 million person-year observations rather than the CPS-LEHD linked sample. The contribution of the between-industry variance to total variance levels and growth from the full LEHD match the contribution from the weighted CPS-LEHD linked sample very closely. We regard Table 5 as showing a remarkable result. The between-industry variance growth in the full LEHD (18-states) is 64.5%, and we can essentially replicate this statistic from our linked CPS-LEHD data when using CPS earnings and a LEHD 4-digit industry measure.

5.2 A full variance decomposition of the human capital earnings equation

Table 6 shows the variance decomposition of the human capital earnings equation with columns corresponding to Table 5 (there is no column 7 with results from the full LEHD because the full LEHD does not have measures of the CPS explanatory variables Z). We focus on the decomposition of the growth of inequality.[16] Following the numbering from Table 5, we are primarily interested in columns (5)-(6) that use the linked, common-coded CPS-LEHD data with the detailed LEHD industry codes. Column (5) shows results using CPS earnings, and column (6) shows results using LEHD earnings.

The results from columns (5) and (6) have some important similarities and differences. One key similarity that we already know from Table 5 is that the overall between-industry contribution to rising dispersion is very similar using either CPS or LEHD earnings and is very large. Another similarity is that person characteristics within industries contribute only modestly to rising dispersion. Person characteristics (inclusive of occupation) are more important through their contributions to between-industry sorting and segregation. However, the relative importance of industry premia and segregation are very different when using CPS or LEHD earnings. Segregation is much more important using CPS earnings (34.9% compared to 15.3%) while industry pay premia is much more important using LEHD

[16]Level decompositions for the two subperiods are available in Appendix Tables A5 and A6, and additional growth decompositions are in A7. One striking finding in Appendix Table A7 is that moving from CPS to LEHD industry codes substantially reduces the unexplained portion of the increase in CPS earnings inequality (compare 36.1% in column (3) to either 24.1% or 26.7% in columns (4) and (5)).

Table 6: Variance decomposition of the human capital earnings equation

	(5)	(6)
Data	Linked CPS-LEHD	Linked CPS-LEHD
Sample	Common coded	Common coded
Earnings measure	CPS	LEHD
Industry measure	LEHD 299	LEHD 299
Change from 1996-02 to 2012-18		
Variance growth	0.050	0.100
Within-industry:	34.5%	33.8%
Age, education & occupation:	7.8%	13.6%
Age and education	12.2%	14.9%
Occupation	-1.8%	-0.4%
Covariance: age+educ. & occ.	-2.6%	-0.9%
Residual	26.7%	20.3%
Between-industry:	65.5%	66.2%
Segregation	34.9%	15.3%
Age and education	12.7%	6.2%
Occupation	4.6%	1.9%
Covariance: age+educ. & occ:	17.5%	7.1%
Pay premia	-3.4%	21.4%
Sorting	34.1%	29.5%
Covariance: age+educ. & ind.	20.5%	19.4%
Covariance: industry & occ.	13.5%	10.1%

Notes: See Table 4 for further definitions applied to select a "common coded" sample. Further details about the construction of the CPS-LEHD matched dataset are provided in Section 4.1. "Earnings measure" indicates whether CPS or LEHD earnings is used in the decomposition. "Industry measure:" "LEHD 299" refers to 299 4-digit NAICS industries from the LEHD. See equation (10) for definitions.

earnings (21.4% compared to -3.4%). Another important difference is the contribution of unexplained factors is larger using the CPS compared to the LEHD earnings. In interpreting these similarities and differences in percent contributions, it is also important to remember that the increase in dispersion for the CPS earnings measure is only about half that of the LEHD earnings measure.

We further provide a breakout of the contribution of age by education and occupation effects in Table 6. Whether using either CPS or LEHD earnings, age by education is more important than occupation but in both cases more than half of the contribution of each is accounted for by between-

industry sorting and segregation effects.

To help put these findings from the linked CPS-LEHD into perspective, it is instructive to discuss differences between these findings and those in HHS that focus on using an AKM decomposition of earnings. Estimating AKM effects requires the full LEHD. HHS find, using the full LEHD (restricted to firms with 20 or more employees), that 62% of rising earnings inequality over the same time periods is accounted for by rising between-industry dispersion. Using AKM firm and worker effects, they find that 28% of the between-industry contribution is due to sorting (high person effects increasing working in high industry premium industries), 25% is due to segregation (high person effect workers increasingly working together) and 9% due to the industry pay premium. These findings are broadly similar to those reported here but with a more substantial role for between-industry segregation.As discussed in HHS, this stems from the AKM decomposition of earnings yielding a substantially smaller residual compared to using observable characteristics from the CPS.

6 Geography issues in the CPS-LEHD linked data

Our analysis has mainly focused on the CPS-LEHD linked data. To construct this harmonized and integrated survey and administrative dataset, we linked the national CPS to the LEHD constructed from 18 states.[17] The linked data permit us to make an apples-to-apples comparison of survey and administrative data for a large sample over an extended period. Still, as we noted above, there are some notable changes when we move from the common-coded CPS for all states to the linked CPS-LEHD sample.

In this Section, we consider the 18 vs 50 state differences for the CPS. In related work (see HHS), we present evidence that key aspects of the inferences from administrative data are robust to using 18 vs 50 states. First, the HHS 18-state total variances and between-firm variances match the Song et al. (2019) 50-state results almost exactly for roughly similar time periods. Second, we compare the percentage of the between-firm variance that is between industries in the 18 state LEHD to the Census Bureau's 18- and 50-state Longitudinal Business Database (LBD). In the 18-state LEHD, 73% of the rising between-firm dispersion is accounted for by rising between-industry dispersion. In the 18-state

[17]We do not restrict the CPS to 18 states before linking because the geography in the CPS is place of residence and geography in the LEHD is place of work. There are plenty of individuals who work and live in different states; examples are New York and New Jersey, and the Washington DC metro area comprised of Maryland, Virginia, and the District of Columbia.

Table 7: Variance decompositions for common-coded CPS, 18 vs. 50 States

	50 states	18 states
Variance level 1996-2002		
Earnings variance	0.703	0.725
Within-industry	86.9%	86.6%
Between-industry	13.1%	13.4%
Variance level 2012-2018		
Earnings variance	0.738	0.762
Within-industry	86.1%	85.3%
Between-industry	13.9%	14.7%
Change from 1996-02 to 2012-18		
Variance growth	0.035	0.036
Within-industry	70.7%	59.6%
Between-industry	29.3%	40.4%

Notes: See Table 4 for further definitions applied to select the "common coded" sample in the CPS dataset. The District of Columbia is included in "50 states."

and 50-state LBD, the analogous statistics are 74% and 73%, respectively.[18] Our focus in this Section is thus the sensitivity of the CPS to using 18 vs 50 states.

To investigate this issue, we first return to the common-coded CPS and compute the components of Table 5 for the 18 states that are in the LEHD (that is, the CPS for 18 states without restricting to being linked to LEHD). The first column of Table 7 repeats the results from column 2 of Table 5 and the second column shows the results for the 18 state CPS sample. While patterns are broadly similar, this exercise shows that the between-industry contribution to the change in the variance is higher in the 18 state sample (40.4%) compared to the 50 state sample (29.3%). The implication is that at least for the CPS there are geographic differences in the contribution of between-industry effects using broad sectoral definitions of industry.

