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

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

July 6, 2022

*Preliminary draft prepared for the 2022 NBER Summer Institute.*

## Abstract

Recent analyses of survey data highlight rising dispersion in earnings by observable person characteristics such as education and occupation as critical factors for rising earnings inequality. Also, most of the increase is accounted for by within cell unobservable factors. Industry plays if anything a dampening role. In contrast, administrative records employer-employee matched data permit a more comprehensive quantification of person, firm, and industry effects. Recent research highlights that rising between firm earnings dispersion inequality and in turn between-industry earnings dispersion dominates the rise in earnings inequality. Increasing sorting of high (low) person effect workers to industries with high (low) premia and segregation of high and low person effect workers into different industries account for most of this rise in between-industry dispersion. To help reconcile these contrasting messages, we construct a novel integrated dataset based upon CPS microdata linked with LEHD administrative records data. Using the integrated data, we find that most of the rise in earnings inequality is accounted for by rising between-industry inequality whether using CPS or LEHD earnings. Part of this finding depends on using a variance decomposition approach that quantifies sorting and segregation contributions of observable person characteristics across industries. Part of this finding also depends on using high quality detailed industry codes from the administrative data. This finding mostly reflects a substantial contribution of increased sorting and segregation of observable person characteristics (including education and occupation) between industries for rising earnings inequality.

JEL Codes: J31, J21

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

[*] 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, Thomas Lemieux, Kristin Sandusky, and participants in the U.S. Census Bureau research lunch and the 2022 Society of Labor Economists conference for helpful comments and suggestions. All results have been reviewed to ensure that no confidential information is disclosed (CBDRB-FY22-159). 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

[§] Center for Economic Studies, U.S. Census Bureau. 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. Matched employer-employee data permits analyzing inequality through the lens of the empirical framework of Abowd, Kramarz, and Margolis (1999), hereafter abbreviated as AKM. Song et al. (2019) demonstrate that increasing earnings inequality in the U.S. is attributable to rising between firm dispersion, as highly paid workers increasingly work for high-paying firms (i.e., sorting) and with each other (i.e., segregation). 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 both increased sorting and segregation at the industry level, along with widening industry-level pay premia, account 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. The role of industry and occupation have also been analyzed (for recent studies see Acemoglu and Autor, 2011, and Hoffman, Lee, and Lemieux, 2020, hereafter HLL). These recent studies find a supporting role for rising between occupation differences in earnings while 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 person effects, occupations, firms 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 including employer characteristics such as firm size and/or industry in the analysis). 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, HHS state, "Most of the rise in overall earnings inequality is accounted for by rising between-industry inequality," whereas HLL state "The between-group variance component linked to industry has been declining over time," and Stansbury and Summers (2020) write "Using the CPS, we show that since the 1980s there has been a decline of about one third in the dispersion of industry wage premia."

To assess these contrasting messages, we construct a novel source of CPS microdata linked with LEHD administrative records data. We assess the role of several competing explanations for why survey versus administrative records seem to tell such different stories regarding how earnings inequality has increased. These range from relatively straightforward measurement issues to fundamental differences in characterizing how changes in workforce composition drive changes in inequality.

There are several measurement issues that are straightforward to document and evaluate. The first explanation that we explore is quite simple: studies that rely on the CPS exclude many more low earners than those that rely on administrative records data. This sample selection criterion may matter, as inequality measures are sensitive to employment among the lowest-paid jobs. Industry measurement may also matter. Studies that rely on the CPS often aggregate industry-level information to a small number of industry sectors based on the Standard Industrial Classification (SIC). Does using detailed industry data matter, and does replacing the SIC with North American Industry Classification System (NAICS) measures of industry matter? Furthermore, does using an individual's industry from the administrative data rather than the industry collected in the CPS matter in accounting for increasing earnings inequality?

Apart from these measurement issues, there is a sharp difference between the way these strands of the inequality literature consider the role of employers to increasing inequality. 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 how workers are segregated or sorted into industries. Studies that rely on administrative records follow the empirical framework of AKM when considering the worker- and firm-level determinants of inequality. In particular, the AKM framework is a natural starting point for considering the roles of sorting and segregation in addition to pay premia in the evolution of inequality. This distinction matters as changing workforce composition manifests itself in earnings inequality mainly through firm- and industry-level differences: that is, though sorting and segregation. Person effects are also potentially quite different in the AKM framework compared to observable person characteristics such as age, education, and occupation.

To address these questions, we have assembled a new and unique data source. We have linked respondents in the CPS Annual Social and Economic Supplement (ASEC) with administrative records from the U.S. Census Bureau's LEHD data. The LEHD data we use here include the AKM worker and firm effects as estimated by HHS. We believe that our linked CPS-LEHD data is the first U.S. microdata to include both the AKM components of wage determination along with the more traditional CPS based measures of age, education, gender, occupation, and industry.

We estimate that most of the rise in earnings inequality is accounted for by rising between-industry inequality with the CPS data. This required three modifications to the typical analysis of CPS data. First, we change the methodology, moving away from a marginal contribution of industry conditional on demographic characteristics and switching to a full variance decomposition of the human capital earnings equation that accounts for covariances that measure segregation and sorting into industries. Second, after common coding both the CPS and the LEHD (which does not substantially change the between-industry contribution to variance growth in the administrative data), the between-industry contribution to variance growth is larger in the linked CPS-LEHD sample than it is in the common coded CPS. Part of this reflects geographic effects that are idiosyncratic to the CPS. And third, replacing the CPS measure of SIC industry sector with a 4-digit NAICS industry measure increases the between-industry contribution to variance growth, with most of this increase coming from moving from sectors to 4-digit industry. Switching from SIC to NAICS and switching from a CPS industry sector measure to a LEHD industry sector measure both result in small increases to the between-industry contribution to variance growth, but the effect of each of these changes to a higher quality industry measure is smaller than the effect of increasing the amount of industry detail. It is only the administrative data that permits this industry detail.

We also find the effect of observable person characteristics in the human capital earnings equation differ substantially relative to person effects from the AKM earnings equation. Using the human capital earnings equation with CPS observables such as age, education and occupation, within-industry variation in these factors account accounts for 18 to 22 percent of the earnings variance, whereas within firm variation in the person effect accounts for about 55 percent of the earnings variance using the AKM earnings equation (and this increases by another ten percent or so including between firm, within-industry variation). This difference is reversed for the unexplained contribution to earnings variance. Using the human capital earnings equation, 55 to 60 percent of the earnings variance is unaccounted for, whereas this is 10 to 15 percent using the AKM earnings equation. These large differences in cross sectional level variances translate into large differences in the role of person effects (and in turn sorting and segregation effects) in the growth of earnings dispersion.

Even though observable person characteristics don't capture as much variation as AKM person effects, we find similar messages on the role of between-industry dispersion using either observable person characteristics or AKM person effects. When we use administrative data earnings and detailed industry codes, by construction overall industry effects are the same. However, the decomposition of these effects into between-industry sorting, segregation and differ depend on the use of observable person characteristics vs. the AKM person effects. Interestingly, we find that sorting contribution is large (about 30 percent) and almost identical using either approach. Using observable person characteristics yields a smaller contribution of between-industry segregation but a larger role for between-industry premia but either way these two contributions add up to about 35 percent.

The paper proceeds as follows. We provide an overview in section 2 of how inequality has been characterized in administrative records data to identify the roles of sorting, segregation, and pay premia. We review some of the main findings regarding the contribution of industry from Haltiwanger, Hyatt, and Spletzer (2022). Also, in section 2, we review the findings from HLL in more detail as this paper provides an excellent synthesis of the literature using household surveys and independent confirming evidence from the CPS. In section 3, we 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 4, we adjust for differences in the sample composition of how analysts often use the CPS and various administrative records. In section 5, 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 6. Sensitivity analysis is in section 7. Concluding remarks are in section 8.

# 2 Literature review

## 2.1 Analysis of inequality using linked employer-employee data

The recent literature on inequality that uses administrative data takes AKM as its starting point. In the AKM framework, log earnings $y_t^{i,j,k,p}$ of worker $i$ at firm $j$ in industry $k$ at time $t$ in time interval $p$ are a linear function of a worker effect $\theta^{i,p}$, a firm effect $\psi^{j,k,p}$, and time-varying observable characteristics $X_t^{i,p}$ (usually including age and time effects) that have marginal effects according to the vector $\beta^p$.[1] This is expressed as

$$\underbrace{y_t^{i,j,k,p}}_{\substack{\text{log real}\\\text{earnings}}} = \underbrace{X_t^{i,p}\beta^p}_{\substack{\text{time-varying}\\\text{observables}}} + \underbrace{\theta^{i,p}}_{\substack{\text{person}\\\text{effect}}} + \underbrace{\psi^{j,k,p}}_{\substack{\text{firm}\\\text{effect}}} + \underbrace{\varepsilon_t^{i,j,k,p}}_{\text{residual}}. \tag{1}$$

Following Card, Heining, and Kline (2013), this model is estimated separately by time intervals indexed by $p$. Estimation of this model on matched employer-employee data provides a comprehensive description of the extent to which earnings inequality is determined by observable characteristics, worker effects, firm effects, and sorting. Changes in inequality are captured through a variance decomposition. The variance of labor earnings $\text{var}(y_t^{i,j,k,p})$ can be written as follows:

$$\begin{aligned}\underbrace{\text{var}(y_t^{i,j,k,p})}_{\substack{\text{earnings}\\\text{variance}}} = & \underbrace{\text{var}(X_t^{i,p}\beta) + \text{var}(\theta^{i,p}) + 2\text{cov}(X_t^{i,p}\beta, \theta^{i,p})}_{\text{person effects and observables}} + \underbrace{\text{var}(\psi^{j,k,p})}_{\text{pay premia}} + \\ & \underbrace{2\text{cov}(\theta^{i,p}, \psi^{j,k}) + 2\text{cov}(X_t^{i,p}\beta, \psi^{j,k})}_{\text{total sorting}} + \underbrace{\text{var}(\varepsilon_t^{i,j,k,p})}_{\text{AKM residual}}.\end{aligned} \tag{2}$$

The contribution of workers to earnings dispersion is the sum of the variance of the worker effects, that of observable characteristics, and their covariance: $\text{var}(X_t^{i,p}\beta) + \text{var}(\theta^{i,p}) + 2\text{cov}(X_t^{i,p}\beta, \theta^{i,p})$. The contribution of firms is given by the dispersion of the firm pay premia $\text{var}(\psi^{j,k,p})$. Sorting is the extent to which firms with low- versus high-effects employ workers with low- vs. high-effects, as well as low- versus high-observables: $2\text{cov}(\theta^{i,p}, \psi^{j,k,p}) + 2\text{cov}(X_t^{i,p}\beta, \psi^{j,k,p})$.

Card, Heining, and Kline (2013) used a variance decomposition of this form to study the evolution of inequality in West Germany. To do so, they estimated the model separately over a number of time periods (e.g., 1985-1991), and compared the contribution of these components between different periods (e.g., relative to 2002-2009) to understand the contribution of workers, firms, and sorting to changes in inequality. They found that earnings dispersion increased due to increases in the dispersion of worker and firm effects, as well as increases in sorting.

More recently, Song et al. (2019) considered the evolution of earnings dispersion in the U.S. These authors were motivated by the fact that increases in earnings inequality occurred between rather than

[1] We include here the industry $k$ superscript here introduced by HHS. For a formal treatment of the differences between the variance decompositions of Card, Heining, and Kline (2013), Song et al. (2019), and HHS, see Appendix A.

within firms. A key lesson from Song et al. (2019) is that there is a difference between the between-firm component of earnings dispersion estimated from a simple within and between firm variance decomposition and the cumulative contributions of firm pay premia and sorting estimated from the AKM model. Song et al. reconciled this discrepancy by expressing the between-worker component var($\theta^{i,p}$) in terms of that which is due to segregation (that is, the tendency for workers with low vs. high effect to work with each other), denoted by var($\bar{\theta}^{j,k,p}$), relative to worker-driven dispersion, denoted by var($\theta^{i,p} - \bar{\theta}^{j,k,p}$).

In the Song et al. (2019) variance decomposition, firm segregation is defined as var($\bar{\theta}^{j,k,p} + \bar{X}^{j,k,p}\beta^{p}$), and measures differences among firms in terms of the workers that they employ.[2] If all firms employ, on average, workers with similar worker effects and observable characteristics, then the contribution of segregation to earnings inequality will be small. By contrast, if some firms employ workers only with high effects, and other firms employ workers only with small effects, the contribution of segregation to inequality can be substantial. Note that increasing inequality due to segregation does not imply that firms do not play a role in terms of increasing inequality. For example, recent work by Herkenhoff et al. (2018) and Jarosch, Oberfield, and Rossi-Hansberg (2019) indicate that workers learn from their co-workers. Increasing segregation across firms could therefore provide fewer opportunities for low-paid workers to learn new skills that could allow them to increase their pay.

The extension of AKM considered by Song et al. (2019) offers a powerful framework in which to consider the between-firm contribution to increasing inequality. Workers at some firms earn more than others. They find that firm pay premia (var($\psi^{j,k,p}$)) have a small effect, but most of the firm effects on increasing dispersion are due to differences in the workers that firms they employ, through both segregation (var($\bar{\theta}^{j,k,p} + \bar{X}^{j}\beta$)) and sorting (2cov($\bar{\theta}^{j,k,p} + \bar{X}^{j,k,p}\beta^{p}, \psi^{j,k,p}$)). They find a role for both in increasing inequality.

HHS present an extension of the Song et al. (2019) framework to consider differences that are within- versus between-industry. This analysis was motivated by the findings of Haltiwanger and Spletzer (2020), who showed that rising between-firm inequality in the U.S. has been driven by between-industry inequality. The pay associated with jobs in low-paying industries such as restaurants and general merchandise stores has been declining, while that of jobs in high-paying industries such as information services and financial investment activities has been increasing. HHS consider

[2] Note that var($\bar{\theta}^{j,k,p} + \bar{X}^{j,k,p}\beta^{p}$) = var($\bar{\theta}^{j,k,p}$) + var($\bar{X}^{j,k,p}\beta^{p}$) + 2cov($\bar{\theta}^{j,k,p}, \bar{X}^{j,k,p}\beta^{p}$).

the extent to which this increasing between-industry inequality is due to sorting, segregation, and pay premia.

HHS document that over the time period 1996-2002 to 2012-2018, the variance of annual log earnings increased from 0.794 to 0.915. Between-industry differences account for 61.9% of this increase. This 61.9% between-industry effect is accounted for by increased industry sorting (28.0%), increased industry segregation (25.2%), and increased industry-specific pay premia (8.7%). While the changing industry earnings differentials are relatively small (8.7%), the changing composition of the workforce across industries (sorting and segregation) is very important. These findings imply that the increased segregation and sorting accounting for most of the increase in earnings inequality is between industries. Strikingly, HHS find that it is only a relatively small fraction of industries (about 10%) in the tails of the earnings distribution that account for virtually all of the increasing role of between-industry dispersion.

