The Role of Worker, Firm, Industry and Occupation Effects in Increasing Inequality: Reconciling Evidence from Survey and Administrative Data

John Haltiwanger, University of Maryland

Henry Hyatt, U.S. Census Bureau

James R. Spletzer, U.S. Census Bureau

October 25, 2021

Extended abstract
Preliminary and Incomplete

This document is an extended abstract prepared for the NBER conference "Wage Dynamics in the 21st Century." Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed (CBDRB-FY21-046, CBDRB-FY21-155, and CBDRB-FY21-323). John Haltiwanger was also a Schedule A employee of the U.S. Census Bureau at the time of the writing of this abstract.

1. Introduction

What drives increasing earnings inequality? This question has become increasingly important in recent decades as earnings dispersion has increased. An open question concerns the extent to which increasing inequality is driven by low-paying vs. high-paying jobs. For example, employees of restaurants and general merchandise stores are commonly understood to be among the lowest-paid workers. If the earnings of these workers declines, or if their number increases, inequality will naturally increase. Similarly, information service workers and financial investment workers are among the highest-paid and increases in their earnings or their number will lead to increases in 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 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 (2021), 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 two strands of the inequality literature do not provide a consistent answer regarding the contribution of 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."

Our work in this paper is motivated by the question of why administrative records data indicate that industry-level differences drive increasing inequality, while studies that rely on the CPS suggest a small and opposite effect. To assess this, we rely on 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. As inequality measures are sensitive to employment among the lowest-paid jobs, this sample selection criterion matters. When we apply Song et al.'s (2019) and HHS's sample inclusion criteria on the CPS, this results in including more low-earning workers in the CPS, which furthermore results in finding a greater role for industry-level differences in increasing inequality. Industry measurement also matters. Studies that rely on the CPS usually aggregate industry-level information to a small number of high-level aggregates. Using the detailed industry data present in the CPS shows more of a role for industry-level differences in the evolution of income inequality. And thirdly (in ongoing work not yet completed), we believe that replacing the Standard Industrial Classification (SIC) measures of industry often used in CPS analysis with North American Industry Classification System (NAICS) measures of industry will also increase the role of industry in accounting for increasing earnings inequality.

But 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 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.

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. A full set of estimates have been calculated from this linked CPS-LEHD data, but as of October 2021 they have not yet been cleared through Census Bureau's disclosure. Our estimates assess the roles of (1) sample differences such as including workers with low earnings, (2) the use of SIC codes in the CPS versus NAICS codes in the LEHD, (3) the concordance of CPS and LEHD industry codes for a given individual, (4)

different methodologies for quantifying the role of industry, and (5) exploring the roles of sorting, segregation, and pay premia in the matched CPS-LEHD data.

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 use these data to quantify the relative importance of available data from the CPS and the LEHD for determining the sources of increasing earnings inequality. We seek to determine what the CPS is missing in not having worker and firm fixed effects, along with determining what the LEHD is missing in not having occupation and education.[1]

In the remainder of this extended abstract, in Section 2 we provide an overview 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 from Haltiwanger, Hyatt, and Spletzer (2021) on the evolution of these. In Section 3, 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 4, we list the possible reasons why the contribution of industry appears to differ in studies using administrative records versus studies using the CPS. While our results from the linked CPS-LEHD dataset have not yet been released, we present some tabulations using public-use CPS microdata alone that allow us to begin answering several of the measurement issues involved. In Section 5, we provide an overview of the linked CPS-LEHD data and estimates we have calculated but not yet released from this data.