We ask why the between-industry variance growth is so different in the 50 versus 18 state data (0.0103 in 50 states, 0.0147 in 18 states). The between-industry variance growth can be written as $\Sigma_{k=1}^{18}\Delta[(N_k/N)(\bar{w}_k-\bar{w})^2]$, where k indexes NAICS industry sectors. We list the 18 industry contributions to between-industry variance growth in Table 8. Two sectors stand out as accounting for the

[18]These LEHD and LBD calculations use the restrictions in HHS (e.g., restricting to firms with 20+ employees). We find very similar results removing those restrictions.

Table 8: Industry contributions to between-industry growth, 18 vs. 50 states

Industry	CPS 50 states + DC	CPS 18 states	Difference 18 minus 50 States	Changing employment shares	Changing earnings differentials
Information	0.0054	0.0082	0.0028	-0.0001	0.0029
Retail Trade	0.0035	0.0054	0.0019	0.0002	0.0017
Manufacturing	-0.0021	-0.0013	0.0008	-0.0002	0.0010
Educational Svcs.	-0.0005	-0.0003	0.0003	0.0001	0.0002
Hlth. Cr. & Soc. As.	0.0004	0.0006	0.0002	0.0000	0.0002
Arts, Ent., & Rec.	0.0001	0.0003	0.0002	0.0001	0.0001
Accm. & Food Svcs.	0.0082	0.0084	0.0002	0.0002	0.0000
Finance & Insurance	0.0050	0.0051	0.0001	-0.0002	0.0003
Mining	0.0015	0.0017	0.0001	0.0004	-0.0003
Transp. & Wareh.	-0.0007	-0.0007	0.0001	0.0000	0.0001
Utilities	-0.0001	-0.0001	0.0000	0.0001	0.0000
Real Est. & Rt. & Ls.	0.0002	0.0002	0.0000	0.0000	0.0000
Construction	0.0000	-0.0001	-0.0001	0.0000	-0.0001
Other Svcs.	-0.0005	-0.0006	-0.0001	0.0000	0.0000
Prof. & Bus. Svcs.	0.0010	0.0008	-0.0002	-0.0001	-0.0002
Wholesale Trade	-0.0005	-0.0007	-0.0002	0.0000	-0.0002
Agriculture	-0.0008	-0.0015	-0.0007	-0.0003	-0.0005
Unknown (2nd job)	-0.0097	-0.0106	-0.0009	-0.0010	0.0001
Total	0.0103	0.0147	0.0044	-0.0008	0.0053

Notes: See Table 4 for further definitions applied to select the "common coded" sample in the CPS dataset. The District of Columbia is included in "50 states." See text for the methodology used to create these statistics. Statistics may not add exactly to column totals because of rounding.

18 vs 50 state difference: Information and Retail Trade. The difference in contribution of these two sectors (0.0047 = 0.0028 + 0.0019) exceeds the 50 versus 18 state difference in the total contribution (0.0044 = 0.0147 - 0.0103).

We take this decomposition one step further and ask whether 50 state versus 18 state differences in employment shares or earnings differentials in the retail trade and information sectors are driving the difference in the contributions to between-industry variance growth. We do this by noting that for industry k, $\Delta[(N_k/N)(\bar{w}_k - \bar{w})^2] = \overline{(\bar{w}_k - \bar{w})^2}\Delta(N_k/N) + \overline{(N_k/N)}\Delta(\bar{w}_k - \bar{w})^2$. The first term on the right hand side of this equation is the contribution of changing employment shares, and the second term is the contribution of changing earnings differentials. The calculations are in Table 8. We find that differences in earnings differentials account for most if not all of the different contributions to between-industry variance growth. In retail trade, a large industry in terms of employment share,

Table 9: Comparisons of the Information and Retail Trade sectors, in 18 vs. 50 states

Data	CPS (micro.)		CPS (agg.)		QCEW (agg.)	
	50 states	18 states	50 states	18 states	50 states	18 states
Contribution to variance growth from 1996-2002 to 2012-18:						
Information	0.0054	0.0082	0.0050	0.0076	0.0030	0.0038
Retail Trade	0.0035	0.0054	0.0024	0.0041	0.0066	0.0078
Ratio of 50 state to 18 state:						
Information	65.9%		65.8%		78.7%	
Retail Trade	64.8%		59.3%		84.8%	

Notes: "CPS (micro.)" are tabulations from our "common-coded" CPS microdata as defined in Table 4; "CPS (agg.)" uses the same data but first aggregates real earnings and employment to state by 18 sector level before computing variance decompositions; "QCEW (agg.)" aggregates published QCEW earnings and employment to state by 18 sector level before computing variance decompositions. The contribution to variance growth is the change across intervals. The difference between CPS (micro.) and (agg.) is due to the Micro starting with log wages at the person level and aggregating while the (agg.) starts with wages at the state by sector level and then takes logs. The QCEW (agg.) starts with wages at the state by sector level and then takes logs.

earnings differentials are declining more in the 18 states than in the 50 states (-0.2313 to -0.3094 in the 18 states, -0.2406 to -0.2926 in the 50 states). In the information industry, earnings differentials are rising faster in the 18 states than in the 50 states (0.4083 to 0.5789 in the 18 states, 0.3951 to 0.5362 in the 50 states).

A question then is whether these large differences in the contribution of these two sectors is idiosyncratic to the CPS or holds more broadly. To investigate this question, we turn to the QCEW at the state by sectoral level (using the same definitions of sectors as in the CPS). An advantage of the QCEW is that it is from comprehensive administrative data covering all 50 states. It is notable that the QCEW public domain data have the same underlying source data as the LEHD data.

Tabulations of between-industry earnings differentials from the public domain QCEW at the state by sectoral level and from the micro CPS are not directly comparable given the CPS differentials reflect employment-weighted means of logs while the QCEW reflects the log of the employment-weighted means. To facilitate an apples-to-apples comparison with the QCEW, we aggregate the levels of the micro common-coded CPS to the state by sectoral level. Results from this exercise are reported in Table 9 focusing on these two key sectors. We refer to the state by sector level data for the CPS and QCEW as "aggregate" (taking logs after aggregation, and abbreviated agg.) in this table.

For the CPS, the contribution to between-industry from the micro data vs. the aggregate for these

two sectors is similar. For Retail Trade, the ratio of the 50 to 18 state contribution is about 65% for the micro data and 59% for the aggregate CPS. For Information, the analogous two ratios are 66%. In contrast, the QCEW yields much less of a difference in the between-industry contribution for these two sectors with ratios of 85% for Retail Trade and 79% for Information in comparing the 50 to 18 state contributions. The inference we draw from this exercise is that the CPS idiosyncratically has a low contribution of Retail Trade and Information for the 50 states vs 18 states. We don't know why the CPS is an outlier relative to the administrative data for these two sectors. We think the measurement issues with CPS industries discussed above is a likely explanation. Our findings imply that there are substantial gains in inference integrating administrative data based industry codes into the CPS.