## 2.2 Analysis of inequality using CPS data

It is beyond the scope of this study to provide an in-depth synthesis of the voluminous literature on inequality from household surveys, and in particular from the CPS for the U.S. Fortunately, HLL provide such an in-depth synthesis, so our discussion here draws heavily on the insights from that paper.[3] 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 and synthesize the large literature using the CPS. This independent analysis is an excellent synthesis of the existing literature and serves as the starting point for our own analysis.[4]

HLL's Figure 3 summarizes the contributions of worker characteristics. They find that most of the inequality growth is due to the sum of four variance components: 1) rising within-group dispersion for high-school-educated workers, 2) rising within-group dispersion for college-educated workers that is greater than the increase in growth in dispersion for high-school-educated workers, 3) rising between group dispersion for education, and 4) composition effects linked to the shift from high-school-educated to college-educated workers. In interpreting these findings, it is important to empha-

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

[4] We will use the CPS-ASEC data posted by HLL to the *Journal of Economic Perspectives* website to replicate and extend their results. Unless otherwise stated, all references in this paper to the "CPS" refer to the "CPS-ASEC."

size the important role of rising within-group inequality. Put differently, residual within-group effects play a much larger role than the residual inequality from the matched employer-employee analysis using the AKM approach as in Song et al. (2019) and HHS.

Firms don't play 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 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.[5] 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).[6] We build on HLL's approach using the CPS by quantifying such covariance effects directly. The covariance between observable worker characteristics and industry is related to the sorting and segregation effects emphasized in Song et al. (2019) and HHS.

# 3 Industry and increasing inequality in the CPS

## 3.1 Replicate CPS results from Hoffman, Lee, and Lemieux (2020)

We estimate the following human capital earnings equation:

$$y_i = AgeEduc_i\beta_1 + Occupation_i\beta_2 + Industry_i\beta_3 + \varepsilon_i, \tag{3}$$

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

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

[6] 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."

Table 1: Estimation of the human capital earnings equation using CPS ASEC data and 5-year intervals

| | 1975-<br>1979 | 1980-<br>1984 | 1985-<br>1989 | 1990-<br>1994 | 1995-<br>1999 | 2000-<br>2004 | 2005-<br>2009 | 2010-<br>2014 | 2015-<br>2018 | Growth<br>1975-79 to<br>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 (4). "Occupation" is the marginal contribution of including occupation, obtained by subtracting the percentage of the variance explained by equation (4) from equation (5). "Industry" is the marginal contribution of industry, obtained by subtracting the percentage of variance explained by equation (5) from the percentage of variance explained by equation (6). Industry is defined using 12 SIC categories. "Residual" is the fraction of the variance that is unexplained when estimating equation (6).

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 {high school dropouts, high school graduates, some college, college graduates, and college post-graduates}. The marginal effects of these demographic categories on earnings is given by $\beta_1$. $Occupation_i$ is a vector of dummy variables that is equal to one if worker $i$ is employed in that occupation, and is equal to zero otherwise. The marginal effect of each of the nine occupation categories is given by the vector $\beta_2$. Analogously, $Industry_i$ is a vector of dummy variables of each of (initially) twelve SIC industries, with marginal effects given by the vector $\beta_3$.

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

$$y_i = AgeEduc_i\beta_1 + \varepsilon_i \tag{4}$$

$$y_i = AgeEduc_i\beta_1 + Occupation_i\beta_2 + \varepsilon_i \tag{5}$$

$$y_i = AgeEduc_i\beta_1 + Occupation_i\beta_2 + Industry_i\beta_3 + \varepsilon_i \tag{6}$$

Equation (4) 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 (5) relative to equation (4). The marginal contribution of industry is obtained last by subtracting the percentage of variance explained by equation (5) from the percentage of variance explained by equation (6).

Table 1 replicates HLL Figure 4, with two exceptions: we use labor income instead of total income, and we pool males and females instead of presenting gender specific results. These exceptions really don't matter (we show this in Appendix Table A2.). We offer our thanks to HLL for making their data and replication code available on the *Journal of Economic Perspectives* website.

We find that in each 5-year interval, the contribution of age by education effects are large and growing over time. Occupation effects contribute positively to both levels and growth. The marginal contribution of industry is positive: 0.017 (5.9% of 0.283) in 1975-1979 and 0.011 (2.7% of 0.398) in 2015-2018. However, the marginal contribution of industry is falling over time: -0.006 (-5.3% of 0.115) over the 1975-1979 to 2015-2018 intervals. Importantly, most of the variation in earnings dispersion in levels and changes is unexplained in these CPS human capital earnings equations.

Table 2 mimics Table 1 with three differences: we use 7-year intervals instead of 5-year intervals

Table 2: Estimation of the human capital earnings equation using CPS-ASEC data and 7-year intervals

| | 1975-<br>1981 | 1982-<br>1988 | 1989-<br>1995 | 1996-<br>2002 | 2004-<br>2010 | 2012-<br>2018 | Growth<br>1975-81 to<br>2012-18 | Growth<br>1996-02 to<br>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% | 58.1% | 65.5% |
| | | | | *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% | 58.8% |

*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 we male and females, the year 2000 is deleted, and the way industry is measured. 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 (4). "Occupation" is the marginal contribution of including occupation, obtained by subtracting the percentage of the variance explained by equation (4) from equation (5). "Industry" is the marginal contribution of industry, obtained by subtracting the percentage of variance explained by equation (5) from the percentage of variance explained by equation (6). "Residual" is the fraction of the variance that is unexplained when estimating equation (6).

(so we match the time intervals used by HHS), we use both SIC and NAICS measures of industry, and we delete the year 2000 from the data.[7] 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). The contribution of person effects, occupation effects, and unexplained factors are similar in Tables 1 and 2.

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.

[7] 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.

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 2 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 results are relatively insensitive to different time periods.

## 3.2 Within- and between-industry variance decomposition

We now estimate a simple within and between-industry variance decomposition:

$$\underbrace{\mathrm{var}(y_{i,k} - \bar{y})}_{\substack{\text{earnings} \\ \text{variance}}} = \underbrace{\mathrm{var}(y_{i,k} - \bar{y}_k)}_{\substack{\text{within-industry} \\ \text{dispersion}}} + \underbrace{\mathrm{var}(\bar{y}_k - \bar{y})}_{\substack{\text{between-industry} \\ \text{dispersion}}} \tag{7}$$

The first term on the right hand side of the above equation 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*. Table 3 presents the results of this decomposition for seven-year intervals, using both SIC and NAICS industries.

In each seven-year interval, the level of between-industry variance $\mathrm{var}(\bar{y}_k - \bar{y})$ using SIC is positive: 4.8% of 0.283 in 1975-1979, 4.2% of 0.360 in 1975-1981, and 5.1% of 0.397 in 2012-2018. In each 7-year interval, the level of between-industry variance using NAICS is larger than the between-industry variance using SIC. The between-industry variance component is increasing over time: 9.2% of 0.113 over the 1975-1979 to 2012-2018 intervals using NAICS sectors, and 23.1% of 0.037 over the 1996-2002 to 2012-2018 intervals using NAICS sectors.

Table 3 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 2 (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 is originating from covariances that are implicit in the within and between estimate but excluded from the marginal contribution of industry.

Table 3: Within- and between-industry variance decomposition, CPS-ASEC data

| | 1975-<br>1981 | 1982-<br>1988 | 1989-<br>1995 | 1996-<br>2002 | 2004-<br>2010 | 2012-<br>2018 | Growth<br>1975-81 to<br>2012-18 | Growth<br>1996-02 to<br>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% |
| Between-industry | 4.8% | 4.8% | 4.6% | 4.2% | 4.7% | 5.1% | 6.0% | 13.8% |
| | | | | *Using 18 NAICS industries* | | | | |
| Within-industry | 93.2% | 93.3% | 93.8% | 94.1% | 93.6% | 92.5% | 90.8% | 76.9% |
| Between-industry | 6.8% | 6.7% | 6.2% | 5.9% | 6.4% | 7.5% | 9.2% | 23.1% |

*Notes*: HLL CPS-ASEC data downloaded from *Journal of Economic Perspectives* website. Pooled males and females. The year 2000 is deleted. Earnings is the 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). Definitions follow equation (7).

## 3.3 Variance decomposition of the human capital earnings equation

We re-write the human capital earnings equation used by HLL as

$$y_{i,k} = Z_{i,k}\beta_Z + Industry_{i,k}\beta_3 + \varepsilon_{i,k}, \tag{8}$$

where $Z$ concatenates the $AgeEduc_i$ and $Occupation_i$ vectors, and $\beta_Z$ concatenates the marginal effects vectors $\beta_1$ and $\beta_2$. Define $\overline{Z_k\beta_Z}$ as the industry mean of $Z_{i,k}\beta_Z$. Taking variances of both sides of the human capital earnings equation results in:

$$\begin{aligned}\underbrace{\mathrm{var}(y_{i,k})}_{\substack{\text{earnings}\\\text{variance}}} = {} & \underbrace{\mathrm{var}(Z_{ik}\beta_Z - \overline{Z_k\beta_Z})}_{\substack{\text{within-industry dispersion}\\\text{from age, education,}\\\text{and occupation}}} + \underbrace{\mathrm{var}(\overline{Z_k\beta_Z})}_{\substack{\text{between-industry}\\\text{segregation}}} + \\ & \underbrace{\mathrm{var}(Industry_{i,k}\beta_3)}_{\substack{\text{between-industry}\\\text{pay premium}}} + \underbrace{2\mathrm{cov}(\overline{Z_k\beta_Z}, Industry_{i,k}\beta_3)}_{\text{between-industry sorting}} + \underbrace{\mathrm{var}(\varepsilon_{i,k})}_{\substack{\text{residual dispersion}\\\text{(within-industry)}}}\end{aligned} \tag{9}$$

Each of the terms on the right hand side of this variance decomposition can be labeled with the terminology of Song et al. (2019) and Haltiwanger, Hyatt, and Spletzer (2022): $\mathrm{var}(Z_{i,k}\beta_Z - \overline{Z_k\beta_Z})$ is the within-industry effect of observable person characteristics, $\mathrm{var}(\overline{Z_k\beta_Z})$ is industry segregation, defined as how persons with similar observables ($Z\beta_Z$) concentrate within industries, $\text{var}(Industry_{i,k}\beta_3)$ is the industry pay premium, 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-paid industries, and how low-paid workers in terms of observable characteristics work for low-paid industries.

Table 4 presents the results of this variance decomposition for 7-year intervals, using both SIC and NAICS industries. By construction, the between-industry overall contribution is equal to industry segregation + industry pay premium + industry sorting. Looking at variance growth from 1996-2002 to 2012-2018, and using NAICS sectors (our preferred specification in the bottom right panel of Table 4), industry segregation and industry sorting are positive (14.8% and 7.3% respectively), with the industry pay premium 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. Looking at variance growth from 1996-2002 to 2012-2018, 58.8% of variance growth is unexplained when using NAICS. This is very different than the -3.9% in HHS.

Tables 2 and 4 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 2 and 4. In the former, age by education plus occupation effects account for 40.5% of the increase in earnings inequality. In Table 4, 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 4, 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 2. While it is not an identity that the marginal contribution of industry in Table 2 is equal to the industry pay premium in Table 4 (the covariance structures underlying the two tables are different), they are similar in magnitude.

Table 4: Variance decomposition of the human capital earnings equation, CPS-ASEC data

| | 1975-<br>1981 | 1982-<br>1988 | 1989-<br>1995 | 1996-<br>2002 | 2004-<br>2010 | 2012-<br>2018 | Growth<br>1975-81 to<br>2012-18 | Growth<br>1996-02 to<br>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: | | | | | | | | |
| 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: | | | | | | | | |
| Segregation | 2.2% | 2.4% | 2.8% | 3.0% | 3.5% | 4.0% | 8.5% | 13.3% |
| Pay premium | 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: | | | | | | | | |
| 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: | | | | | | | | |
| Segregation | 2.9% | 2.9% | 3.2% | 3.3% | 3.8% | 4.4% | 8.1% | 14.8% |
| Pay premium | 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 (9) for definitions.

Table 5: Variance decomposition of the human capital earnings equation, CPS-ASEC data, with detail on age, education, and occupation

| | | | | | | | Growth | Growth |
|---|---|---|---|---|---|---|---|---|
| | 1975- | 1982- | 1989- | 1996- | 2004- | 2012- | 1975-81 to | 1996-02 to |
| | 1981 | 1988 | 1995 | 2002 | 2010 | 2018 | 2012-18 | 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: | | | | | | | | |
| 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% | 58.8% |
| Between-industry: | | | | | | | | |
| 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 premium | 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 (9).

The person characteristics, sorting, and segregation effects in Table 4 reflect the contribution of both age by education effects and occupation effects. We provide guidance about the relative contribution of age by education vs occupation effects in Table 5. Focusing on results using NAICS sectors, 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 5), 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 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.

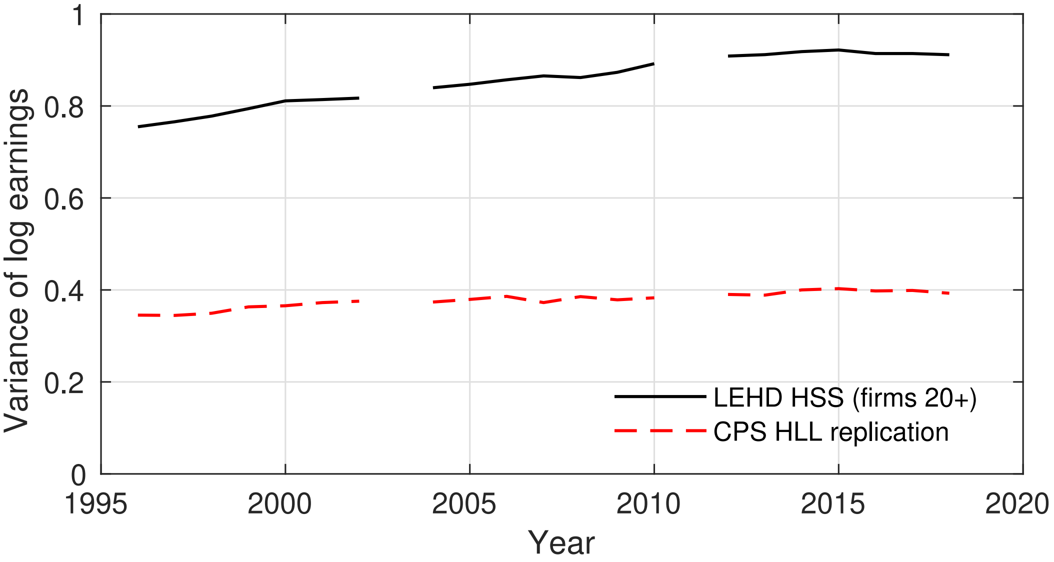
It is also worth emphasizing that the contribution of occupation is mostly via sorting and segregation effects between industries. Within-industry occupation effects for 6.3% of rising dispersion including both the direct and covariance effects. Between-industry segregation effects from occupation contribute 10.9% including both direct and covariance effects. Between-industry sorting effects from occupation account for 5.3%.