2. Matched employer-employee administrative records evidence on sorting, segregation, and pay premia

The recent literature on inequality using administrative data takes AKM as its starting point. In this framework, the earnings $y_{i,j,t}$ of worker i at firm j at time t is a linear function of a worker effect θ_i, a firm effect ψ_j, and time-varying observable characteristics $X_{i,t}$ (usually including age and time effects) that have marginal effects according to the vector β:

$$y_{i,j,t} = \theta_i + \psi_j + X_{i,t}\beta + \varepsilon_{i,j,t}$$

Estimation of this model on matched employer-employee data allows researchers to evaluate the extent to which earnings inequality is determined by observable characteristics, worker effects, firm effects, and sorting. Specifically, it is possible to express the variance of earnings as

$$Var(y_{i,j,t}) = Var(\theta_i) + Var(\psi_j) + Var(X_{i,t}\beta) + 2Cov(\theta_i, \psi_j) + 2Cov(\theta_i, X_{i,t}\beta) + 2Cov(\psi_j, X_{i,t}\beta) + Var(\varepsilon_{i,j,t})$$

[1] The LEHD does have a measure of education but it is imputed for 88 percent of observations (Vilhuber 2018, page 5-1).

The contribution of workers to earnings dispersion is usually thought of as the sum of the variance of the worker effects, that of observable characteristics, and their covariance: $Var(\theta_i) + Var(X_{i,t}\beta) + 2Cov(\theta_i, X_{i,t}\beta)$. The contribution of firms is given by the dispersion of the firm pay premia $Var(\psi_j)$. Sorting is the extent to which firms with low- versus high-effects employ workers with low- versus high-effects, as well as low- versus high-observables: $2Cov(\theta_i, \psi_j) + 2Cov(\psi_j, X_{i,t}\beta)$.

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.

Using this framework, Song et al. (2019) considered the evolution of earnings dispersion in the U.S. These authors were motivated by the fact that earnings inequality was driven by the between-firm component. A key lesson from Song et al. 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 $var(\theta^i)$ 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($\bar{\theta}^j$), relative to worker-driven dispersion, denoted by $var\left(\theta^i - \bar{\theta}^j\right)$.

In the Song et al. (2019) variance decomposition, firm segregation is defined as $var\left(\bar{\theta}^j + \overline{X\beta}^j\right)$.[2] Firm segregation measures differences among firms in terms of the workers that they employ. 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 θ^i, 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 inequality. Workers at some firms earn more than others. This may be due either to firm pay premia, or to differences in the workers that they employ. 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

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

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

We reproduce some of the main results of HHS in our Table 1. During the time period 1996-2002 to 2012-2018, the variance of annual log earnings increased from 0.7938 to 0.9152. 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.

3. What do household surveys (e.g., the CPS) tell us about inequality?

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 jumping off point for our own analysis.

Figure 3 of HLL 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 emphasize 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 illustrated in Table 1 from 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,

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

and space effects over and above their baseline analysis of worker characteristics. Their Figure 4 shows that increasing occupation wage differentials play an important but supporting role compared to the baseline contribution from worker characteristics only. The marginal contribution of inter-industry wage differentials is actually negative after controlling for the baseline worker and occupation effects.[4] Location effects contribute little to rising dispersion.

As we discuss below, HLL's marginal approach limits the role of occupation, industry, and location since all of the covariances with baseline worker characteristics are attributed to the latter. 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 effects emphasized in Song et al. (2018) and HHS.

4. Why do the LEHD and the CPS differ on industry?

The HHS results on the contributions of particular industries appear counter to the conclusions of other recent studies including those of HLL and Stansbury and Summers (2020). In this section, we rely on the CPS ASEC replication files made available by HLL.[5]

4a. Industry detail

We quantify how industry detail affects conclusions about how inequality evolves within- versus between-industry. The variance decomposition used here is based on the equation

$$var(Y_{ik}) = var(Y_{ik} - \overline{Y}_k) + var(\overline{Y}_k)$$

where i denotes the individual and k denotes the industry. The first term on the right is the within industry variance, and the second term on the right is the between industry variance.

Our findings are shown in Table 2. Columns 1 and 2 use the HLL CPS data (we use HLL's measure of labor earnings rather than labor earnings plus capital income). Column 1 uses the same 12 industries that HLL use in their Figure 4, and column 2 uses all industries available in the CPS microdata. For reference, we include the analogous tabulations from HHS in column 3.

Before looking at the decomposition results in Table 2, it is useful to note that the variance of earnings in the LEHD is more than double that in HLL's CPS. The CPS variance of earnings is .3604 in the 1996-2002 interval, while the LEHD shows .7938. In the more recent 2012-2018 interval, the CPS shows .3966 while the LEHD shows .9152. There is also a substantial discrepancy in the growth of earnings variance: .0361 in the CPS versus .1214 in the LEHD.