7 Concluding remarks

Research into rising dispersion of earnings has proceeded along two mostly independent paths. Most of the literature uses household survey data. The messages from that line of literature are well-known. There is an important role of rising dispersion across observable person characteristics including age, education, and occupation. Age by education effects are relatively more important but occupation plays an important supporting role. Changing industry differentials on the margin (that is after controlling for person characteristics) play little if any role.

Longitudinal matched employer-employee data has enabled an alternative look at the determinants of rising earnings inequality. Most of the rise in earnings inequality is accounted for by rising between-firm dispersion. Moreover, most of the rising between-firm dispersion is accounted for by between-industry dispersion.

We have used a novel integrated survey and administrative data to help reconcile these two quite different perspectives. An important part of the reconciliation is methodological. The substantial role of industry from the matched employer-employee data stems from using a full variance decomposition rather than focusing on the marginal effects of industry after controlling for other factors. When using the CPS, a full variance decomposition shows that a substantial fraction of rising CPS earnings dispersion is accounted for by rising between-industry dispersion.

However, the CPS has a significant limitation in terms of the quality and detail of industry codes. Using our integrated CPS-LEHD data, we show that if one uses the CPS for all variables except for industry from the administrative data, we find that overall between-industry variation accounts

for about 65% of rising dispersion whether using household survey (CPS) earnings or administrative (LEHD) earnings.

We show that between-industry dispersion can be decomposed into industry premia, sorting, and segregation using a human capital earnings equation. We find that most of the increase is due to sorting and segregation. Bringing this back to the main conclusions from the studies using household survey data, most of the contribution of observable person characteristics that have been emphasized in the literature reflects increased sorting and segregation of these characteristics between industries.

There are some quantitative differences in our findings depending on whether CPS earnings or administrative data earnings are used. Given issues in the tails with CPS earnings, the CPS has substantially less variance of earnings in the cross section and in terms of growth over time. These differences don't impact the inference that between-industry effects dominate rising earnings inequality, but they raise caution about using CPS earnings in isolation for quantifying changes in the earnings distribution.

References

[1] Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. High Wage Workers and High Wage Firms. *Econometrica* 67 (2): 251-333.

[2] Abraham, Katharine G., John Haltiwanger, Kristin Sandusky, and James R. Spletzer. 2013. Exploring Differences in Household vs. Establishment Measures of Employment. *Journal of Labor Economics*, 31 (2, Pt. 2): S129-S172.

[3] Acemoglu, Daron and David H. Autor. 2011. Skills, Tasks and Technologies: Implications for Employment and Earnings. In *Handbook of Labor Economics*, Volume 4, ed. Orley Ashenfelter and David Card, Amsterdam: Elsevier-North Holland, 1043-1171.

[4] Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. 2019. Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch. *Journal of Political Economy*, 127 (5): 2143-2185.

[5] Card, David, Jesse Rothstein, and Moises Yi. 2022. Industry Wage Differentials: A Firm-Based Approach. Unpublished draft, Department of Economics, University of California, Berkeley.

[6] Dey, Matthew, Susan Houseman, and Anne Polivka. 2010. What Do We Know About Contracting Out in the United States? Evidence from Household and Establishment Surveys in *Labor in the New Economy*, Katharine G. Abraham, James R. Spletzer, and Michael Harper, eds., Chicago: University of Chicago Press, pp. 267-304.

[7] Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, and Michael Westberry. 2021. Integrated Public Use Microdata Series, Current Population Survey: Version 9.0 [dataset]. Minneapolis, MN: IPUMS, `https://doi.org/10.18128/D030.V9.0`

[8] Haltiwanger, John, Henry R. Hyatt, and James R. Spletzer. 2022. Industries, Mega Firms, and Increasing Inequality. Unpublished draft, Department of Economics, University of Maryland.

[9] Hoffmann, Florian, David S. Lee, and Thomas Lemieux. 2020. Growing Income Inequality in the United States and Other Advanced Economies. *Journal of Economic Perspectives* 34 (4): 52-78.

[10] Mellow, Wesley, and Hal Sider. 1983. Accuracy of Response in Labor Market Surveys: Evidence and Implications. *Journal of Labor Economics* 1 (4): 331-344.

[11] Pollard, Emily. 2019. New Approach to Industry and Occupation Recoding in the CPS. Federal Reserve Bank of Kansas City Technical Briefing No. 19-02.

[12] Song, Jae, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till von Wachter. 2019. Firming Up Inequality. *Quarterly Journal of Economics* 134 (1): 1-50.

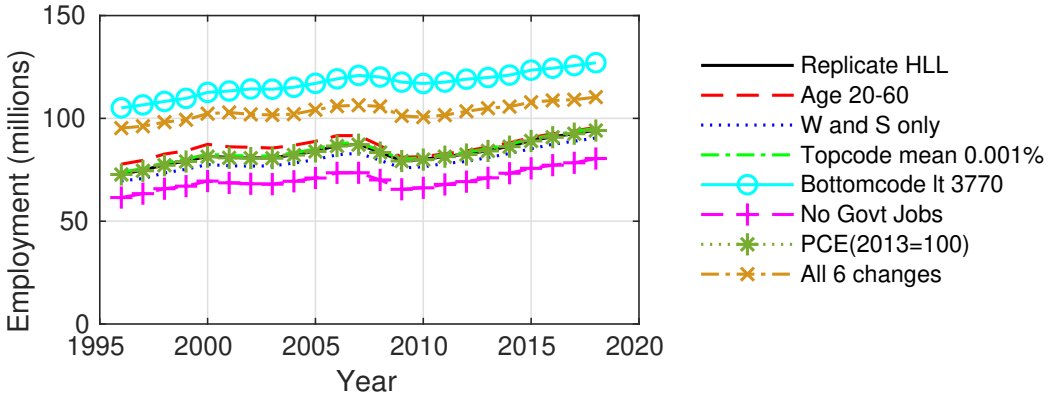
[13] Stansbury, Anna, and Lawrence H. Summers. 2020. The Declining Worker Power Hypothesis: An Explanation for the Recent Evolution of the American Economy. *Brookings Papers on Economic Activity*, Spring, 1-77.