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

We seek to understand what underlies the different 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 5) 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 0.330 in 2018 while the variance off 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

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



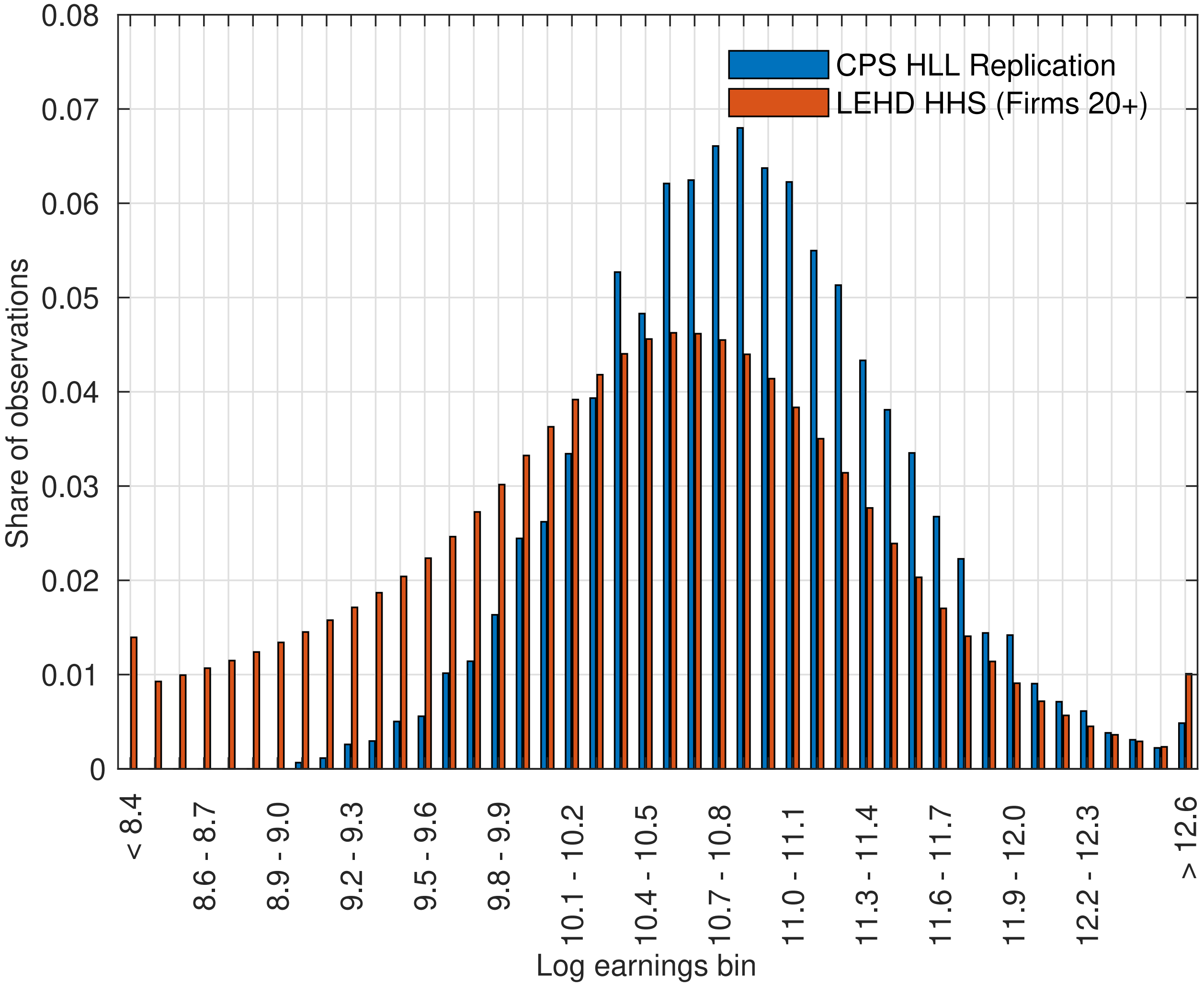
than the HLL CPS. The HLL CPS data is bottom coded at annual earnings=\$7840 (weeks worked $> 49$, usual hours $> 40$, and a real hourly wage $>$ \$4 using a CPI 2018=100 deflator), whereas the HLL LEHD data is bottom coded at annual earnings =\$3770 (weeks worked $> 13$ and a real hourly wage $>$ \$7.25 using a PCE 2013=100 deflator). Figures 1 and 2 clearly show that HLL and HHS are not analyzing increasing inequality using the same annual earnings distributions.

Table 6 shows the eight identifiable differences in the HLL CPS and the HHS LEHD data. The middle column of Table 6 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 2018=100 CPI to 2013=100 PCE in the CPS; (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 2: PDFs of HLL CPS-ASEC and HHS LEHD earnings, all years pooled



tributor to these changes is the bottom coding, where we add many lower earnings individuals to the HLL data. Appendix Figure A1 shows the effects, one-by-one, of the common coding on the HLL data. 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 quite 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 al. (2013) finding that low earnings is one characteristic predicting having an LEHD earnings record and not being measured as employed in the CPS.

Table 6: 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 $>$ 20 |
| Years | 1975-2018 | 1996-2002, 2004-2010, & 2012-2018 | 1996-2002, 2004-2010, & 2012-2018 |

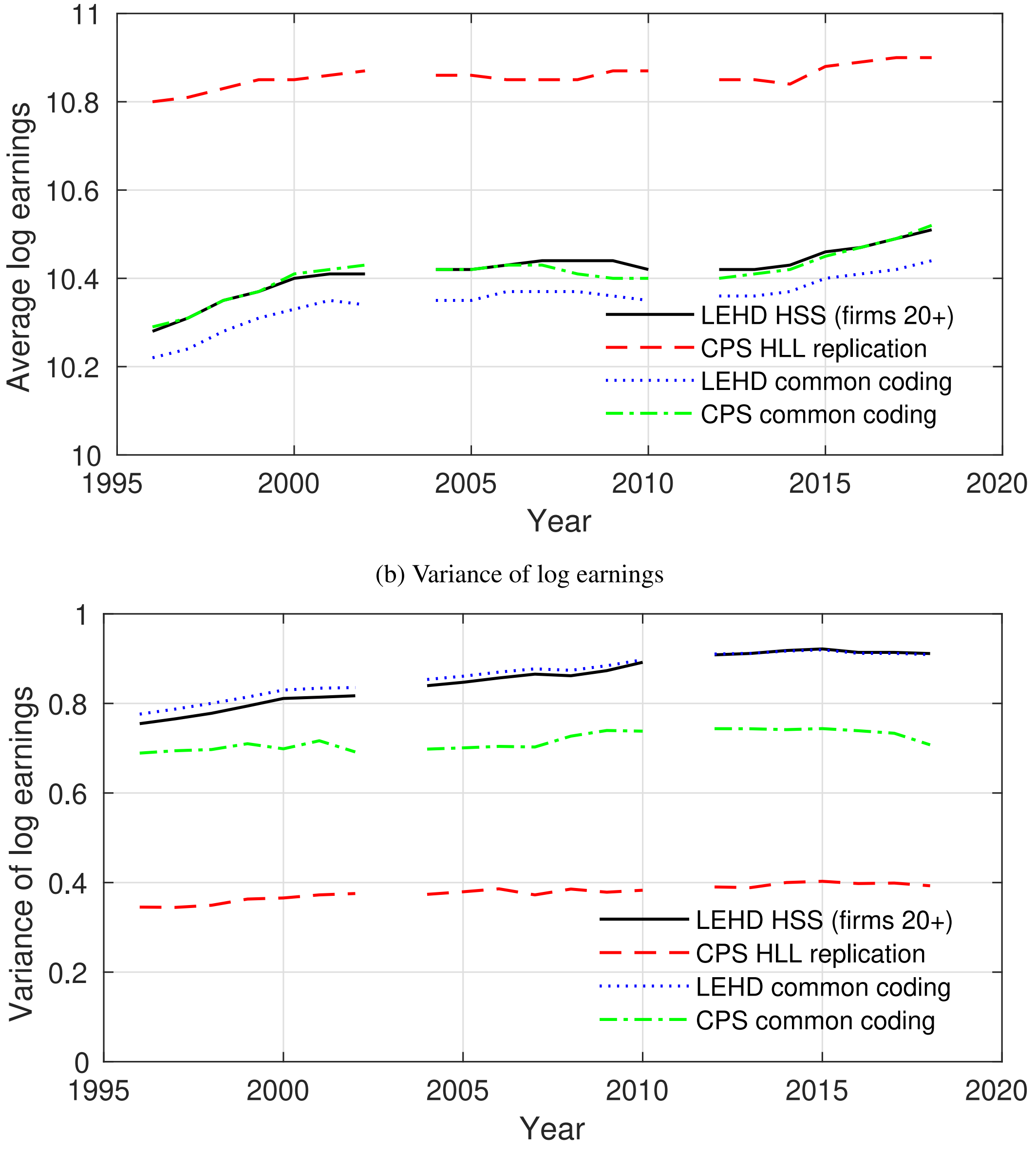
Figures 1 and 2 show us that HLL and HSS analyzed different earnings distributions. Figure 4 shows us that the common coded CPS and the common coded LEHD have similar earnings distributions. But as we will show in the next section, common coding does not substantially change the between-industry contribution 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.

# 5 A linked CPS-LEHD dataset

## 5.1 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

Figure 3: Mean and variance of HLL CPS-ASEC, HHS LEHD, common-coded, CPS-ASEC, and common-coded LEHD earnings, by year

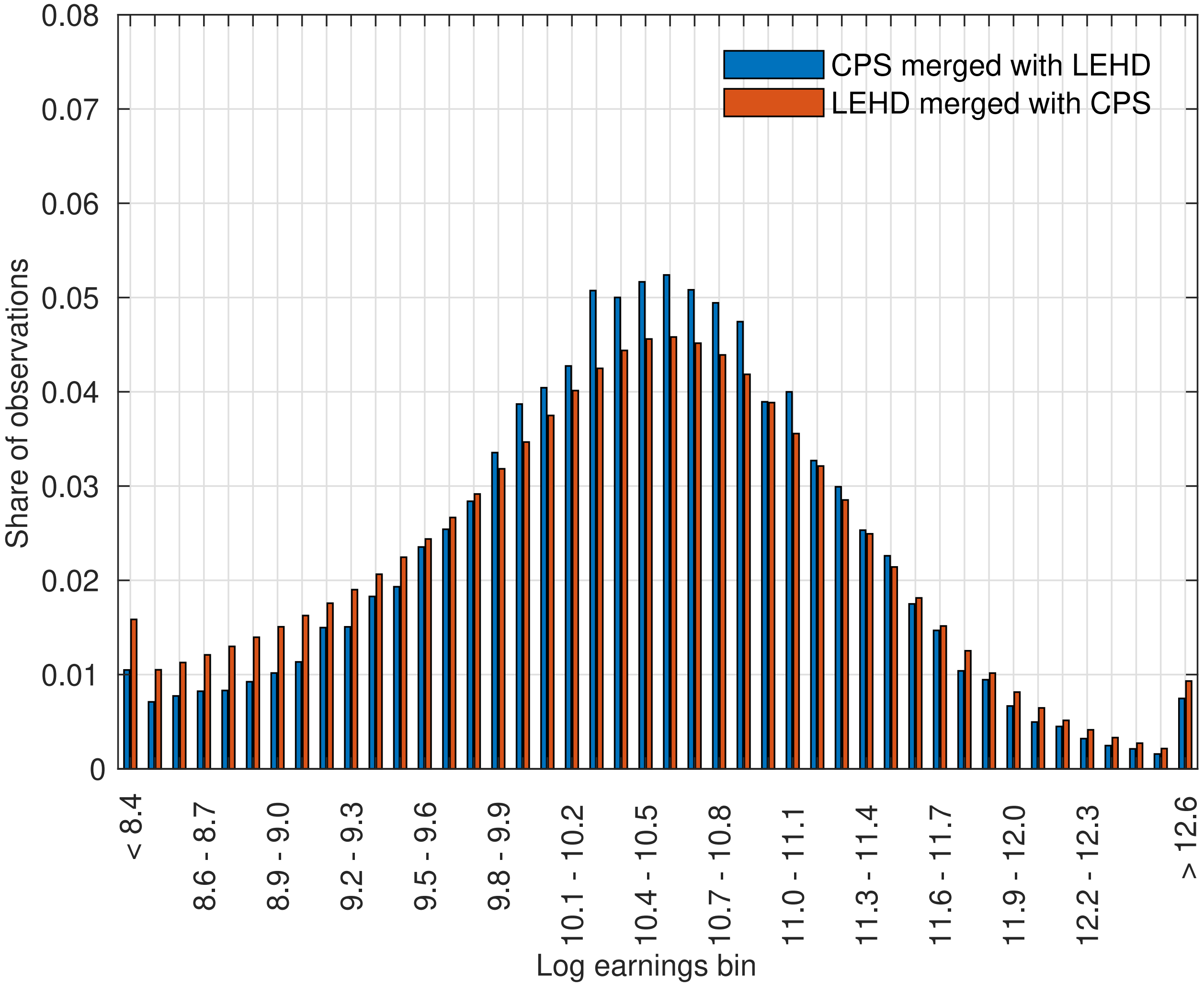

(a) Average log earnings

(b) Variance of log earnings

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

Figure 4: PDFs of common-coded CPS-ASEC and common-coded LEHD earnings, all years pooled



are poor quality for earnings reference year 2000, and we do not link the 2000 CPS-ASEC with the LEHD. The year-specific PIK rates for the common coded CPS-ASEC are given in Appendix Figure A3.

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 “1 if the CPS-ASEC record has a PIK, 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 dividing the original weight by the predicted value.