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

[5] We are very grateful to Hoffman, Lee, and Lemieux for making their replication data and programs available on the website of the *Journal of Economic Perspectives*, as well as for providing detailed answers to our questions about their replication files.

Table 2 shows that industry aggregation matters for the measurement of inequality. The HLL replication file with 12 aggregated industries indicates that 13.2% of the growth of inequality is between industries. When we use all available CPS industries (which changes over time – from 236 in the earlier interval to 221 in the later), we find that 31.0% of the growth occurred between industries. In the LEHD data used by HHS, 61.9% of variance growth is between 301 4-digit NAICS industries.

One other detail warrants mentioning here. The 200+ industry codes from the CPS ASEC used in Table 2 are based on the SIC. We know that the switch from the SIC to the NAICS added quite a bit of detail, particularly in the services sector. It is possible that the some of the difference between 31.0% in the CPS and 61.9% in the LEHD could be accounted for by this SIC-NAICS difference. It is on our "to-do" list to create consistent 1996-2018 NAICS based industry codes in the CPS microdata and re-do the CPS within- and between-industry variance decomposition with this new measure of industry.

4b. Sample differences

The large difference in the level of cross-sectional variances when comparing the CPS and the LEHD suggests that sample differences play a large role. We want to compare the CPS ASEC and the LEHD on an apples-to-apples basis. Seven major differences between the HLL CPS data and the HHS LEHD data are listed in the following table, along with the common coding we will use for comparison:

	HLL CPS	Common Coding	HHS LEHD
Earnings	Wage & Salary + Self Employment + Farm	Wage & Salary	Wage & Salary
Age	25-65	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 $\geq$ 49 & usual hours $\geq$ 40 & real hourly wage $\geq$ \$4 $\rightarrow$ annual real earnings $\geq$ \$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 restriction	None	None	Firm Size $\geq$ 20

A complete description of each of these seven differences will be in an appendix to our forthcoming paper. We have created the common-coded CPS ASEC data. The pdfs of labor earnings in the HLL replication and in the common coded CPS are given in Figure 1 – we see that the earnings distribution in the common coded CPS has added substantial amounts of mass to the left of the HLL CPS distribution.

We present variance decomposition results from the common coded CPS in Table 3. Comparing the common coded CPS (in column 2 of Table 3) to the HLL CPS (in column 1 of Table 3), we see that the variance of earnings has increased substantially, from .3604 to .7035 in the 1996-2002 interval and from .3966 to .7373 in the 2012-2018 interval. This large increase is mostly due to changing the bottom code from $7840 to $3770. Changing from age 25-65 to age 20-60, changing the deflator, and getting rid of the self-employed do essentially nothing to the variance of CPS HLL earnings. Getting rid of longest jobs government jobs and changing the topcode results in a small increase in the variance level. We also note that the variance growth in the common coded sample did not increase relative to the HLL sample (it fell from .0361 to .0338).

Recoding the CPS ASEC to be similar to Song et al. (2019) and HHS, with an emphasis on the inclusion of low-earnings workers, increases the industry contribution to inequality. In the CPS with common coding, and with using full industry detail, we find that 54.2% of variance growth is between industries (column 2 of Table 3). This is substantially higher than the 31.0% in the HLL CPS and much closer to the 61.9% obtained from the HHS LEHD data.