Appendices

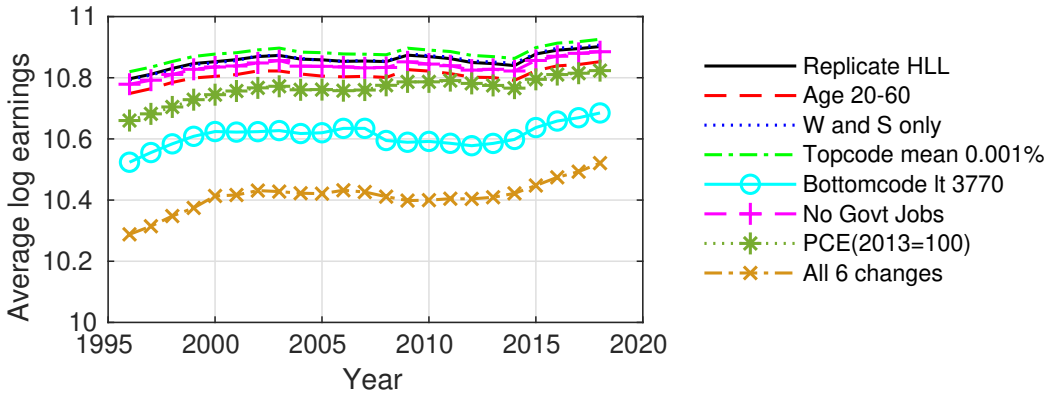
A Supplementary tables and figures

Figure A1: Creating the common-coded CPS-ASEC

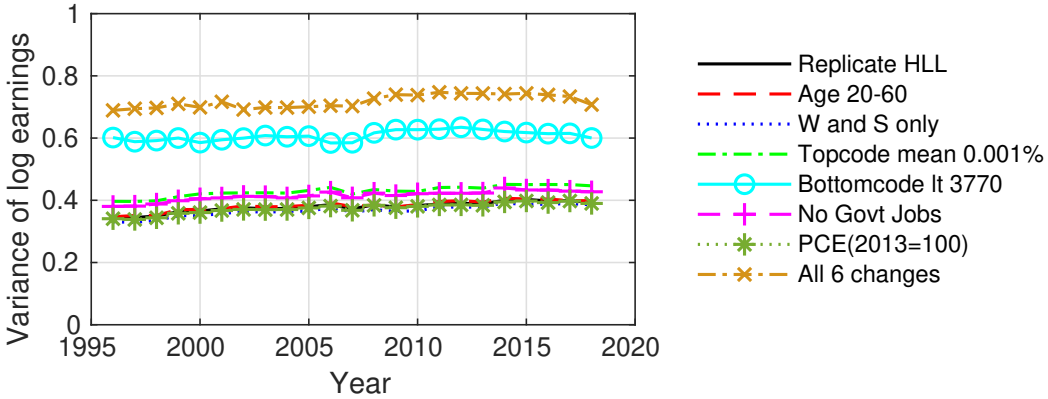
(a) Employment



(b) Average log earnings



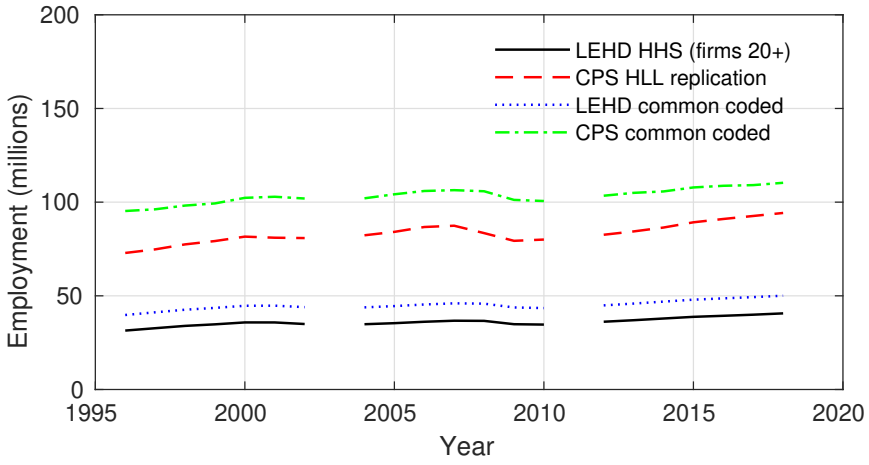
(c) Variance of log earnings



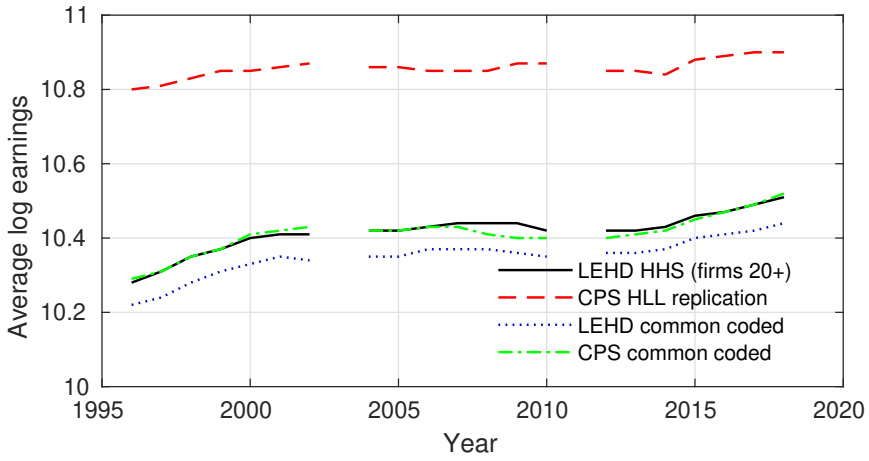
Notes: Authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. Pooled males and females. See Table 4 for the criteria applied to select a "common coded" sample.

Figure A2: Mean, variance, and employment of HLL CPS-ASEC, HHS LEHD, common-coded CPS-ASEC, and common-coded LEHD earnings, by year