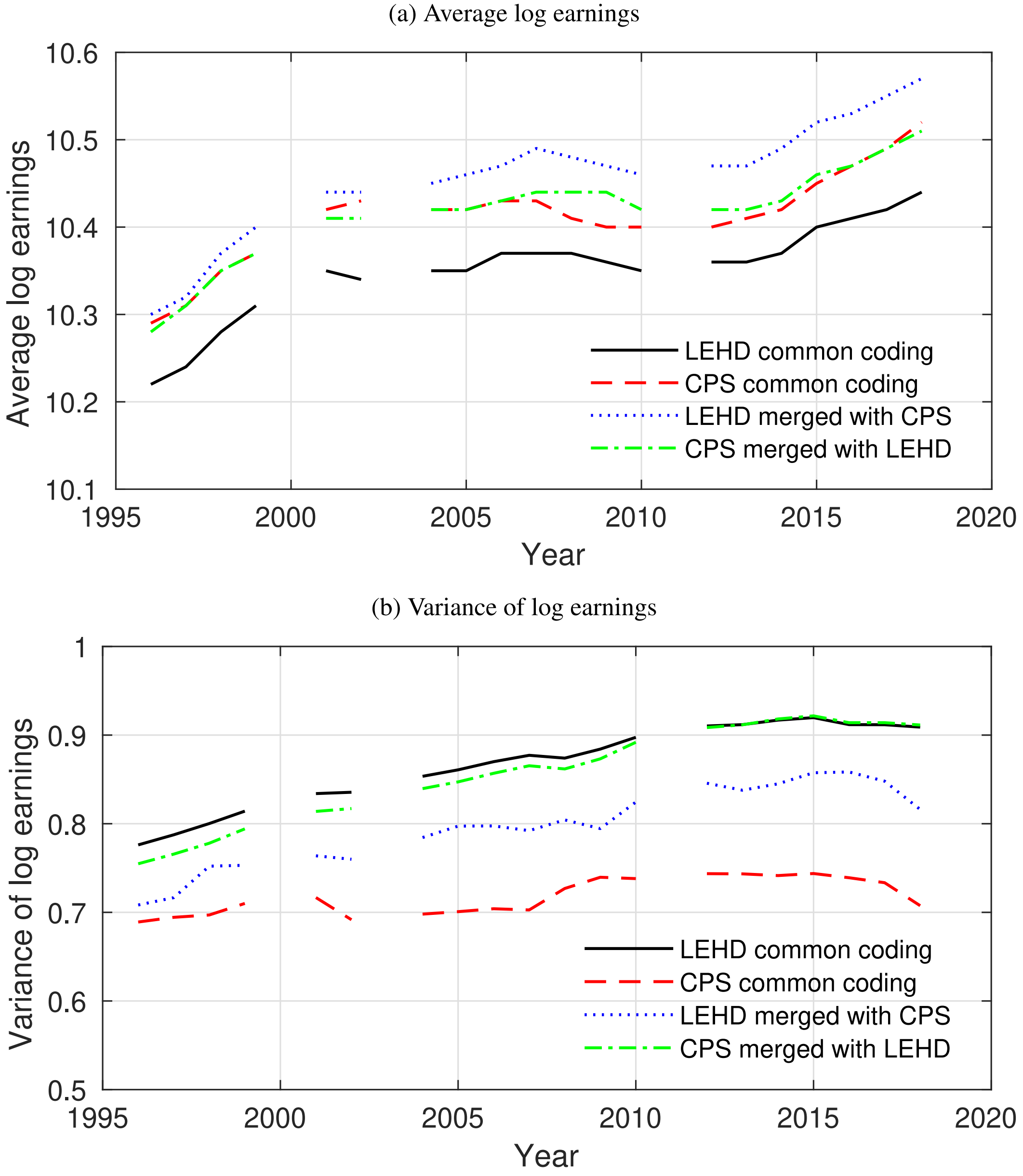
We then merge the PIKed common coded CPS-ASEC data with the common coded LEHD data. We only keep observations where an individual is in both the CPS-ASEC and in the LEHD.[9] We run another set of year-specific logistic regressions where the dependent variable is "1 if PIKed CPS-ASEC matches to the LEHD, 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. 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.

Two notes about the CPS-LEHD linked data warrant mention. First, the linked CPS-LEHD microdata contains the AKM parameters $\{\theta^{i,p}, \psi^{j,k,p}, \beta^{p}\}$ that were estimated on the full LEHD data with 946 million person-year observations. This will be important later. Second, 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.

To illustrate the similarities and differences in earnings in the CPS and LEHD in the CPS-LEHD linked data, Figure A4 shows the equivalent of Figure 4 from the linked data. The two earnings distributions are roughly similar, but Figure A5 illustrates substantial differences when computing the pdf of CPS minus LEHD earnings at the individual level. In Figure A5, there is substantial mass near zero, but there are clear differences in the tails. Figure A6 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 A6. However, even after removing imputed cases the pattern in the tails remains. These earnings differences in the

[9] 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 is built from 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 whereas the CPS is a survey, so there are many individuals in the LEHD who are not sampled in the CPS.

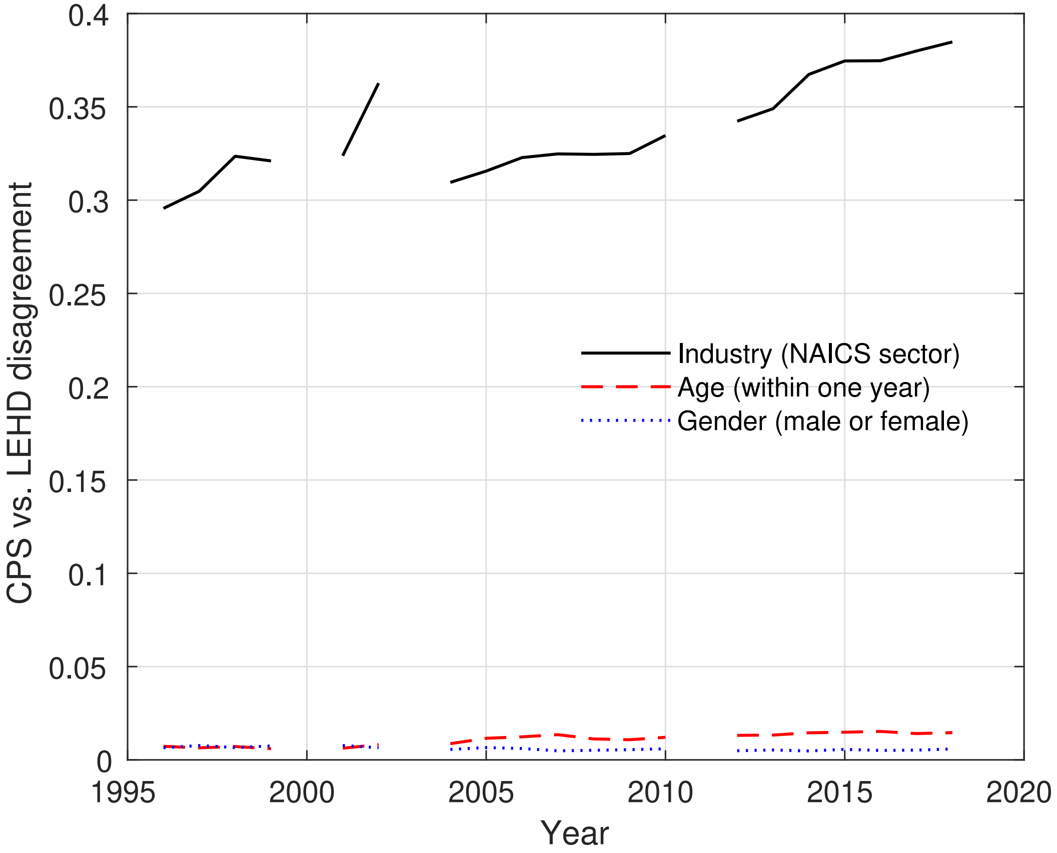
Figure 5: Mean and variance of HLL CPS-ASEC, HHS LEHD, common-coded, CPS-ASEC, and common-coded LEHD earnings, by year


(a) Average log earnings

(b) Variance of log earnings

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.

Figure 6: Linked CPS-ASEC and LEHD differences



*Notes*: 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.

## 5.2 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 6 shows the disagreement between age and gender in the CPS and the LEHD. 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. Disagreement 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)).

# 6 Increasing inequality in the linked CPS-LEHD dataset

## 6.1 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 7 presents the variance decomposition using our minor modifications of HLL's CPS-ASEC sample (this is the same sample used in Tables 2 to 5). 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%. As we discuss below in section 7, several factors underlie this finding. Details are below but 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).

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

Table 7: Within and between-industry variance decompositions

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Data | CPS | CPS | Linked | Linked | Linked | Linked | |
| | | | CPS-LEHD | CPS-LEHD | CPS-LEHD | CPS-LEHD | LEHD |
| Sample | HLL JEP | Common | Common | Common | Common | Common | Common |
| | | coded | coded | coded | coded | coded | 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:* The rows titled “Data” and “Sample” indicate the data used for the variance decomposition (see text for description). The row titled “Earnings measure” indicates whether CPS or LEHD earnings is used in the decomposition. In the row titled “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 (7).

between-industry variance growth is unaffected. Note 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. 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 7 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.

## 6.2 A full variance decomposition of the human capital earnings equation

Table 8 shows the variance decomposition with columns corresponding to Table 7 (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 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. Before turning to these differences, one striking finding is that moving from CPS to LEHD industry codes substantially reduces the unexplained portion of the increase in CPS earnings inequality (compare 38.3% in column (3) to either 25.9% or 28.5% in columns (4) and (5)).

The results from columns (5) and (6) have some important similarities and differences. One key similarity that we already know from Table 7 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 (31.7% compared to 14.8%) while industry pay premia is much more important using LEHD earnings (22.0% compared to -1.2%). 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 increase in dispersion in the

Table 8: Variance decomposition of the human capital earnings equation

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Data | CPS | CPS | Linked | Linked | Linked | Linked |
| | | | CPS-LEHD | CPS-LEHD | CPS-LEHD | CPS-LEHD |
| Sample | HLL JEP | Common | Common | Common | Common | Common |
| | | coded | coded | coded | coded | 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: | | | | | | |
| Age, educ. & occ. | 28.5% | 24.4% | 27.7% | 27.9% | 22.2% | 18.7% |
| Residual | 65.6% | 62.5% | 60.0% | 63.8% | 60.3% | 59.6% |
| Between-industry: | | | | | | |
| Segregation | 3.3% | 2.2% | 2.2% | 2.0% | 4.1% | 3.1% |
| Pay premium | 6.0% | 11.1% | 8.7% | 4.3% | 8.2% | 12.2% |
| Sorting | -3.4% | -0.1% | 1.5% | 2.0% | 5.2% | 6.4% |
| | | | *Variance level 2012-2018* | | | |
| Earnings variance | 0.397 | 0.738 | 0.717 | 0.717 | 0.717 | 0.845 |
| Within-industry: | | | | | | |
| Age, educ. & occ. | 27.5% | 24.5% | 26.8% | 27.5% | 21.1% | 18.1% |
| Residual | 65.0% | 61.6% | 58.5% | 61.2% | 58.0% | 55.0% |
| Between-industry: | | | | | | |
| Segregation | 4.4% | 3.7% | 4.2% | 3.9% | 6.1% | 4.5% |
| Pay premium | 5.5% | 8.4% | 7.0% | 3.9% | 7.5% | 13.4% |
| Sorting | -2.4% | 1.8% | 3.5% | 3.6% | 7.3% | 9.1% |
| | | | *Change from 1996-02 to 2012-18* | | | |
| Variance growth | 0.037 | 0.035 | 0.050 | 0.050 | 0.050 | 0.100 |
| Within-industry: | | | | | | |
| Age, educ. & occ. | 18.2% | 27.0% | 15.1% | 21.9% | 6.0% | 13.3% |
| Residual | 58.8% | 43.6% | 38.8% | 25.9% | 28.5% | 20.6% |
| Between-industry: | | | | | | |
| Segregation | 14.8% | 34.9% | 31.5% | 28.5% | 31.7% | 14.8% |
| Pay premium | 1.0% | -45.8% | -14.9% | -1.4% | -1.2% | 22.0% |
| Sorting | 7.3% | 40.2% | 29.5% | 25.1% | 35.1% | 29.4% |

*Notes:* The rows titled "Data" and "Sample" indicate the data used for the variance decomposition (see text for description). The row titled "Earnings measure" indicates whether CPS or LEHD earnings is used in the decomposition. In the row titled "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 (9) for definitions.

Table 9: Variance decomposition of the AKM earnings equation

| | Linked CPS-LEHD<br>LEHD earnings<br>299 LEHD industries | | | Common Coded LEHD | | |
|---|---|---|---|---|---|---|
| | 1996-02 | 2012-18 | Growth | 1996-2002 | 2012-2018 | Growth |
| Earnings variance | 0.746 | 0.845 | 0.100 | 0.811 | 0.914 | 0.103 |
| Between-firm: | 18.7% | 19.1% | 22.7% | 17.5% | 18.3% | 24.0% |
| Segregation | 9.7% | 9.9% | 11.0% | 9.2% | 9.6% | 12.1% |
| Pay premium | 6.4% | 5.9% | 2.3% | 6.1% | 5.8% | 4.1% |
| Sorting | 2.5% | 3.3% | 9.3% | 2.3% | 2.9% | 7.8% |
| Between-industry: | 21.7% | 26.9% | 66.2% | 20.4% | 25.4% | 64.5% |
| Segregation | 7.4% | 9.7% | 26.9% | 6.9% | 9.0% | 25.9% |
| Pay premium | 4.1% | 4.8% | 9.6% | 4.1% | 4.6% | 9.3% |
| Sorting | 10.2% | 12.5% | 29.6% | 9.5% | 11.7% | 29.2% |
| Within-firm: | 59.7% | 54.0% | 11.1% | 62.0% | 56.4% | 11.6% |
| Person effect & obs. | 47.6% | 44.0% | 17.8% | 46.5% | 42.7% | 13.2% |
| Residual | 12.1% | 9.9% | -6.6% | 15.6% | 13.6% | -1.7% |

*Notes:* Definitions follow Appendix equation (A5). "Between-firm" measures between-firm, within-industry dispersion. "Person effect & obs." refers to heterogeneity attributed to the person effect and from observable characteristics in an AKM regression as described in equation (A5).

CPS earnings measure is only about half that for the LEHD earnings measure.

## 6.3 Unexplained contribution in the CPS vs. the LEHD data

An important factor in reconciling the contributing factors to rising earnings inequality between the CPS and LEHD data is the substantially larger role of unexplained factors in the CPS when using observable person characteristics. Several components are at work here. As highlighted in our discussion of Table 8, the unexplained contribution to variance growth is reduced substantially when using industry codes from the LEHD. It is reduced further in the linked CPS-LEHD (column 6 of Table 8) when using LEHD earnings along with CPS observable person characteristics and detailed LEHD industry codes. Still, compared to the findings in Song et al. (2019) and HHS, the unexplained component of rising earnings inequality is much greater using CPS observable person characteristics than using the person and firm characteristics from an AKM decomposition of earnings. In this section, we compare and contrast the latter with the CPS observable person characteristics.