4c. Substantial residual in the CPS

We now describe the evolution of the variance of earnings in the CPS. We start with a regression motivated by Hoffman, Lee, and Lemieux (2021):

$$LN(W_{it}) = Age*Educ_{it}\beta 1 + Occupation_{it}\beta 2 + Industry_{it}\beta 3 + \varepsilon_{it}$$

We can characterize the contribution of industry to rising variance as follows:

$$\begin{aligned} Var[LN(W_{it})] = {} & var[Age*Educ_{it}\beta 1] + var[Occupation_{it}\beta 2] + var[Industry_{it}\beta 3] \\ & +2cov[Age*Educ_{it}\beta 1, Occupation_{it}\beta 2] + 2cov[Age*Educ_{it}\beta 1, Industry_{it}\beta 3] \\ & +2cov[Occupation_{it}\beta 2, Industry_{it}\beta 3] + var[\varepsilon_{it}] \end{aligned}$$

This variance decomposition is presented in Table 4 (using 18 industry sectors from the CPS SIC-based industry variable ind90ly). First note that the explanatory variables only explain about 33 to 35 percent of cross-sectional CPS earnings variation. This contrasts with the AKM methodology which explains about 85 percent of cross-sectional LEHD earnings variation.

Looking at the effect of industry in Table 4 and looking at variance growth in the bottom third of Table 4, we see that var(industry) has a negative effect: -5.7% in the HLL CPS and -30.8% in the common coded CPS. However, the covariances involving industry are positive and large in magnitude. These covariances are not mentioned in most CPS studies and are the beginning of a crosswalk to the sorting and segregation stories in the administrative records literature.[6]

[6] Stansbury and Summers (2020) note the potential confounding influences of demographic characteristics and occupation. Specifically, when describing the falling dispersion of the industry wage premia, they note "our result does not conflict with the result of Haltiwanger and Spletzer (2020), who find that the dispersion of average log earnings across industries has risen over 1997-2013: this pattern also exists in our raw CPS data, but is reversed once occupation and individual characteristics are controlled for."

4d. High quality industry measures in the LEHD

One final explanation for the smaller industry effect in the CPS relative to the LEHD is the quality of the industry data. The LEHD industry codes are derived from the QCEW, which the Bureau of Labor Statistics (BLS) uses as the establishment-level universe for all sampling and publications. The BLS has strong incentives to assign and maintain detailed and accurate industry codes at the establishment-level. On the other hand, CPS industry codes are assigned based on a respondents answer to the questions "47a) For whom did you work at your longest job last year?" and "47b) What kind of business or industry is this? For example: TV and radio manufacturing, retail shoe store, farm." The limitations of the self-reported industry information in the CPS are well-known (see, e.g., Mellow and Sider (1983) and Dey, Houseman and Polivka (2010)).

The way to analyze the effect of higher quality industry codes in the LEHD versus the CPS is to link the CPS and LEHD microdata and directly substitute one industry code for another in a variance decomposition. We have done this but have not cleared the results through disclosure. In the following section, we describe the linked CPS-LEHD data that will allow for such an analysis.

5. Matched CPS-LEHD data

Matching the CPS and the LEHD microdata will allow for an analysis of measurement error as described in the previous section. The linked CPS-LEHD data will also allow for a determination of the relative explanatory power of various sets of variables when explaining increasing earnings inequality.

5a. Creating the linked CPS-LEHD data

We have linked the CPS ASEC and LEHD microdata. The following is a brief description of the linkage algorithm; a more detailed description will be available when our results are cleared by the Census Bureau's Disclosure Review Board.

The Census Bureau has Protected Identification Keys (PIKs) in the CPS ASEC for all survey years 1996-2020. Since the CPS ASEC asks about earnings in the prior calendar year, PIKs have been attached for earnings reference years 1995-2019. Not every record in the CPS ASEC has a PIK attached. As noted in the appendix of Bollinger et al. (2019):

> "The Census Bureau changed its consent protocol to link respondents to administrative data beginning with the 2006 ASEC. Prior to this CPS collected respondent SSNs and an affirmative agreement allowing a link to administrative data; i.e., an "opt-in" consent option. Beginning with the 2006 ASEC, respondents not wanting to be linked to administrative data had to notify the Census Bureau through the survey field representative, website or use a mail-in response in order to "opt-out". This opt-out rate is a very small 0.5 percent of the ASEC sample. If the respondent doesn't opt out, they are assigned an SSN using the Person

Validation System. Under the prior "opt-in" consent option in the 2005 ASEC, the linkage rate among earners was only 61 percent, as compared to 86 percent for our full 2006-2011 ASEC sample."