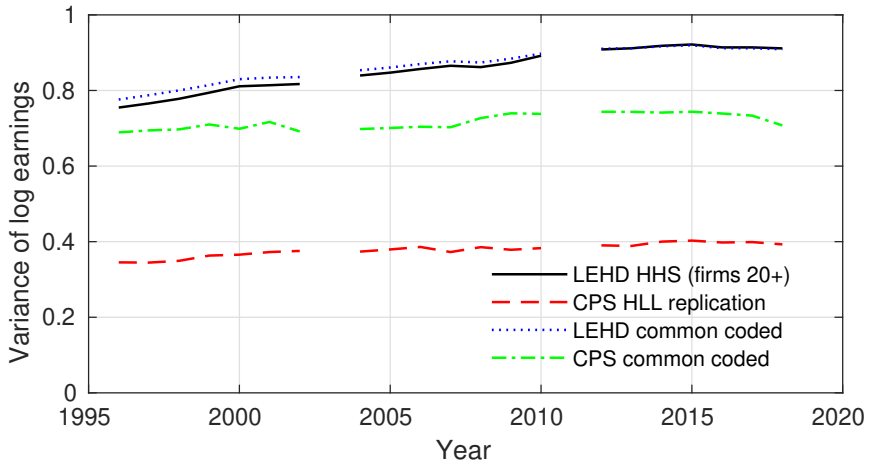
(a) Employment



(b) Average log earnings



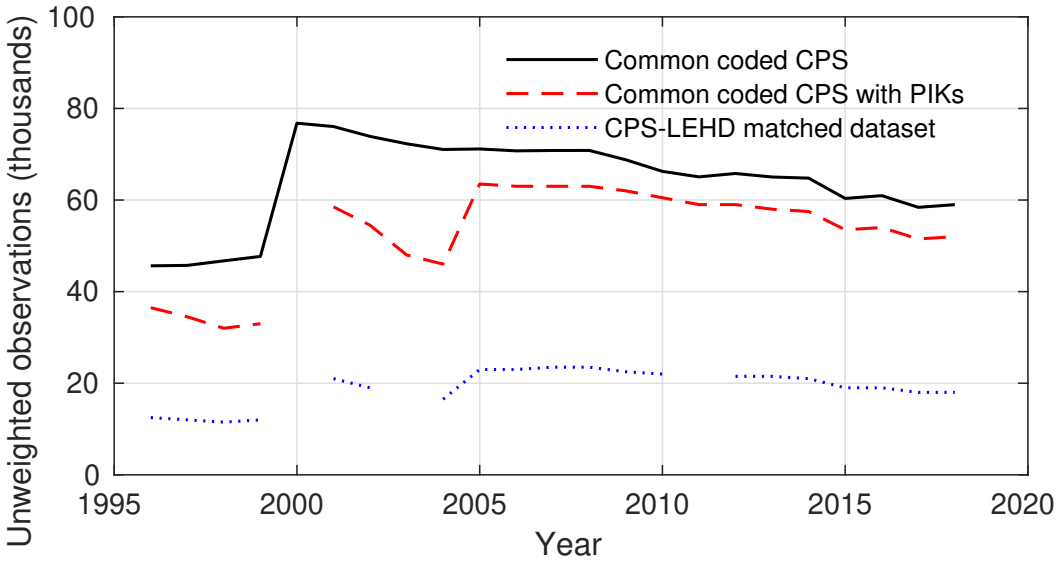
(c) Variance of log earnings



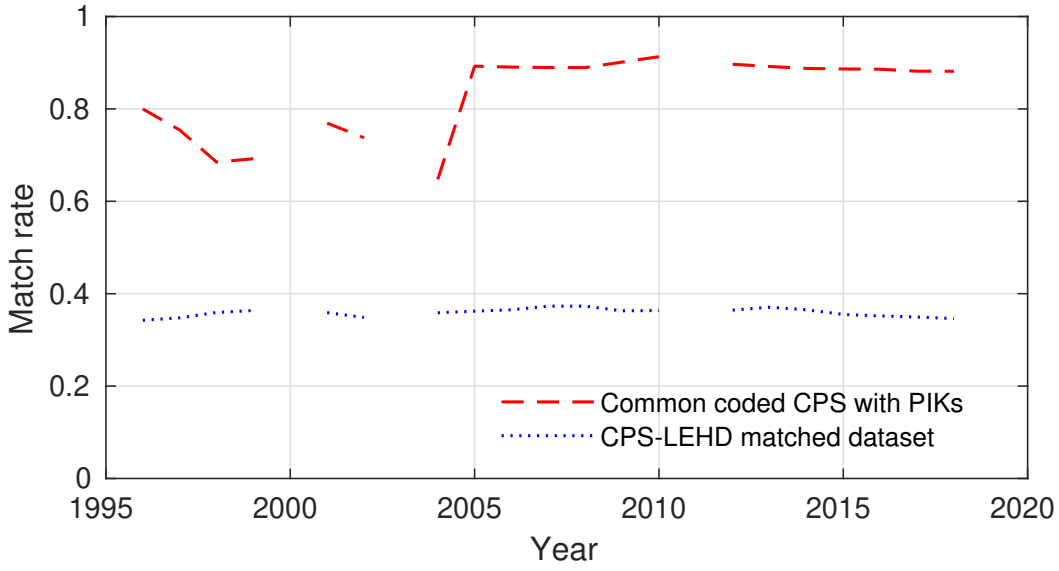
Notes: "CPS HLL replication" shows authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. "LEHD HHS (firms 20+)" shows authors' tabulations of LEHD administrative records as utilized in HHS, which considers only people who work at firms that employ at least twenty people. "Common coding" applies a consistent set of sample selection criteria to the HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website and the HHS LEHD administrative records dataset. See Table 4 for further definitions applied to select the sample in each dataset.