The AKM earnings equation is stated above in equation 1. We have estimated $\{\theta_i, \psi_j, X_{i,t}\beta\}$ for each 7-year time period from the full 946 million observation common coded LEHD, and we have used these estimates to create the industry-enhanced variance decomposition reported in Table 9. These estimated AKM components are also in the CPS-LEHD linked data, which allows us to create the industry-enhanced variance decomposition from the linked CPS-LEHD data. These variance decompositions are also reported in Table 9.

The industry enhanced variance decompositions reported in Table 9 are quite similar for the linked data and the full LEHD data. Between-industry variance growth is 66.2% in the linked CPS-LEHD and is 64.5% in the full LEHD (and is 61.9% in HHS). Within-industry between-firm variance growth is 22.7% in the linked CPS-LEHD and is 24.0% in the full LEHD (and is 23.1% in HHS). Within-industry between-person variance growth is 11.1% in the linked CPS-LEHD and is 11.6% in the full LEHD (and is 14.9% in HHS). The industry segregation, pay premium, and sorting terms are also amazingly similar across datasets. Looking at variance growth, industry segregation is 26.9% in the linked CPS-LEHD, 25.9% in the full LEHD, and is 25.2% in HHS. The industry pay premium is 9.6% in the linked CPS-LEHD, 9.3% in the full LEHD, and is 8.7% in HHS. Industry sorting is 29.6% in the linked CPS-LEHD, 29.2% in the full LEHD, and is 28.0% in HHS.

It is instructive to compare the results from Table 9 to Table 8 more directly. In Table 10, we repeat columns 5 and 6 of Table 8, and we repeat components from Table 9 in the third column of Table 10. All three columns of Table 10 use the linked CPS-LEHD data and all three use the detailed industry codes from LEHD.

The contribution of total between-industry is by construction identical in the right two columns and very similar in the first column. However, the allocation of the total between-industry contribution into segregation, sorting, and pay premium differs substantially on some dimensions. At the core of these differences is the difference between using observable person characteristics in the human capital equation in the middle column compared to using person and firm effects from an AKM decomposition in the last column. Even though the person effects in the last column reflect only the within firm contribution of such effects, this component still accounts for more than (or equal to) the within-industry observable person characteristics from the CPS. If we add the within-firm person components and the within-industry between-firm components from the AKM earnings regression, this difference is substantially larger than the observable person characteristics effect from the CPS.[10] This greater role for within-industry contributions in the far right column is not coming from the

[10] The within-industry between firm effects include within-industry between firm segregation and sorting as well as firm effects. The segregation effects reflect a regrouping of person effects and are the most important component.

Table 10: Variance decomposition of the human capital earnings equation (from Table 8) vs. the AKM earnings equation (from Table 9)

| | (5) | (6) | (7) |
|---|---|---|---|
| Earnings measure | CPS | LEHD | LEHD |
| Specification | Human capital | Human capital | AKM |
| *Variance level 1996-2002* | | | |
| Earnings variance | 0.667 | 0.746 | 0.746 |
| Within-industry: | | | |
| Age, educ. & occ. | 22.2% | 18.7% | |
| Person effect & obs. | | | 47.6% |
| Residual | 60.3% | 59.6% | 12.1% |
| Between-firm | | | 18.7% |
| Between-industry: | | | |
| Segregation | 4.1% | 3.1% | 7.4% |
| Pay premium | 8.2% | 12.2% | 4.1% |
| Sorting | 5.2% | 6.4% | 10.2% |
| *Variance level 2012-2018* | | | |
| Earnings variance | 0.717 | 0.845 | 0.845 |
| Within-industry: | | | |
| Age, educ. & occ. | 21.1% | 18.1% | |
| Person effect & obs. | | | 44.0% |
| Residual | 58.0% | 55.0% | 9.9% |
| Between-firm | | | 19.1% |
| Between-industry: | | | |
| Segregation | 6.1% | 4.5% | 9.7% |
| Pay premium | 7.5% | 13.4% | 4.8% |
| Sorting | 7.3% | 9.1% | 12.5% |
| *Change from 1996-02 to 2012-18* | | | |
| Variance growth | 0.050 | 0.100 | 0.100 |
| Age, educ. & occ. | 6.0% | 13.3% | |
| Person effect & obs. | | | 17.8% |
| Residual | 28.5% | 20.6% | -6.6% |
| Between-firm | | | 22.7% |
| Between-industry: | | | |
| Segregation | 31.7% | 14.8% | 26.9% |
| Pay premium | -1.2% | 22.0% | 9.6% |
| Sorting | 35.1% | 29.4% | 29.6% |

*Notes:* Linked CPS-LEHD dataset, common coded sample. 299 4-digit NAICS industries. Columns 5 and 6 are copied directly from Table 8. Column 7 is copied from Table 9.

between-industry contribution but rather the unexplained contribution is much smaller in the far right column compared to the middle column.

Looking at variance levels in Table 10, the contribution of observable person characteristics differs dramatically using the human capital earnings equation relative to the person effects using the AKM earnings equation. Using the human capital earnings equation, the within-industry variation in the Z vector (age*education and occupation, multiplied by $\beta$) accounts for 18 to 22 percent of the earnings variance, whereas the within firm person effects accounts for 44 to 48 percent of the earnings variance in the AKM earnings equation (and another 19% inclusive of the within-industry between firm effects). This difference is reversed for the unexplained contribution to the level of earnings variance. In the human capital earnings equation, 55 to 60 percent of the earnings variance is unaccounted for, whereas this is 10 to 12 percent in the AKM earnings equation.

Turning to the between-industry contribution and focusing on growth, the overall between-industry contribution in the latter two columns in Table 10 is the same by construction. Interestingly, sorting is basically identical whether using the human capital equation approach or the AKM earnings equation approach: the former is 29.4% and the latter is 29.6%. Segregation using the human capital equation approach is 14.8 percent of variance growth, whereas it is 26.9% of variance growth using the AKM equation approach. The industry pay premium is 22.0 percent in the human capital equation, versus 9.6 percent in the AKM equation. In other words, when accounting for the determinants of between-industry variance growth, the industry dummies are more important than $Zb$ in the human capital equation, whereas the AKM individual fixed effects $\theta_i$ (plus $X\beta$) are more important than the industry mean of the AKM firm fixed effects in the AKM earnings equation.

An implication from Table 10 is that one can go pretty far with a hybrid approach based on administrative data on earnings, detailed high-quality industry codes from administrative data, and survey data on person characteristics such as age, education and occupation. By construction, the overall role of industry is the same, and its decomposition into sorting, segregation, and industry pay premia are broadly similar whether using observable person characteristics or AKM person effects.[11] There is still substantial unexplained variation in earnings using such a hybrid approach but such unexplained variation is substantially diminished compared to using only survey data on earnings and industry.

[11] In the next draft, we will include a breakout of the contribution of age*education and occupation for the results in this section.

Table 11: Correlation matrix, all years pooled, linked CPS-ASEC and LEHD data

| | Std. Deviation | CPS earnings | LEHD earnings | Age, educ. & occ. (AEO) | AEO ind. segregation | AEO ind. pay premium | AKM person & obs. | AKM firm segregation | AKM firm pay premium | AKM ind. segregation | AKM ind. pay premium | CPS AEO residual | LEHD AKM residual |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CPS earnings | 0.832 | 1.000 | 0.759 | 0.467 | 0.366 | 0.333 | 0.516 | 0.285 | 0.211 | 0.425 | 0.404 | 0.768 | 0.148 |
| LEHD earnings | 0.896 | | 1.000 | 0.426 | 0.394 | 0.371 | 0.659 | 0.348 | 0.296 | 0.480 | 0.467 | 0.452 | 0.346 |
| Age, educ. & occ. (AEO) | 0.389 | | | 1.000 | 0.000 | 0.000 | 0.467 | 0.242 | 0.122 | 0.000 | 0.000 | 0.000 | 0.009 |
| AEO ind. segregation | 0.191 | | | | 1.000 | 0.412 | 0.001 | 0.001 | -0.003 | 0.818 | 0.709 | 0.000 | -0.003 |
| AEO ind. pay premium | 0.255 | | | | | 1.000 | -0.047 | 0.003 | 0.000 | 0.698 | 0.783 | 0.000 | -0.006 |
| AKM person & obs. | 0.607 | | | | | | 1.000 | -0.007 | -0.004 | -0.005 | -0.011 | 0.388 | -0.029 |
| AKM firm segregation | 0.282 | | | | | | | 1.000 | 0.172 | -0.001 | -0.004 | 0.222 | -0.008 |
| AKM firm pay premium | 0.222 | | | | | | | | 1.000 | -0.005 | -0.004 | 0.120 | -0.001 |
| AKM ind. segregation | 0.263 | | | | | | | | | 1.000 | 0.885 | 0.000 | -0.006 |
| AKM ind. pay premium | 0.197 | | | | | | | | | | 1.000 | 0.000 | -0.005 |
| CPS AEO residual | 0.639 | | | | | | | | | | | 1.000 | 0.190 |
| LEHD AKM residual | 0.333 | | | | | | | | | | | | 1.000 |

*Notes:* "Age, education, & occ. (AEO)" is defined as $Z_{i,k}\beta - \overline{Z_k\beta}$, i.e., within-industry, where $Z$ is *AgeEduc* and occupation. "AEO ind. segregation" is defined as $\overline{Z_k\beta}$, i.e., the first moment. "AEO ind. pay premium" is defined as $\text{Industry}_{i,k}\beta_3$. "AKM person & obs." is defined as $\theta^i - \bar{\theta}^{j,k} + X_t^i\beta - \bar{X}^{j,k}\beta$, i.e., within-industry. "AKM firm segregation" is defined as $\bar{\theta}^k) + \bar{X}^k\beta$, i.e., the first moment. "AKM firm pay premium" is defined as $\psi^{j,k} - \bar{\psi}^k)$, i.e., within-industry. "AKM ind. segregation" is defined as $\bar{\theta}^k) + \bar{X}^k\beta)$, i.e., the first moment. "AKM ind. pay premium" is defined as $\text{var}(\bar{\psi}^k)$. "CPS residual" is defined as $Y_{i,k} - Z_{i,k}\beta - Industry_{i,k}\beta_3$, corresponding to column 5 in Table 7. "LEHD AKM residual" is defined as $y_t^{i,j,k,p} - \theta^{i,p} - \psi^{j,k,p} - X_t^{i,p}\beta^p$.

To shed more light on these patterns, Table 11 presents the correlations between the various components contributing to rising earnings inequality as measured from the CPS human capital earnings relationships versus the AKM decomposition of earnings. While we find strong relationships between the CPS person and LEHD (AKM) person effects (a correlation of 0.467), it is apparent there is much more variation in the LEHD person effects (standard deviations of 0.607 and 0.389). Relatedly, we find that the CPS residual is strongly positively correlated with LEHD person effect, the LEHD within-industry between firm segregation effect, and the LEHD firm pay premium (the correlations of 0.388, 0.222, and 0.200 respectively). The CPS decompositions attribute to the residual what the LEHD decompositions attribute to person, firm segregation, and firm pay premia effects. The correlation patterns in Table 11 help provide more guidance for the much larger role of person effects and firm effects in moving from column (6) to (7) in Table 10.

# 7 Sensitivity analysis

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.[12] 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. First, we note that in related work (see Haltiwanger, Hyatt, and Spletzer (2022)) we present evidence that key aspects of the inferences from administrative data are robust to using 18 vs 50 states.[13] 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 7 for the 18 states that are in the LEHD (that is, the CPS for 18 states without restricting to

[12] 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.

[13] Two exercises warrant mention here. 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, which suggests that the 18 state versus 50 state distinction does not matter in administrative data. Second, in results that have not yet gone through the Census Bureau disclosure process, we compare the percentage of between firm variances that is between industries in the 18 state LEHD to the Census Bureau's 18- and 50-state Longitudinal Business Database (LBD). We expect these LBD results to be available in mid-to-late July 2022.

Table 12: Variance decompositions for common coded CPS, 18 vs. 50 States

| Data | CPS<br>50 states | CPS<br>CPS 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:* DC is included in "50 states."

being linked to LEHD). The first column of Table 12 repeats the results from column 2 of Table 7 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 13. Two sectors stand out as accounting for the 18 vs 50 state difference: Retail Trade and Information. The difference in contribution of these two sectors (0.0047 = 0.0019 + 0.0028) exceeds the 50 versus 18 state difference in the total contribution (0.0043 = 0.0148 - 0.0105).

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

Table 13: Industry contributions to between-industry growth in variance for common coded CPS, 18 vs 50 states

| Industry | CPS<br>50 states | CPS<br>18 states |
|---|---|---|
| Agriculture | -0.0008 | -0.0015 |
| Mining | 0.0016 | 0.0017 |
| Construction | 0.0000 | -0.0001 |
| Manufacturing | -0.0021 | -0.0014 |
| Wholesale Trade | -0.0005 | -0.0007 |
| Retail Trade | 0.0035 | 0.0054 |
| Transport + Ware | -0.0007 | -0.0007 |
| Utilities | -0.0001 | 0.0000 |
| Information | 0.0054 | 0.0082 |
| Finance + Insuran | 0.0050 | 0.0051 |
| real Estat + Rent | 0.0002 | 0.0002 |
| Prof + Bus Serv | 0.0010 | 0.0008 |
| Educational Serv | -0.0005 | -0.0003 |
| Healthcare + Soc | 0.0004 | 0.0006 |
| Arts, Ent, Rec | 0.0001 | 0.0003 |
| Accom + Food Serv | 0.0082 | 0.0084 |
| Other Services | -0.0005 | -0.0006 |
| Unknown (2nd job) | -0.0097 | -0.0106 |
| Total | 0.0105 | 0.0148 |

*Notes:* DC is included in "50 states."

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 Appendix Table A2. 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, 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

Table 14: Comparisons of the Retail Trade and Information sectors in alternative 18 and 50 state samples

| Data | CPS (Micro) | CPS (Micro) | CPS (Agg) | CPS (Agg) | QCEW (Agg) | QCEW (Agg) |
|---|---|---|---|---|---|---|
| | 50 states | 18 states | 50 states | 18 States | 50 States | 18 States |
| Contribution to variance growth | | | | | | |
| from 1995-2002 to 2012-18: | | | | | | |
| Retail Trade | 0.0035 | 0.0054 | 0.0024 | 0.0041 | 0.0066 | 0.0078 |
| Information | 0.0054 | 0.0082 | 0.0050 | 0.0076 | 0.0030 | 0.0038 |
| | | | | | | |
| Ratio of 50 State to 18 State (%): | CPS (Micro) | | CPS (Agg) | | QCEW (Agg) | |
| Retail Trade | 64.8% | | 59.3% | | 84.8% | |
| Information | 65.9% | | 65.8% | | 78.7% | |

*Notes:* CPS Micro are tabulations from Common Coded CPS Micro data; 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 of sector in each interval is given by: where is the employment share, is the mean of the wage for the sector, and is the grand mean. 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.