This difference across years in the probability of having 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 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 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 PIK'ed CPS data with the HHS LEHD data, where both have the common coding system. We only keep observations where an individual is in both the CPS-ASEC and in the LEHD. We run another set of year-specific logistic regressions where the dependent variable is "1 if a CPS 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 explanatory variables in these regressions since the CPS is national but the HHS LEHD is 18 states. We output the predicted values for each person-year observation, and then adjust the (already adjusted) CPS ASEC weights in the matched sample by dividing by the predicted value.

One important note is that the linked CPS-LEHD microdata contains the AKM components $\{\theta_i, \psi_j, X_{i,t}\beta\}$ that were estimated on the full HHS LEHD data with 938.9 million person-year observations.

5b. Analysis of the linked CPS-LEHD data

We have done extensive analysis of this linked CPS-LEHD data, and are planning to do much more. No results have yet been put through the Census Bureau's disclosure process.

- We have done a series of quality checks on the linked CPS-LEHD data to show that analysis of this data replicates the findings from the larger CPS data and the much larger HHS LEHD data.
- We have analyzed CPS-LEHD differences in both earnings and industry.
- We have estimated within- and between- industry variance decompositions, replacing CPS industry with LEHD industry to see whether the between-industry component of variance growth increases with the higher-quality LEHD industry measure
- We have estimated the HLL earnings regression with age*education, occupation, and industry on the right-hand side, and we then ask how the R-squared of this regression, the estimated coefficients, and the variance decomposition change when the AKM components $\{\theta_i, \psi_j, X_{i,t}\beta\}$ are added to this regression.
- Similar to the above, we ask how the R-squared and the variance decomposition change when CPS demographics and occupation are added to an AKM earnings regression. This analysis provides a specification check for AKM permitting us to evaluate whether

worker and firm effects already capture the variation associated with key variables such as education and occupation.

- We regress each of the AKM components $\{\theta_i, \psi_j, X_{i,t}\beta\}$ on CPS demographics and occupation to begin to understand what the AKM components measure in terms of CPS observable characteristics.

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.

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.

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 Hyatt, and James Spletzer. 2021. "Industries, Mega Firms, and Increasing Inequality." Unpublished paper, presented at the May 2021 NBER conference "Wage Dynamics in the 21st Century," revised October 2021.

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. 4:331–44.

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.

Vilhuber, Lars. 2018. "LEHD Infrastructure S2014 files in the FSRDC." Census Bureau Working Paper #CES 18-27R, September 2018.

Table 1: Industry Enhanced Variance Decomposition (Pooled Males and Females)

	1996-2002	2012-2018	Growth 1996-2002 to 2012-2018
Total $var(y^{i,j,k})$	.7938	.9152	.1214
Between firm within industry $var(\bar{y}^{j,k} - \bar{y}^{k})$	14.0%	15.3%	23.1%
Firm segregation	6.3%	7.0%	11.6%
Firm pay premium	3.1%	3.1%	2.9%
Firm sorting	4.6%	5.1%	8.6%
Between industry $var(\bar{y}^{k})$	21.4%	26.8%	61.9%
Industry segregation	7.4%	9.7%	25.2%
Industry pay premium	4.2%	4.8%	8.7%
Industry sorting	9.9%	12.3%	28.0%
Within firm $var(y^{i,j,k} - \bar{y}^{j,k})$	64.6%	58.0%	14.9%
Person effect	48.2%	44.3%	18.8%
Residual	16.4%	13.7%	-3.9%

This table is taken directly from Table 4 of Haltiwanger, Hyatt, and Spletzer (2021). We include this in the extended abstract for expositional convenience. The full draft of the current paper will provide a summary of results from HHS (2021) instead of reproducing a table from that paper.

Notes: (1) Persons with annual real earnings >$3770 in EINs with 20 or more employees.

(2) $y^{i,j,k,t} = \theta^{i} + \Psi^{j} + X_{i,t}\beta + \varepsilon^{i,j,k,t}$

(3) Firm segregation defined as $var(\bar{\theta}^{j,k} - \bar{\theta}^{k}) + var(\overline{X\beta}^{j,k} - \overline{X\beta}^{k}) + 2cov[(\bar{\theta}^{j,k} - \bar{\theta}^{k}), (\overline{X\beta}^{j,k} - \overline{X\beta}^{k})]$.