Figure A3: Linking common-coded CPS-ASEC and common-coded LEHD data

(a) Observations

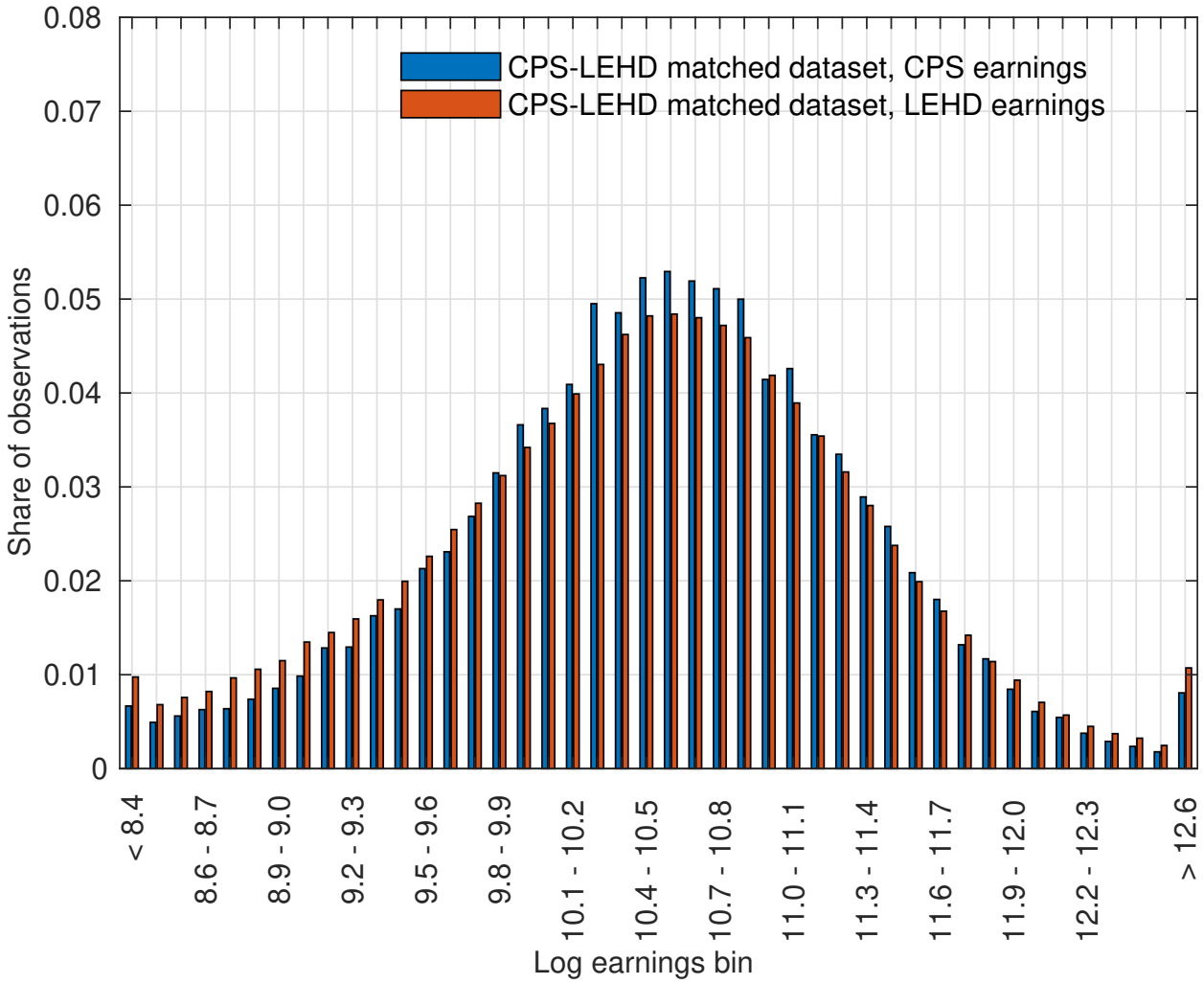


(b) Match rates



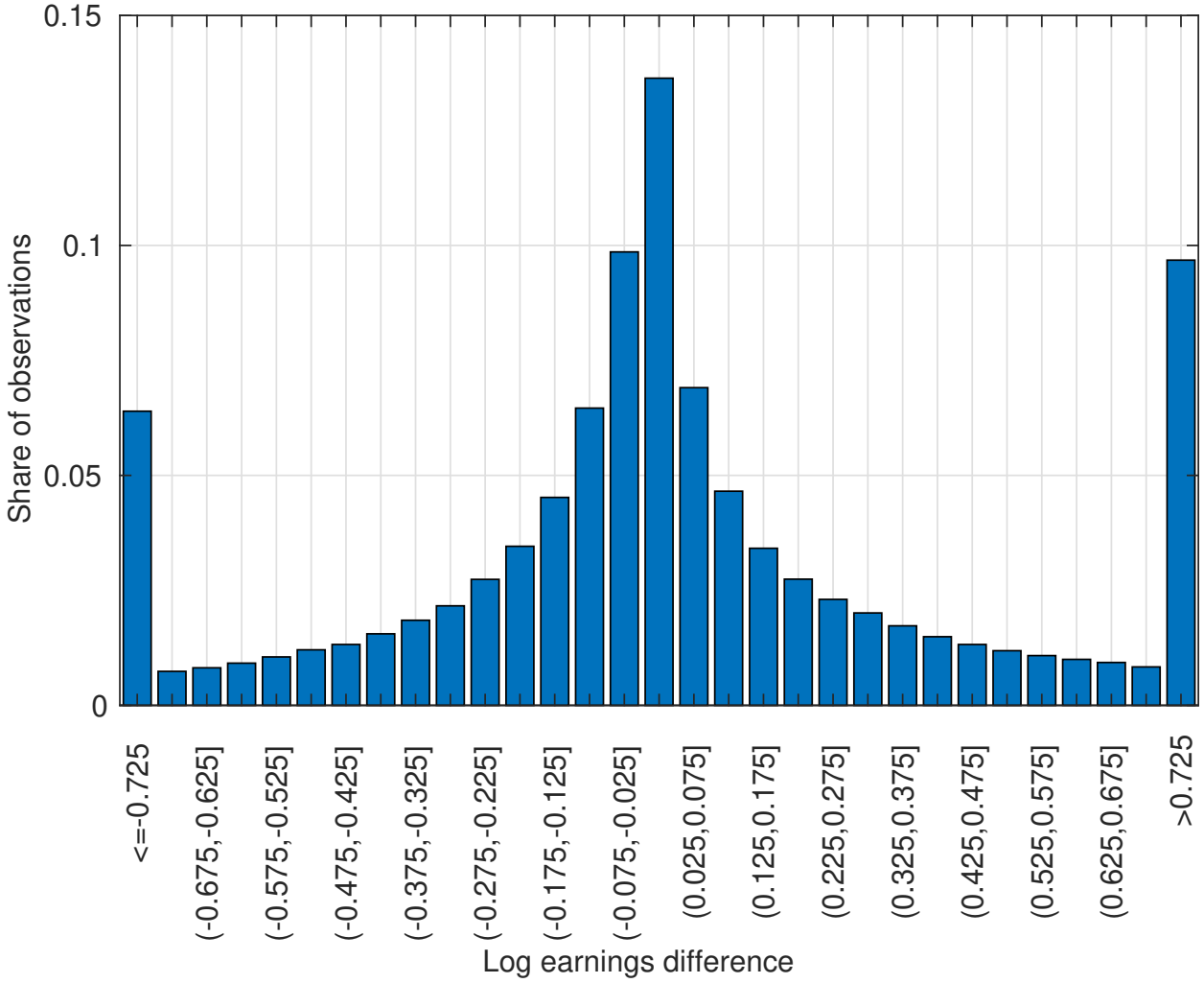
Notes: See Table 4 for the criteria applied to select a "common coded" sample. Further details about the construction of the CPS-LEHD matched dataset are provided in Section 4.1.

Figure A4: Earnings distributions of linked CPS-ASEC and LEHD



Notes: Results pool all years. See Table 4 for the criteria applied to select a "common coded" sample. Further details about the construction of the CPS-LEHD matched dataset are provided in Section 4.1.

Figure A5: PDF of earnings difference, linked CPS-ASEC and LEHD data



Notes: See Table 4 for the criteria applied to select a "common coded" sample. Further details about the construction of the CPS-LEHD matched dataset are provided in Section 4.1.

Table A1: Estimation of the human capital earnings equation using CPS ASEC data and 5-year intervals

	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014	2015-2018	Growth 1975-79 to 2015-18
Earnings variance	0.283	0.292	0.310	0.332	0.349	0.372	0.380	0.390	0.398	0.115
Contributions to total variance, in levels:										
Age and education	0.045	0.049	0.060	0.071	0.079	0.088	0.092	0.094	0.093	0.048
Occupation	0.023	0.022	0.023	0.021	0.023	0.024	0.026	0.026	0.027	0.004
Industry	0.017	0.017	0.016	0.015	0.014	0.012	0.011	0.011	0.011	-0.006
Residual	0.198	0.205	0.220	0.225	0.234	0.249	0.251	0.259	0.267	0.069
Contributions to total variance, in percentages:										
Age and education	15.9%	16.8%	18.8%	21.5%	22.6%	23.5%	24.2%	24.1%	23.5%	41.9%
Occupation	8.3%	7.4%	7.1%	6.3%	6.6%	6.4%	6.9%	6.8%	6.8%	3.3%
Industry	5.9%	5.7%	5.1%	4.4%	3.9%	3.1%	2.8%	2.8%	2.7%	-5.3%
Residual	69.9%	70.1%	69.0%	67.8%	66.8%	66.9%	66.0%	66.4%	67.0%	60.1%

Notes: We downloaded the HLL CPS-ASEC data from the *Journal of Economic Perspectives* website. Our earnings variable is the natural log of real annual labor earnings. Our regression specification is based on HLL Figure 4, except we use labor earnings instead of total income, and we pool male and females. "Age and education" is the fraction of the variance of labor earnings explained by equation (2). "Occupation" is the marginal contribution of including occupation, obtained by subtracting the percentage of the variance explained by equation (2) from that of equation (3). "Industry" is the marginal contribution of industry, obtained by subtracting the percentage of variance explained by equation (3) from that of equation (4). Industry is defined using 12 SIC categories. "Residual" is the fraction of the variance that is unexplained when estimating equation (4).