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 14 focusing on these two key sectors. We refer to the state by sector level data for the CPS and QCEW as "Agg" in this table.

For the CPS, the contribution to between-industry from the micro data vs. the "Agg" for these two sectors is similar. For Retail Trade, the ratio of the of the 50 to 18 state contribution is about 65% for the micro data and 59% for the "Agg" 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. The measurement issues with CPS industries discussed above is one possible explanation.

# 8 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. Finally, and importantly, most of the rise in inequality is within cell – i.e., unexplained.

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. Using an AKM decomposition of earnings into person and firm effects, most of the rising between firm dispersion is accounted for by increases in the segregation of workers by person effects across firms and sorting of high person effect workers to high firm effect firms. A recent refinement of this message is that between-industry effects dominate the between firm channels. That is, most of the rising between firm dispersion is accounted for by between-industry dispersion. Moreover, most of the rising segregation and sorting reflects between-industry segregation and sorting. Also, importantly, the AKM decomposition leaves little of earnings dispersion in levels or changes unexplained.

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. We show that the sorting and segregation interpretation from the recent administrative data literature can be used with the household data when applied to between-industry variation. In combination with using the high quality and detailed industry codes 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. The decomposition of between-industry dispersion into industry premia, sorting and segregation is sensitive to using CPS vs LEHD earnings as well as sensitive to using observable person characteristics or AKM person effects. Still, using administrative data earnings, the messages are broadly similar using either observable person or AKM person effects. Between-industry sorting accounts for about 30% of rising dispersion using either approach. Industry premia is more important when using observable person characteristics and segregation more important using AKM person effects. These patterns are not surprising given that AKM person effects capture important variation not captured by observable person characteristics.

One message that emerges from this integrated approach is that interpreting the role of observable person characteristics is enhanced considerably by quantifying the contribution of within-industry vs between-industry effects. This permits quantifying the role of sorting and segregation of observable person characteristics across industries. Such quantification is important empirically. We find that most of the contribution of observable person characteristics including education and occupation reflects sorting and segregation between industries.

# References

Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. “High Wage Workers and High Wage Firms.” *Econometrica*, Vol. 67, No. 2, pp, 251-333.

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*, Vol. 31, No. 2, Pt. 2, April 2013, pp. S129-S172.

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

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*, Vol. 127, No. 5, pp. 2143-2185.

Card, David, Jorg Heining, and Patrick Kline. 2013. “Workplace Heterogeneity and the Rise of West German Wage Inequality.” *Quarterly Journal of Economics*, Vol. 128, No. 3, pp. 967-1015.

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.

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

Haltiwanger, John and James Spletzer. 2020. “Between Firm Changes in Earnings Inequality: The Dominant Role of Industry Effects.” NBER Working Paper No. 26786, February.

Haltiwanger, John, Henry R. Hyatt, and James R. Spletzer. 2022. “Industries, Mega Firms, and Increasing Inequality.” NBER Working Paper No. 29920, April.

Herkenhoff, Kyle, Jeremy Lise, Guido Menzio, and Gordon Phillips. 2018. “Production and Learning in Teams.” NBER Working Paper No. 25179, October.

Hoffman, Florian, David S. Lee, and Thomas Lemieux. 2020. “Growing Income Inequality in the United States and Other Advanced Economies.” *Journal of Economic Perspectives*, Vol. 34, No. 4, pp. 52-78.

Jarosch, Gregor, Ezra Oberfield, and Esteban Rossi-Hansberg. 2019. “Learning from Coworkers.” NBER Working Paper No. 25418, January.

Mellow, Wesley, and Hal Sider. 1983. “Accuracy of response in labor market surveys: Evidence and implications.” *Journal of Labor Economics* 1, no. 4:331-344.

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

Song, Jae, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till von Wachter. 2019. “Firming Up Inequality.” *Quarterly Journal of Economics*, Vol. 134, No. 1, pp. 1-50.

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, pp. 1-77.

# Appendices

## A Literature review: mathematical details

This Appendix provides formal details on recent empirical studies of earnings inequality that use matched employer-employee data. Much of the following is discussed in greater detail in HHS.

### A.1 Card, Heining, and Kline (2013)

Card, Heining, and Kline (2013) estimate the earnings model of Abowd, Kramarz, and Margolis (1999). They estimate this by separate intervals, which we will denote by $p$. The Card, Heining, and Kline (2013) earnings specification can be expressed as

$$y_t^{i,j,p} = X_t^{i,p}\beta^p + \theta^{i,p} + \psi^{j,p} + \varepsilon_t^{i,j,p}. \tag{A1}$$

In other words, Card, Heining, and Kline (2013) assume that earnings $y_t^{i,j,p}$ are the sum of an effect $\theta^{i,p}$ of worker $i$ in interval $p$, a firm effect $\psi^{j,p}$ when employed by employer $j$ during interval $p$, and a vector of time-varying observable characteristics $X_t^{i,p}$ for worker $i$ at time $t$, which have distinct marginal effects $\beta^p$ in each interval $p$.

Following Card, Heining, and Kline (2013), the variance of labor earnings can be decomposed into the following components in each period $p$:

$$\begin{aligned} \operatorname{var}(y_t^{i,j,p}) = & \underbrace{\operatorname{var}(X_t^{i,p}\beta) + \operatorname{var}(\theta^{i,p}) + 2\operatorname{cov}(X_t^{i,p}\beta, \theta^{i,p})}_{\text{person effects and observables}} + \underbrace{\operatorname{var}(\psi^j)}_{\substack{\text{total pay}\\ \text{premia}}} + \\ & \underbrace{2\operatorname{cov}(\theta^{i,p}, \psi^j) + 2\operatorname{cov}(X_{i,t,p}\beta, \psi^j)}_{\text{total sorting}} + \underbrace{\operatorname{var}(\varepsilon_t^{i,j,p})}_{\text{AKM residual}} . \end{aligned} \tag{A2}$$

Earnings inequality is a function of worker heterogeneity, firm heterogeneity, and the relationship between the two (sorting). Worker heterogeneity contributes to inequality through dispersion in the person effects $\operatorname{var}(\theta^{i,p})$, dispersion in the effects of observable characteristics $\operatorname{var}(X_t^{i,p}\beta)$, and the covariance between the two $2\operatorname{cov}(X_t^{i,p}\beta, \theta^{i,p})$. Firm heterogeneity contributes to inequality through dispersion in firm pay premia var($\psi^j$). Residual dispersion in earnings that is not attributed to worker or firm heterogeneity is var($\varepsilon_t^{i,j}$).

The relationship between worker heterogeneity and firm heterogeneity also contributes to inequality. The covariance between worker-driven pay differentials and firm-specific pay premia is called "sorting." Specifically, sorting is the covariance between firm pay premia and the effects of observable characteristics, $2\text{cov}(X_{i,t,p}\beta, \psi^j)$, as well as person effects $2\text{cov}(\theta^{i,p}, \psi^j)$. Sorting can enhance or mitigate total inequality. For example, if low-earning workers tend to work for high-paying firms, then sorting will lead to lower inequality. In contrast, if low-earnings workers tend to work for low-paying firms, then sorting will lead to higher inequality.

## A.2 Song et al. (2019)

Song et al. (2019) rely on the same earnings specification, as Card, Heining, and Kline (2013), equation (A1). They extend the Card, Heining, and Kline (2013) decomposition of the variance of labor earnings as expressed in equation (A2). Specifically, Song et al. (2019) distinguish worker heterogeneity that occurs within vs. between firms. The tendency for similar workers to be employed among each other is what they call "segregation."

Following Song et al. (2019), the variance of earnings can be written as (for ease of exposition, we now drop the time interval superscript $p$):

$$\begin{aligned}
\text{var}(y_t^{i,j}) = & \underbrace{\text{var}(\theta^i - \bar{\theta}^{j,k}) + \text{var}(X_t^i\beta - \bar{X}^j\beta) + 2\text{cov}(\theta^i - \bar{\theta}^j, X_t^i\beta - \bar{X}^j\beta)}_{\text{within-firm person effects and observables}} + \\
& \underbrace{\text{var}(\bar{\theta}^j) + \text{var}(\bar{X}^j\beta) + 2\text{cov}(\bar{\theta}^j, \bar{X}^j\beta)}_{\text{total segregation}} + \underbrace{\text{var}(\psi^j)}_{\substack{\text{total pay}\\ \text{premia}}} + \\
& \underbrace{2\text{cov}(\bar{\theta}^j, \psi^j) + 2\text{cov}(\bar{X}^j\beta, \psi^j)}_{\text{total sorting}} + \\
& \underbrace{2\text{cov}(\theta^i - \bar{\theta}^j, \varepsilon_t^{i,j}) + 2\text{cov}(X_t^i\beta - \bar{X}^j\beta, \varepsilon_t^{i,j}) + \text{var}(\varepsilon_t^{i,j})}_{\text{within-firm residual and covariances}}.
\end{aligned} \tag{A3}$$

The Song et al. (2019) extension of the standard AKM variance as decomposition is useful for analyzing within- vs. between-firm earnings dispersion.

Between-firm dispersion is the sum of the contributions of sorting, segregation, and firm premia. The firm premia term $\mathrm{var}(\psi^{j,k})$ is the standard expression used by Card, Heining, and Kline (2013) and many others.

Another componet of between-firm dispersion is sorting: the covariance between worker and firm effects. Sorting is $2\mathrm{cov}(\bar{\theta}^{j,k}, \psi^{j,k}) + 2\mathrm{cov}(\bar{X}^{j,k}\beta, \psi^{j,k})$. Sorting reflects the extent to which low- vs. high-earnings workers work for low- vs. high-paying firms. Note that Song et al.(2019) consider the covariance between the firm pay premium $\psi^{j,k}$ and firm-level averages $\bar{\theta}^{j,k}$ and $\bar{X}^{j,k}\beta$. In contrast, Card, Heining, and Kline (2013) use the more standard $\theta^{i,k}$ and $X_t^{i,k}\beta$. Note that these expressions of sorting yield equivalent results after the estimation of the AKM model.

The final component of between-firm dispersion is segregation, which reflects the concentration within firms of workers of the same type (captured by person effects). Segregation is $\mathrm{var}(\bar{\theta}^{j,k}) + \mathrm{var}(\bar{X}^{j,k}\beta) + 2\mathrm{cov}(\bar{\theta}^{j,k}, \bar{X}^{j,k}\beta)$. Note that segregation reflects the worker characteristics that are not predicted by the firm effect, i.e., that are not due to sorting.

The remaining dispersion is within-firm dispersion. Worker-level effects are given by $\mathrm{var}(\theta^i - \bar{\theta}^{j,k}) + \mathrm{var}(X_t^i\beta - \bar{X}^{j,k}\beta) + 2\mathrm{cov}(\theta^i - \bar{\theta}^{j,k}, X_t^i\beta - \bar{X}^{j,k}\beta)$. Finally, some earnings dispersion involves the residual $\varepsilon_t^{i,j,k}$, and the terms that enter into the variance decomposition are $2\mathrm{cov}(\theta^i - \bar{\theta}^{j,k}, \varepsilon_t^{i,j,k}) + 2\mathrm{cov}(X_t^i\beta - \bar{X}^{j,k}\beta, \varepsilon_t^{i,j,k}) + \mathrm{var}(\varepsilon_t^{i,j,k})$. Note that the covariance terms that include the residual are necessary for an exhaustive decomposition of the variance of earnings. The estimated residual from Equation (A1) is by construction orthogonal to worker effects, as well as the effects of worker characteristics. But the estimated residual can be correlated with the deviation of worker effects and the effects of observable characteristics from their respective firm-level averages because they are not explicitly controlled for in Equation (A1).

## A.3 Haltiwanger, Hyatt, and Spletzer (2022)

Haltiwanger, Hyatt, and Spletzer (2022) consider the role of industries indexed by $k$ in the evolution of inequality. They therefore introduce the superscript $k$ into the baseline AKM equation (A1). This yields:

$$y_t^{i,j,k,p} = X_t^{i,p}\beta^p + \theta^{i,p} + \psi^{j,k,p} + \varepsilon_t^{i,j,k,p}. \tag{A4}$$

HHS propose a tractable framework for the study of inequality in terms of effects that occur within- and between-industries. To explore cross-industry differences, HHS calculate industry-level averages. They define the average worker effect in industry $k$ in interval $p$ as $\bar{\theta}^k$, the average effect of observable characteristics as $\bar{X}^k\beta$, and the average firm effect as $\bar{\psi}^k$.