Firm pay premium defined as $var(\Psi^{j,k} - \bar{\Psi}^{k})$.

Firm sorting defined as $2cov[(\bar{\theta}^{j,k} - \bar{\theta}^{k}), (\Psi^{j,k} - \bar{\Psi}^{k})] + 2cov[(\Psi^{j,k} - \bar{\Psi}^{k}),](\overline{X\beta}^{j,k} - \overline{X\beta}^{k})$.

Industry segregation defined as $var(\bar{\theta}^{k}) + var(\overline{X\beta}^{k}) + 2cov(\bar{\theta}^{k}, \overline{X\beta}^{k})$.

Industry pay premium defined as $var(\bar{\Psi}^{k})$.

Industry sorting defined as $2cov(\bar{\theta}^{k}, \bar{\Psi}^{k}) + 2cov(\bar{\Psi}^{k}, \overline{X\beta}^{k})$.

Person effect defined as $var(\theta^{i} - \bar{\theta}^{j,k}) + var(X\beta^{i,j,k} - \overline{X\beta}^{j,k}) + 2cov(\theta^{i} - \bar{\theta}^{j,k}, X\beta^{i,j,k} - \overline{X\beta}^{j,k})$.

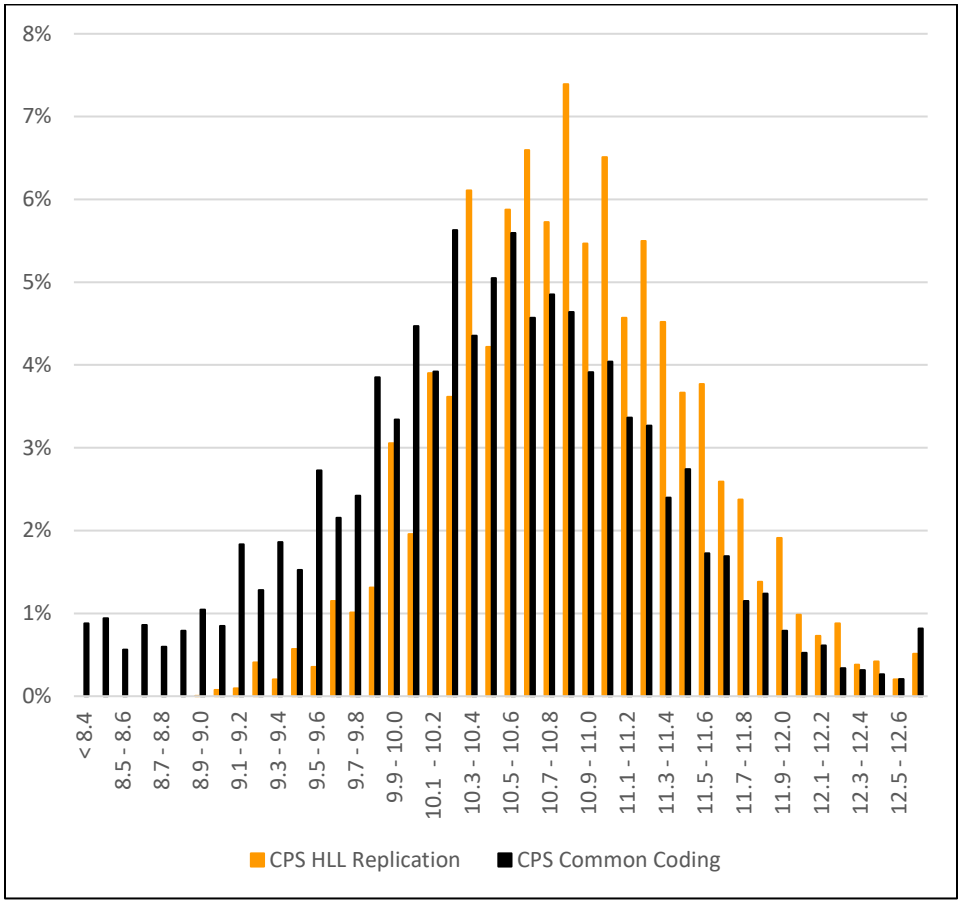
Residual defined as $var(\epsilon^{i,j,k}) + 2cov(\theta^{i} - \bar{\theta}^{j,k}, \epsilon^{i,j,k}) + 2cov(X\beta^{i,j,k} - \overline{X\beta}^{j,k}, \epsilon^{i,j,k})$.