Table A2: First and second panels replicate HLL Figure 4, third panel is HLL with pooled genders, fourth panel is pooled genders with labor earnings rather than total income (Table A1 of this paper)

	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014	2015-2018	Growth 1975-79 to 2015-18
HLL Figure 4, males										
Total income variance	0.262	0.286	0.329	0.358	0.389	0.416	0.427	0.444	0.457	0.195
Age and education	0.052	0.061	0.076	0.091	0.102	0.112	0.119	0.124	0.124	0.072
Occupation	0.009	0.010	0.013	0.016	0.021	0.023	0.025	0.026	0.028	0.019
Industry	0.009	0.010	0.011	0.011	0.010	0.009	0.008	0.008	0.008	-0.001
Residual	0.193	0.206	0.229	0.240	0.256	0.272	0.274	0.286	0.297	0.105
HLL Figure 4, females										
Total income variance	0.190	0.208	0.250	0.273	0.301	0.311	0.333	0.351	0.374	0.185
Age and education	0.038	0.041	0.054	0.066	0.077	0.081	0.089	0.096	0.102	0.065
Occupation	0.012	0.013	0.017	0.019	0.019	0.020	0.023	0.023	0.025	0.013
Industry	0.007	0.008	0.010	0.010	0.009	0.007	0.006	0.006	0.006	-0.001
Residual	0.133	0.147	0.170	0.179	0.195	0.204	0.215	0.226	0.241	0.108
HLL Figure 4, pooled										
Total income variance	0.298	0.305	0.337	0.352	0.380	0.396	0.407	0.421	0.436	0.138
Age and education	0.050	0.057	0.069	0.081	0.091	0.100	0.102	0.105	0.106	0.056
Occupation	0.025	0.022	0.023	0.022	0.024	0.025	0.028	0.028	0.029	0.005
Industry (SIC 12)	0.016	0.016	0.016	0.015	0.014	0.012	0.011	0.011	0.011	-0.005
Residual	0.207	0.211	0.230	0.235	0.250	0.262	0.266	0.277	0.289	0.083
Table A1 this paper, pooled										
Labor earnings variance	0.283	0.292	0.318	0.332	0.349	0.372	0.380	0.390	0.398	0.115
Age and education	0.045	0.049	0.060	0.071	0.079	0.088	0.092	0.094	0.093	0.048
Occupation	0.023	0.022	0.023	0.021	0.023	0.024	0.026	0.026	0.027	0.004
Industry (SIC 12)	0.017	0.017	0.016	0.015	0.014	0.011	0.011	0.011	0.011	-0.006
Residual	0.198	0.205	0.220	0.225	0.234	0.249	0.251	0.259	0.267	0.069

Notes: The top two panels replicate columns {B, E, H, P, W, X} of HLL's figure_4.xlsx downloaded from the *Journal of Economic Perspectives* website.

Table A3: Variance decomposition of the human capital earnings equation, CPS-ASEC data

	1975-1981	1982-1988	1989-1995	1996-2002	2004-2010	2012-2018	Growth 1975-81 to 2012-18	Growth 1996-02 to 2012-18
Earnings variance	0.283	0.310	0.333	0.360	0.380	0.397	0.113	0.037
Using 12 SIC industries								
Within-industry:	95.2%	95.2%	95.4%	95.8%	95.3%	94.9%	94.0%	86.2%
Age, educ., & occ.	24.9%	25.7%	27.5%	28.8%	29.3%	28.0%	35.8%	20.5%
Residual	70.3%	69.5%	67.8%	67.0%	66.0%	66.9%	58.2%	65.7%
Between-industry:	4.8%	4.8%	4.6%	4.2%	4.7%	5.1%	6.0%	13.8%
Segregation	2.2%	2.4%	2.8%	3.0%	3.5%	4.0%	8.5%	13.3%
Pay premia	7.9%	6.7%	5.5%	4.5%	3.5%	3.4%	-7.8%	-6.9%
Sorting	-5.3%	-4.3%	-3.7%	-3.3%	-2.4%	-2.3%	5.2%	7.4%
Using 18 NAICS industries								
Within-industry:	93.2%	93.3%	93.8%	94.1%	93.6%	92.5%	90.8%	76.9%
Age, educ., & occ.	24.7%	25.9%	27.6%	28.5%	29.1%	27.5%	34.6%	18.2%
Residual	68.5%	67.5%	66.2%	65.6%	64.5%	65.0%	56.2%	58.8%
Between-industry:	6.8%	6.7%	6.2%	5.9%	6.4%	7.5%	9.2%	23.1%
Segregation	2.9%	2.9%	3.2%	3.3%	3.8%	4.4%	8.1%	14.8%
Pay premia	10.2%	9.0%	7.3%	6.0%	5.4%	5.5%	-6.2%	1.0%
Sorting	-6.3%	-5.2%	-4.3%	-3.4%	-2.8%	-2.4%	7.3%	7.3%

Notes: Authors' tabulations of HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. Pooled males and females. The year 2000 is deleted. Earnings is natural log of real annual labor earnings. The 12 SIC aggregate industries are defined following the Standard Industrial Classification system. The 18 NAICS aggregate industries are defined following the North American Industrial Classification System. Coding of CPS industry data (indly) into NAICS industries follows Table C-5 of Pollard (2019). See equation (10) for definitions.

Table A4: Industry sector agreement rate in the CPS-LEHD linked data

CPS NAICS	CPS NAICS industry distribution	% LEHD with same industry
Agriculture	1.0%	58.8%
Mining	0.9%	51.5%
Construction	6.9%	70.2%
Manufacturing	15.7%	66.7%
Wholesale Trade	4.1%	**46.9%**
Retail Trade	13.3%	71.3%
Transportation & Warehousing	4.2%	64.4%
Utilities	0.9%	53.6%
Information	6.1%	**46.6%**
Finance & Insurance	6.5%	*77.0%*
Real Estate & Rental	1.9%	50.9%
Professional &+ Business Services	10.3%	67.1%
Educational Services	2.8%	**44.2%**
HealthCare & Social Assistance	12.8%	*83.4%*
Arts, Entertainment, & Recreation	1.5%	51.5%
Accommodation & Food Services	6.7%	*77.9%*
Other Services	3.2%	**48.0%**
Unknown (2nd job)	1.2%	**0.0%**
Total	100%	66.1%

Notes: See Table 4 for the criteria applied to select a "common coded" sample. Further details about the construction of the CPS-LEHD matched dataset are provided in Section 4.1. *Italics* shows agreement >75%. **Bold** shows agreement <50%.