Given this notation, it is possible to measure how firm-level pay premia relate to within- vs. between-industry earnings dispersion. The HHS industry-enhanced variance decomposition is:

$$\begin{aligned}
\mathrm{var}(y_t^{i,j,k}) = & \underbrace{\mathrm{var}(\theta^i - \bar{\theta}^{j,k}) + \mathrm{var}(X_t^i\beta - \bar{X}^{j,k}\beta) + 2\mathrm{cov}(\theta^i - \bar{\theta}^{j,k}, X_t^i\beta - \bar{X}^{j,k}\beta)}_{\text{within-firm person effect and observables}} + \\
& \underbrace{\mathrm{var}(\bar{\theta}^k) + \mathrm{var}(\bar{X}^k\beta) + 2\mathrm{cov}(\bar{\theta}^k, \bar{X}^k\beta)}_{\text{between-industry segregation}} + \\
& \underbrace{\mathrm{var}(\bar{\theta}^{j,k} - \bar{\theta}^k) + \mathrm{var}(\bar{X}^{j,k}\beta - \bar{X}^k\beta) + 2\mathrm{cov}[(\bar{\theta}^{j,k} - \bar{\theta}^k), (\bar{X}^{j,k}\beta - \bar{X}^k\beta)] +}_{\text{within-industry, between-firm segregation}} \\
& \underbrace{\mathrm{var}(\bar{\psi}^k)}_{\substack{\text{between-industry} \\ \text{pay premia}}} + \underbrace{\mathrm{var}(\psi^{j,k} - \bar{\psi}^k)}_{\substack{\text{within-industry,} \\ \text{between-firm} \\ \text{pay premia}}} + \underbrace{2\mathrm{cov}(\bar{\theta}^k, \bar{\psi}^k) + 2\mathrm{cov}(\bar{\psi}^k, \bar{X}^k\beta)}_{\text{between-industry sorting}} + \\
& \underbrace{2\mathrm{cov}[(\bar{\theta}^{j,k} - \bar{\theta}^k), (\psi^{j,k} - \bar{\psi}^k)] + 2\mathrm{cov}[(\psi^{j,k} - \bar{\psi}^k), (\bar{X}^{j,k}\beta - \bar{X}^k\beta)]}_{\text{within-industry, between-firm sorting}} + \\
& \underbrace{2\mathrm{cov}(\theta^i - \bar{\theta}^{j,k}, \varepsilon_t^{i,j,k}) + 2\mathrm{cov}(X_t^i\beta - \bar{X}^{j,k}\beta, \varepsilon_t^{i,j,k}) + \mathrm{var}(\varepsilon_t^{i,j,k})}_{\text{within-firm residual and covariances}}
\end{aligned} \tag{A5}$$

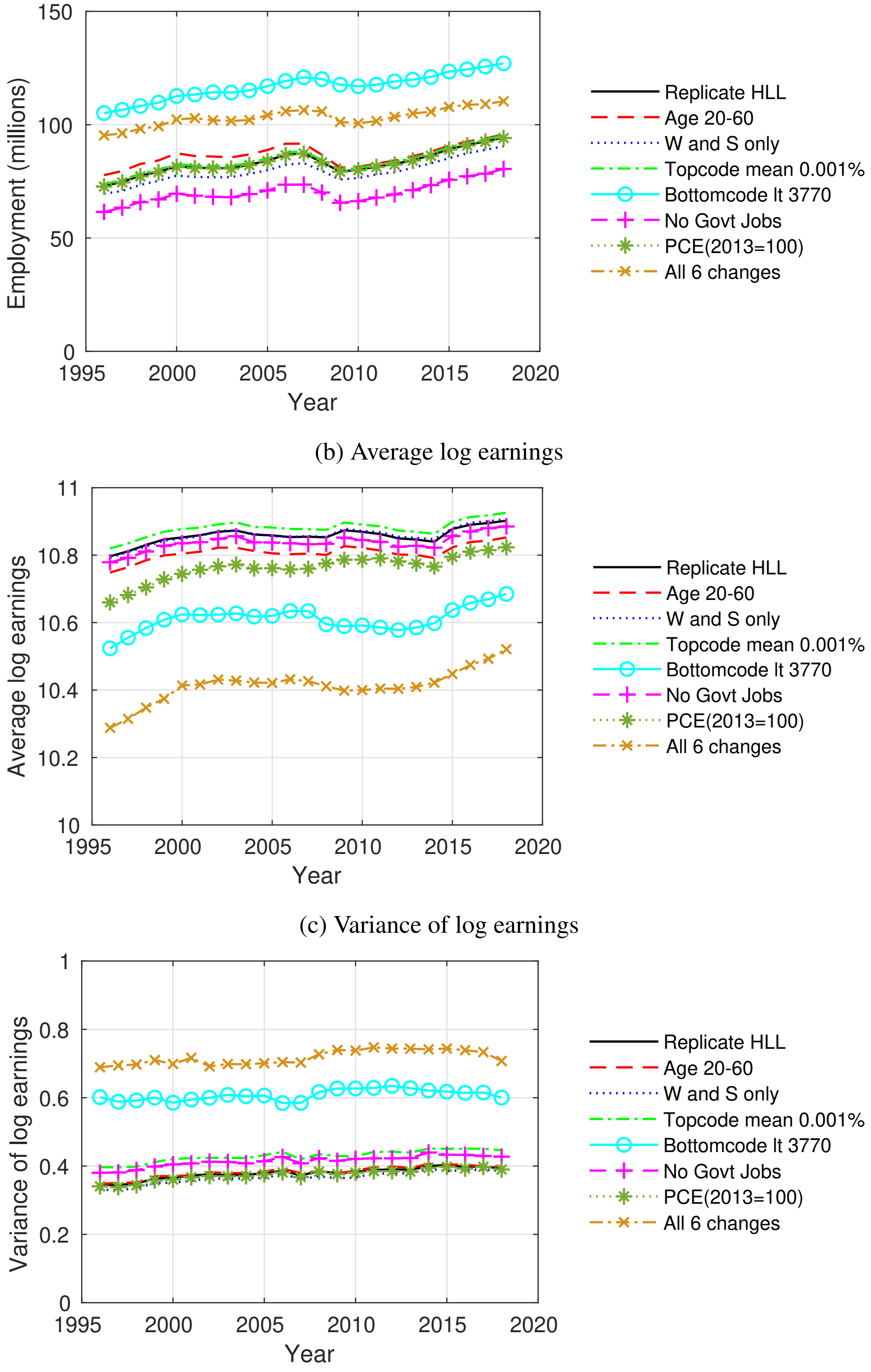
The within-firm dispersion in HHS is exactly as in Song et al. (2019). The differences are in the between-firm components. $\mathrm{var}(\psi^{j,k}) = \mathrm{var}(\bar{\psi}^k) + \mathrm{var}(\psi^{j,k} - \bar{\psi}^k)$, where $\bar{\psi}^k$ reflects the between-industry dispersion in average firm effects, i.e. industry-level pay premia. The remaining term $\mathrm{var}(\psi^{j,k} - \bar{\psi}^k)$ captures the within-industry dispersion of firm-level pay premia. In addition to pay premia, we can distinguish between the within- vs. between-industry components of sorting and segregation.

Between-industry sorting is defined as $2\mathrm{cov}(\bar{\theta}^k, \bar{\psi}^k) + 2\mathrm{cov}(\bar{\psi}^k, \bar{X}^k\beta)$. It therefore reflects the extent to which highly-paid workers are employed in industries with a high pay premium. This is distinct from within-industry sorting $2\mathrm{cov}[(\bar{\theta}^{j,k} - \bar{\theta}^k), (\theta^{j,k} - \bar{\theta}^k)] + 2\mathrm{cov}[(\theta^{j,k} - \bar{\theta}^k), (\bar{X}^{j,k}\beta - \bar{X}^k\beta)]$ This is the component of sorting where relatively highly-paid workers tend to work at high-paying firms, apart from industry-level differences. The between-industry component reflects these industry level differences. The within-industry component reflects the extent to which relatively low- vs. high-paid workers work for relatively low-vs. high-paying firms in those industries.

Segregation also can be decomposed into its within- vs. between-industry components. Between-industry segregation is given by industry-level average worker effects. Formally, this is expressed as $\text{var}(\bar{\theta}^k) + \text{var}(\bar{X}^k\beta) + 2\text{cov}(\bar{\theta}^k, \bar{X}^k\beta\ )$. This is the extent to which low- vs. highly-paid workers tend to work with each other. Segregation that occurs within industries is $\text{var}(\bar{\theta}^{j,k} - \bar{\theta}^k) + \text{var}(\bar{X}^{j,k}\beta - \bar{X}^k\beta) + 2\text{cov}[(\bar{\theta}^{j,k} - \bar{\theta}^k), (\bar{X}^{j,k}\beta - \bar{X}^k\beta)]$.