Table 2: Industry aggregation and within- versus between-industry inequality

	(1) HLL CPS	(2) HLL CPS	(3) HHS LEHD
Variance 1996-2002	.3604	.3604	.7938
Within Industry	95.7%	85.6%	78.6%
Between Industry	4.3%	14.4%	21.4%
Number Industries	12	236	301
Variance 2012-2018	.3966	.3966	.9152
Within Industry	94.9%	84.1%	73.2%
Between Industry	5.1%	15.9%	26.8%
Number Industries	12	221	301
Variance Growth	.0361	.0361	.1214
Within Industry	86.8%	69.0%	38.1%
Between Industry	13.2%	31.0%	61.9%

Notes: The earnings measure in the CPS is labor earnings.
Column 3 is from Haltiwanger, Hyatt, and Spletzer (2021).

Figure 1: PDFs of the HLL CPS and the CPS with Common Coding



Notes: The earnings measure in the CPS is labor earnings.
These pdfs are computed from the 2012-2018 data.

Table 3: Accounting for Sample Differences

	(1) HLL CPS	(2) CPS with common coding	(3) HHS LEHD
Variance 1996-2002	.3604	.7035	.7938
Within Industry	85.6%	81.7%	78.6%
Between Industry	14.4%	18.3%	21.4%
Number Industries	236	236	301
Variance 2012-2018	.3966	.7373	.9152
Within Industry	84.1%	80.1%	73.2%
Between Industry	15.9%	19.9%	26.9%
Number Industries	221	221	301
Variance Growth	.0361	.0338	.1214
Within Industry	69.0%	45.8%	38.0%
Between Industry	31.0%	54.2%	62.0%

Notes: The earnings measure in the CPS is labor earnings.
Common coding is defined in the text.
Column 3 is from Haltiwanger, Hyatt, and Spletzer (2021).

Table 4: Variances from regression $Y = Age*Educ\beta_1 + Occupation\beta_2 + Industry\beta_3 + \varepsilon$

	(1) HLL CPS	(2) CPS with common coding
Variance 1996-2002	.3604	.7035
var(Age*Educβ_1)	14.6%	12.4%
var(Occupationβ_2)	7.9%	6.7%
var(Industryβ_3)	4.8%	6.4%
var(ε)	66.5%	67.1%
2cov(A*Eβ_1,Occβ_2)	9.3%	6.6%
2cov(A*Eβ_1,Indβ_3)	-1.9%	0.4%
2cov(Occβ_2,Indβ_3)	-1.2%	0.4%
Variance 2012-2018	.3966	.7373
var(Age*Educβ_1)	14.7%	13.7%
var(Occupationβ_2)	7.7%	6.8%
var(Industryβ_3)	3.9%	4.7%
var(ε)	66.4%	65.2%
2cov(A*Eβ_1,Occβ_2)	9.2%	7.6%
2cov(A*Eβ_1,Indβ_3)	-1.4%	1.0%
2cov(Occβ_2,Indβ_3)	-0.5%	1.1%
Variance Growth	.0361	.0338
var(Age*Educβ_1)	16.5%	40.5%
var(Occupationβ_2)	5.4%	8.5%
var(Industryβ_3)	-5.7%	-30.8%
var(ε)	65.3%	25.2%
2cov(A*Eβ_1,Occβ_2)	8.2%	27.4%
2cov(A*Eβ_1,Indβ_3)	3.7%	13.0%
2cov(Occβ_2,Indβ_3)	6.6%	16.1%

Notes: The earnings measure in the CPS is labor earnings.
Common coding is defined in the text.