Table A5: Variance decomposition of the human capital earnings equation

	(1)	(2)	(3)	(4)	(5)	(6)
Data	CPS	CPS	Linked CPS-LEHD	Linked CPS-LEHD	Linked CPS-LEHD	Linked CPS-LEHD
Sample	HLL JEP	Common coded	Common coded	Common coded	Common coded	Common coded
Earnings measure	CPS	CPS	CPS	CPS	CPS	LEHD
Industry measure	CPS 18	CPS 18	CPS 18	LEHD 18	LEHD 299	LEHD 299
Variance level 1996-2002						
Earnings variance	0.360	0.703	0.667	0.667	0.667	0.746
Within-industry:	94.1%	86.9%	87.6%	91.7%	82.5%	78.3%
Age, education & occupation	28.5%	24.7%	27.8%	28.1%	22.4%	18.8%
Age and education	13.0%	12.3%	14.0%	14.3%	11.3%	10.3%
Occupation	7.2%	6.1%	6.5%	6.6%	5.6%	4.3%
Covariance: age+educ. & occ.	8.2%	6.3%	7.4%	7.2%	5.5%	4.2%
Residual	65.6%	62.2%	59.8%	63.6%	60.1%	59.5%
Between-industry:	5.9%	13.1%	12.4%	8.3%	17.5%	21.7%
Segregation:	3.3%	2.2%	2.2%	2.1%	4.2%	3.1%
Age and education	1.8%	1.0%	1.1%	1.1%	1.8%	1.5%
Occupation	0.5%	0.4%	0.4%	0.3%	0.8%	0.5%
Covariance: age+educ. & occ.	1.1%	0.8%	0.8%	0.7%	1.6%	1.1%
Pay Premia	6.0%	11.0%	8.6%	4.3%	8.1%	11.2%
Sorting:	-3.4%	-0.1%	1.5%	1.8%	5.2%	7.4%
Covariance: age+educ. & ind.	-2.2%	-0.2%	0.9%	0.9%	3.3%	5.4%
Covariance: industry & occ.	-1.2%	0.1%	0.6%	0.9%	1.9%	1.9%

Notes: The rows titled "Data" and "Sample" indicate the data used for the variance decomposition (see text for description). "Earnings measure" indicates whether CPS or LEHD earnings is used in the decomposition. "Industry measure:" "CPS 18" refers to 18 NAICS sectors from the CPS-ASEC (recoding CPS-ASEC variable indly following Table C-5 of Pollard (2019)), "LEHD 18" refers to NAICS sectors from the LEHD, and "LEHD 299" refers to 299 4-digit NAICS industries from the LEHD. See equation (10) for definitions.

Table A6: Variance decomposition of the human capital earnings equation (continued)

	(1)	(2)	(3)	(4)	(5)	(6)
Data	CPS	CPS	Linked CPS-LEHD	Linked CPS-LEHD	Linked CPS-LEHD	Linked CPS-LEHD
Sample	HLL JEP	Common coded	Common coded	Common coded	Common coded	Common coded
Earnings measure	CPS	CPS	CPS	CPS	CPS	LEHD
Industry measure	CPS 18	CPS 18	CPS 18	LEHD 18	LEHD 299	LEHD 299
Variance level 2012-2018						
Earnings variance	0.397	0.738	0.717	0.717	0.717	0.845
Within-industry:	92.5%	86.1%	85.3%	88.6%	79.1%	73.1%
Age, education, & occupation:	27.5%	24.9%	27.2%	27.8%	21.4%	18.2%
Age and education	12.9%	12.9%	14.3%	14.5%	11.3%	10.9%
Occupation	6.9%	5.7%	5.9%	6.3%	5.1%	3.7%
Covariance: age+educ. & occ.	7.7%	6.3%	6.9%	7.0%	4.9%	3.6%
Residual	65.0%	61.2%	58.1%	60.8%	57.8%	54.9%
Between-industry:	7.5%	13.9%	14.7%	11.4%	20.9%	26.9%
Segregation:	4.4%	3.9%	4.4%	4.1%	6.4%	4.6%
Age and education	2.0%	1.5%	1.7%	1.8%	2.6%	2.0%
Occupation	0.7%	0.7%	0.7%	0.6%	1.0%	0.7%
Covariance: age+educ. & occ.	1.8%	1.7%	2.0%	1.7%	2.7%	1.8%
Pay premia	5.5%	8.3%	6.9%	3.7%	7.3%	13.2%
Sorting:	-2.4%	1.7%	3.4%	3.5%	7.2%	9.1%
Covariance: age+educ. & ind.	-1.8%	0.8%	2.0%	1.9%	4.5%	6.2%
Covariance: industry & occ.	-0.6%	1.0%	1.5%	1.6%	2.7%	2.9%

Notes: The rows titled "Data" and "Sample" indicate the data used for the variance decomposition (see text for description). "Earnings measure" indicates whether CPS or LEHD earnings is used in the decomposition. "Industry measure:" "CPS 18" refers to 18 NAICS sectors from the CPS-ASEC (recoding CPS-ASEC variable indly following Table C-5 of Pollard (2019)), "LEHD 18" refers to NAICS sectors from the LEHD, and "LEHD 299" refers to 299 4-digit NAICS industries from the LEHD. See equation (10) for definitions.

Table A7: Variance decomposition of the human capital earnings equation

	(1)	(2)	(3)	(4)	(5)	(6)
Data	CPS	CPS	Linked CPS-LEHD	Linked CPS-LEHD	Linked CPS-LEHD	Linked CPS-LEHD
Sample	HLL JEP	Common coded	Common coded	Common coded	Common coded	Common coded
Earnings measure	CPS	CPS	CPS	CPS	CPS	LEHD
Industry measure	CPS 18	CPS 18	CPS 18	LEHD 18	LEHD 299	LEHD 299
Change from 1996-02 to 2012-18						
Variance growth	0.037	0.035	0.050	0.050	0.050	0.100
Within-industry:	76.9%	70.7%	54.0%	47.8%	34.5%	33.8%
Age, education & occupation:	18.2%	30.0%	17.9%	23.7%	7.8%	13.6%
Age and education	11.8%	25.4%	18.3%	17.7%	12.2%	14.9%
Occupation	4.1%	1.3%	-1.4%	1.4%	-1.8%	-0.4%
Covariance: age+educ. & occ.	2.2%	6.9%	1.0%	4.6%	-2.6%	-0.9%
Residual	58.8%	40.7%	36.1%	24.1%	26.7%	20.3%
Between-industry:	23.1%	29.3%	46.0%	52.2%	65.5%	66.2%
Segregation	14.8%	36.9%	33.7%	30.3%	34.9%	15.3%
Age and education	3.8%	11.7%	11.0%	11.2%	12.7%	6.2%
Occupation	2.5%	5.7%	5.4%	4.0%	4.6%	1.9%
Covariance: age+educ. & occ:	8.4%	19.4%	17.3%	15.1%	17.5%	7.1%
Pay premia	1.0%	-47.0%	-16.1%	-2.6%	-3.4%	21.4%
Sorting	7.3%	39.4%	28.5%	24.5%	34.1%	29.5%
Covariance: age+educ. & ind.	2.0%	20.6%	15.7%	14.1%	20.5%	19.4%
Covariance: industry & occ.	5.3%	18.9%	12.7%	10.4%	13.5%	10.1%

Notes: The rows titled "Data" and "Sample" indicate the data used for the variance decomposition (see text for description). "Earnings measure" indicates whether CPS or LEHD earnings is used in the decomposition. "Industry measure:" "CPS 18" refers to 18 NAICS sectors from the CPS-ASEC (recoding CPS-ASEC variable indly following Table C-5 of Pollard (2019)), "LEHD 18" refers to NAICS sectors from the LEHD, and "LEHD 299" refers to 299 4-digit NAICS industries from the LEHD. See equation (10) for definitions.