# B Supplementary tables and figures

Figure A1: Creating the common coded CPS-ASEC

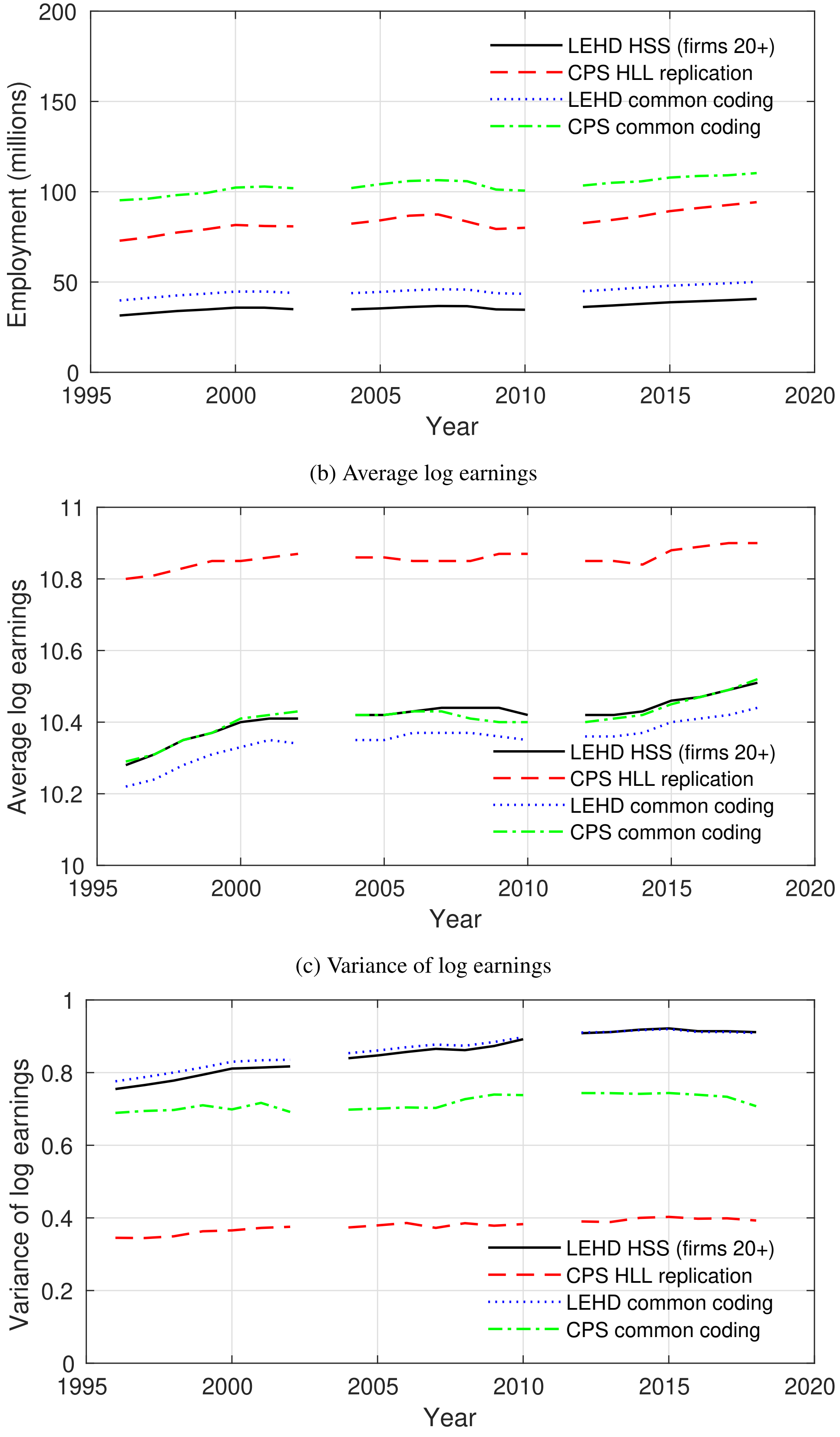


(a) Employment

(b) Average log earnings

(c) Variance of log earnings

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

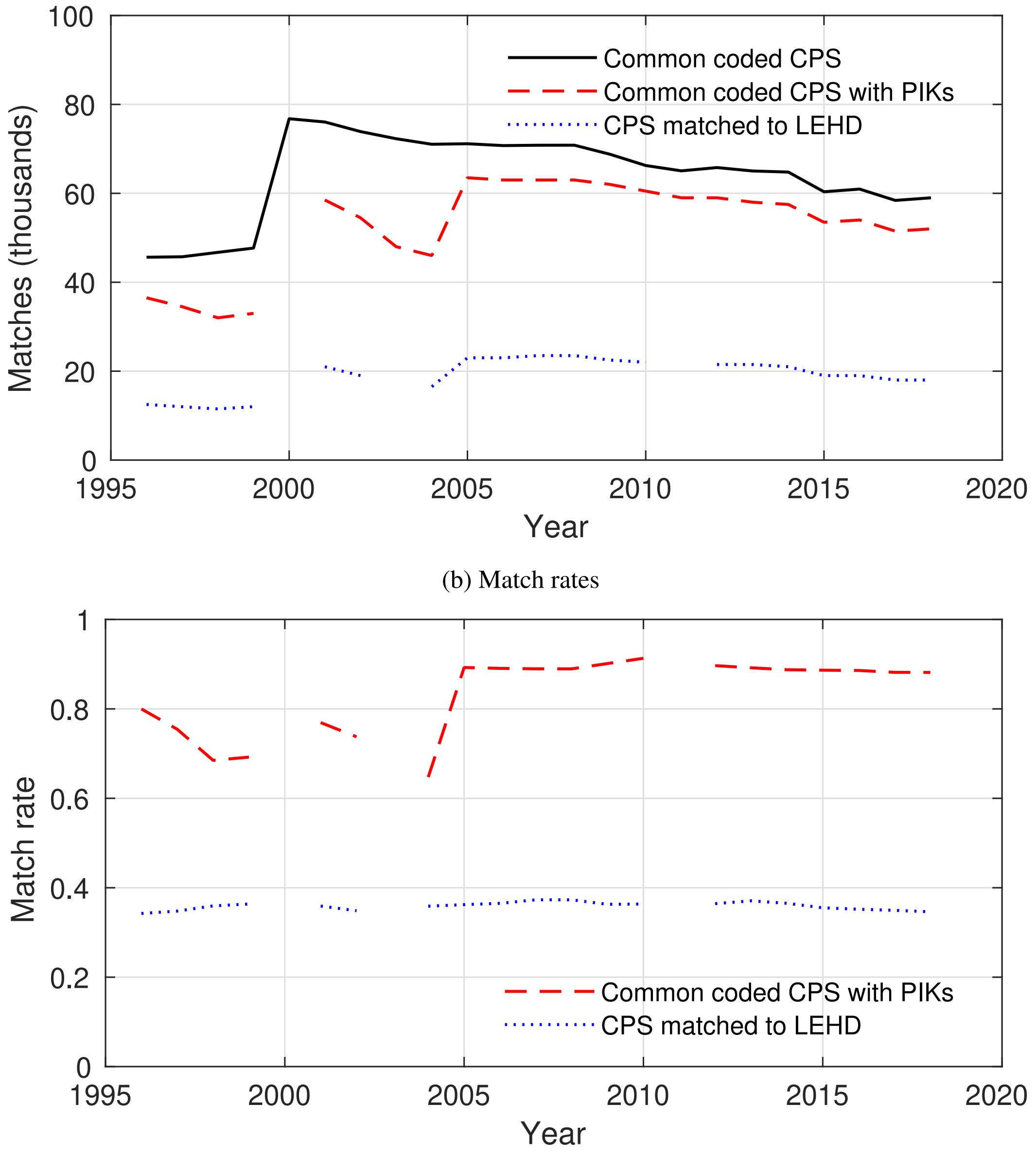


(a) Employment

(b) Average log earnings

(c) Variance of log earnings

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



(a) Observations

(b) Match rates

Figure A4: PDFs of linked CPS-ASEC and LEHD earnings

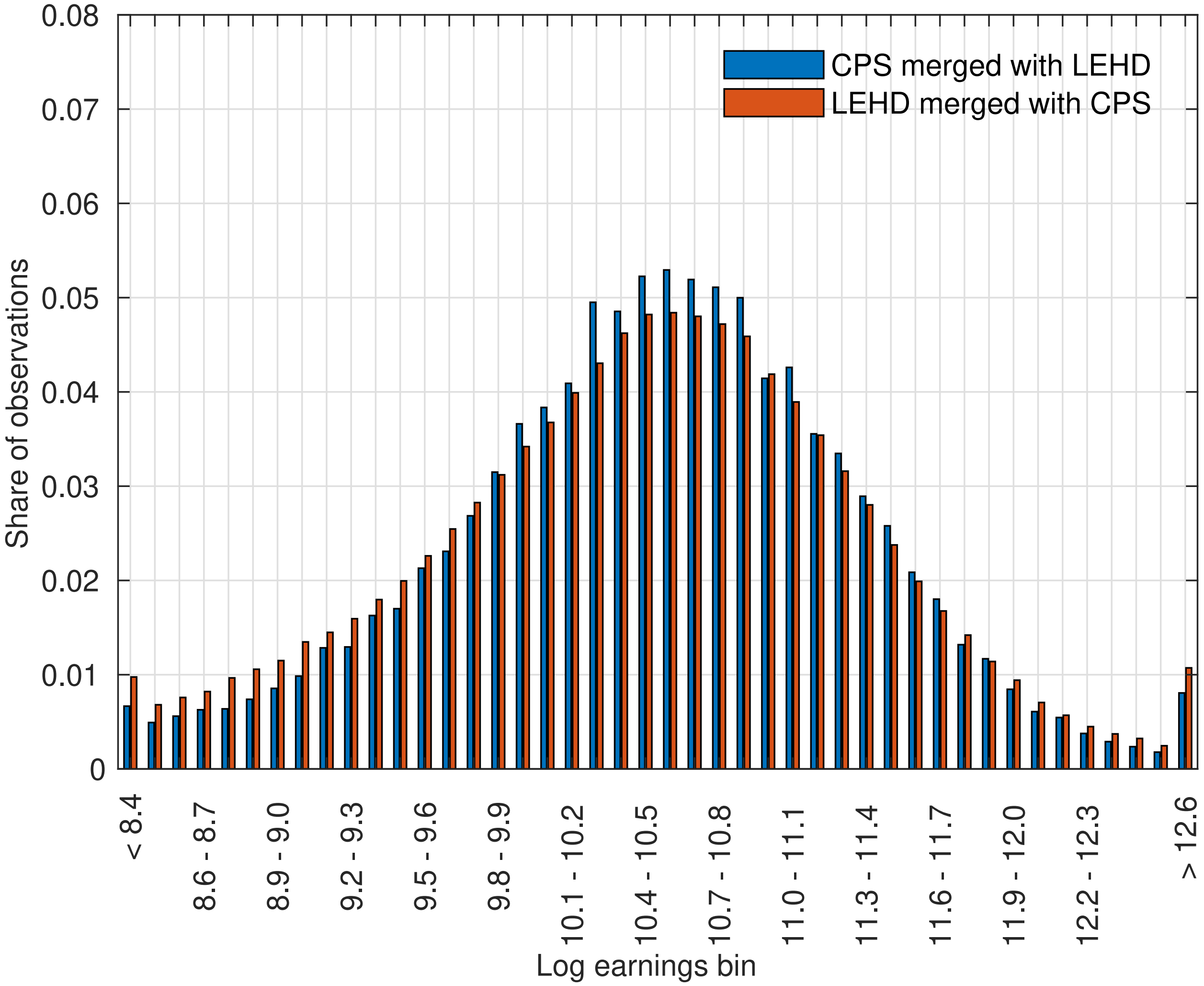


Figure A5: PDF of earnings difference, linked CPS-ASEC and LEHD data

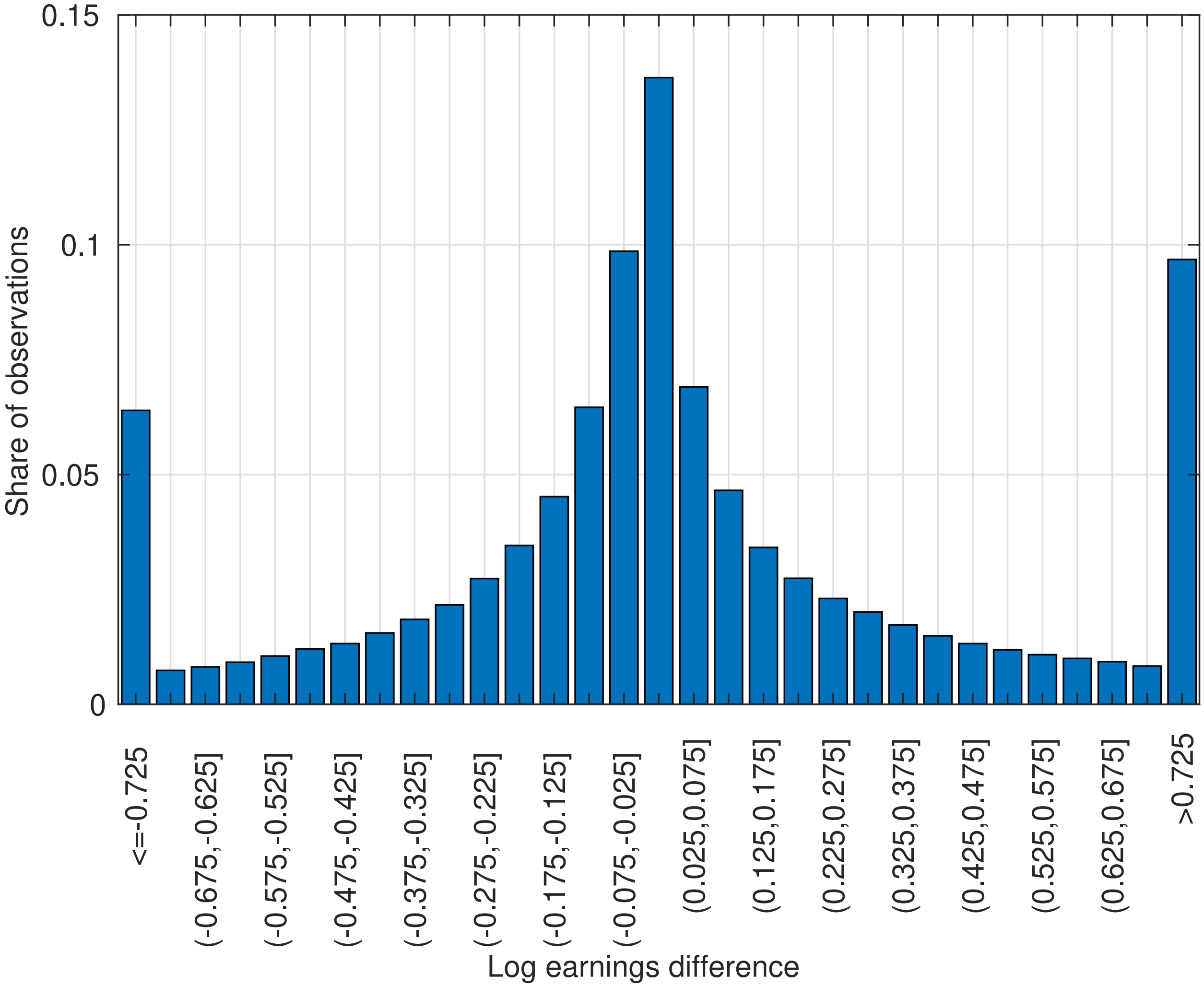


Figure A6: Trouble in the tails in CPS earnings distributions
CPS Earnings by LEHD Earnings

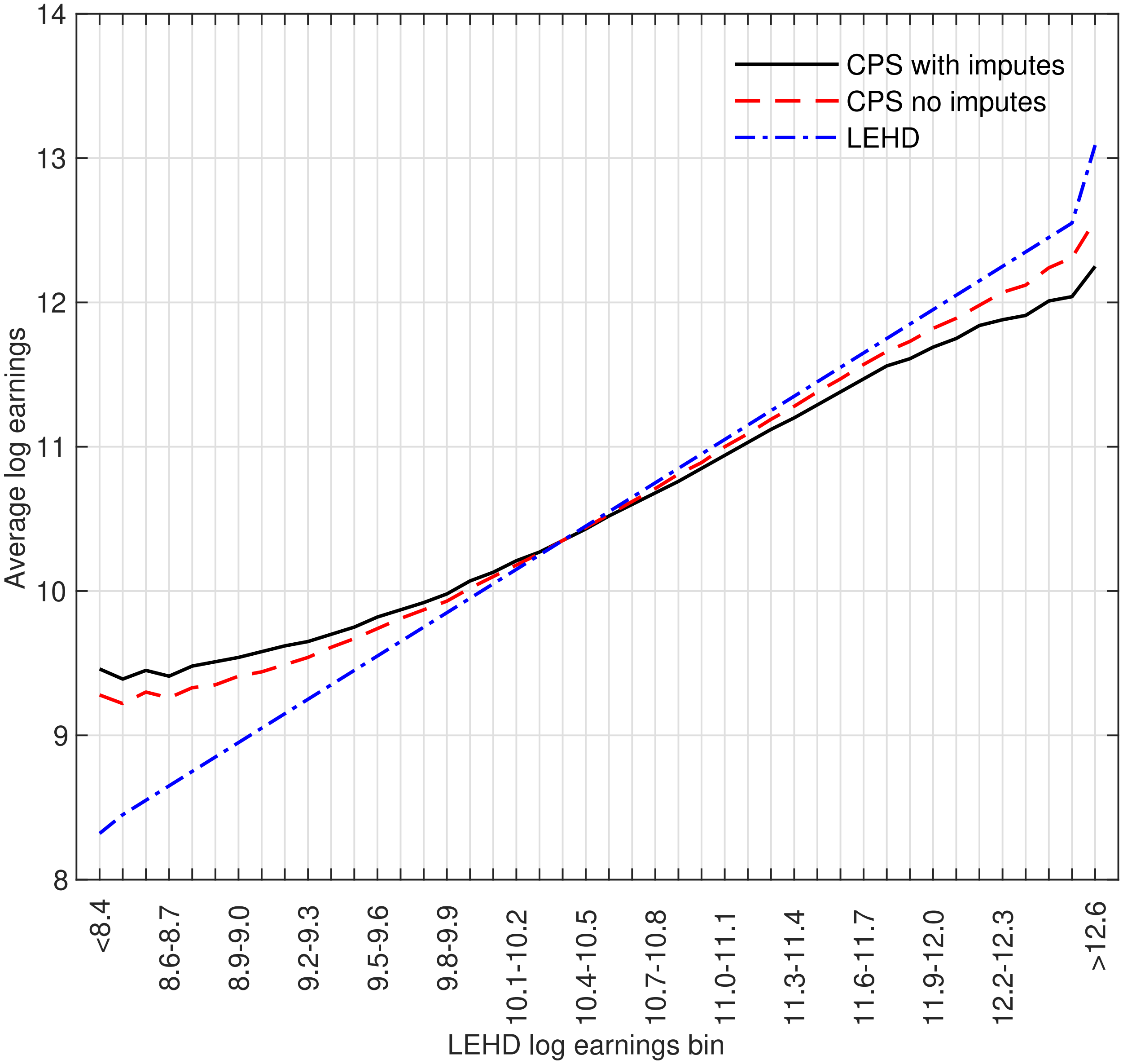


Table A1: First and second panels replicate HLL Figure 4, third panel is HLL with pooled genders, Fourth panel is pooled genders with labor earnings rather than total income (Table 1 of this paper)

| | 1975-<br>1979 | 1980-<br>1984 | 1985-<br>1989 | 1990-<br>1994 | 1995-<br>1999 | 2000-<br>2004 | 2005-<br>2009 | 2010-<br>2014 | 2015-<br>2018 | Growth<br>1975-79 to<br>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*Educ | 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*Educ | 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*Educ | 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 1 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*Educ | 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 on the *Journal of Economic Perspectives* website.

Table A2: Industry contributions to between-industry growth Common Coded CPS, 50 states and 18 states

| Industry | CPS<br>50 states<br>+ DC | CPS<br>18 states | Difference<br>18 minus 50<br>States | Changing<br>employment<br>shares | Changing<br>earnings<br>differentials |
|---|---|---|---|---|---|
| Agriculture | -0.0008 | -0.0015 | -0.0007 | -0.0003 | -0.0005 |
| Mining | 0.0015 | 0.0017 | 0.0001 | 0.0004 | -0.0003 |
| Construction | 0.0000 | -0.0001 | -0.0001 | 0.0000 | -0.0001 |
| Manufacturing | -0.0021 | -0.0013 | 0.0008 | -0.0002 | 0.0010 |
| Wholesale Trade | -0.0005 | -0.0007 | -0.0002 | 0.0000 | -0.0002 |
| **Retail Trade** | **0.0035** | **0.0054** | **0.0019** | **0.0002** | **0.0017** |
| Transport & Warehous | -0.0007 | -0.0007 | 0.0001 | 0.0000 | 0.0001 |
| Utilities | -0.0001 | -0.0001 | 0.0000 | 0.0001 | 0.0000 |
| **Information** | **0.0054** | **0.0082** | **0.0028** | **-0.0001** | **0.0029** |
| Finance & Insurance | 0.0050 | 0.0051 | 0.0001 | -0.0002 | 0.0003 |
| Real Estate & Rental | 0.0002 | 0.0002 | 0.0000 | 0.0000 | 0.0000 |
| Prof & Bus Services | 0.0010 | 0.0008 | -0.0002 | -0.0001 | -0.0002 |
| Educational Services | -0.0005 | -0.0003 | 0.0003 | 0.0001 | 0.0002 |
| Healthcare & Soc Ass | 0.0004 | 0.0006 | 0.0002 | 0.0000 | 0.0002 |
| Arts, Ent, & Rec | 0.0001 | 0.0003 | 0.0002 | 0.0001 | 0.0001 |
| Accom & Food Serv | 0.0082 | 0.0084 | 0.0002 | 0.0002 | 0.0000 |
| Other Services | -0.0005 | -0.0006 | -0.0001 | 0.0000 | 0.0000 |
| Unknown (2nd job) | -0.0097 | -0.0106 | -0.0009 | -0.0010 | 0.0001 |

*Notes*: See text of section 7 for the methodology used to create these statistics. Rows are in bold when columns 1 and 2 differ by more than .001.