Male Earnings Volatility in LEHD Before, During, and After the Great Recession

Kevin L. McKinney and John M. Abowd

ABSTRACT
This article is part of a coordinated collection of papers on prime-age male earnings volatility. Each paper produces a similar set of statistics for the same reference population using a different primary data source. Our primary data source is the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) infrastructure files. Using LEHD data from 1998 to 2016, we create a well-defined population frame to facilitate accurate estimation of temporal changes comparable to designed longitudinal samples of people. We show that earnings volatility, excluding increases during recessions, has declined over the analysis period, a finding robust to various sensitivity analyses.

1. Introduction
This article is part of a coordinated collection of papers on prime-age male earnings volatility. Each paper produces a similar set of statistics for the same reference population using a different primary data source. Our paper uses unemployment insurance (UI) based administrative worker annual earnings data from the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. To put our article in context, we note it is based on sample selection methods developed in Abowd, McKinney, and Zhao (2018; AMZ hereafter). Among other findings, AMZ demonstrated that in order to draw comparisons between administrative, designed cross-sectional, and designed longitudinal estimates of related labor market phenomena like earnings inequality and volatility, one must construct a proper, dynamic universe as a reference population. Armed with that construct, which we also employ in this article, we show that it is essential to pay careful attention to the left tail (bottom) of the earnings distribution.

Using our consistent, dynamic population frame, we estimate earnings volatility trends for prime-age males from 1998 to 2016. Our main result is that volatility has declined over the analysis period, excluding increases during recessions. Although the level of volatility differs across various specifications, the downward trend in volatility is robust to all sensitivity analyses.

Our article is organized as follows. Section 2 describes the methods we used to construct the LEHD analysis samples. Section 3 presents our statistical results. Section 4 concludes.

2. Data
The empirical work in this paper uses earnings information from the Longitudinal Employer-Household Dynamics (LEHD) infrastructure files, developed and maintained by the U.S. Census Bureau (Abowd et al. 2009). From this data source, we construct annual person-level earnings files covering the period 1998–2017.

In contrast with survey data sources such as the CPS, PSID, and ACS, LEHD data contain annual earnings for the virtual universe of private wage and salary workers in the United States. The essentially complete coverage of the universe enables a detailed analysis of worker earnings volatility. However, because of an important artifact documented in Appendix A, supplementary materials, it is essential to use earnings associated with only “eligible” workers—those who are part of our dynamic frame—namely, individuals with identifiers issued by the Social Security Administration, and used by the person to whom they were issued. Adopting this population frame, allows us to consistently estimate calendar-year time trends for our earnings volatility measures.

In the LEHD data infrastructure, a “job” is the statutory employment of a worker by a statutory employer as defined by the Unemployment Insurance (UI) system in a given state. Mandated reporting of UI-covered wage and salary payments between one statutory employer and one statutory employee is governed by the state’s UI system. Reporting covers private employers and state and local government. There are no self-employment earnings unless the proprietor drew a salary, which, for UI earnings data, is indistinguishable from other employees.

States joined the Local Employment Dynamics federal/state partnership that supplies input data to the LEHD program at different dates as illustrated in Appendix Table A1, supplementary materials. When a state joined, the data custodians were asked to produce historical data for as many quarters in the past, back to 1990Q1, as could be reasonably recovered from their information storage systems. As a result, the date that a
data-supplying entity joined the partnership is not the same as the first quarter in which that entity’s data appear in the system. The start date for any state depends primarily on the amount of historical data the state could recover at the time it joined. This potential ignorability (in the sense of Rubin 1987 or Imbens and Rubin 2015) of the start date for a segment of the LEHD data—that is the possibility that state start dates are conceivably not related to data quality or earnings volatility—is the basis for our methods of constructing a time-series of nationally representative estimates.

Although state entry is a potential concern, AMZ show that the annual earnings distributions for the subset of states available by 1995Q1 are almost identical to the complete data annual earnings distribution. However, in this article we are not estimating measures of earnings inequality but are instead measuring earnings volatility. The variance measures used here are especially sensitive to earnings changes at the bottom of the earnings distribution, where a relatively small change in absolute value can result in a large percentage change. Even though AMZ show the annual earnings distributions between the fifth and the ninety-fifth percentiles are not noticeably affected by state entry/exit, we need to be cautious in applying their results to this paper because earnings changes for workers in the bottom 5% of the earnings distribution potentially have an outsized impact on the results.

Although the LEHD data provide a high-quality jobs frame, individual identifier misuse complicates the time-varying many-to-one assignment of jobs to workers. Therefore, it is preferable to have a person frame that covers a known population of interest, such as all persons legally eligible to work in the United States. For our analysis, we create a frame of workers using the Census Bureau’s edited version of the Social Security Administration’s master SSN database (“the Census Numident”), capturing all officially reported employment-eligible workers but removing jobs associated with ineligible workers, as we elaborate below.

LEHD earnings records are reported quarterly by the employing firm. These records contain a nine-digit person identifier, typically assumed to be a Social Security Number. However, at the time the report is received by the state UI office, the nine-digit person identifier is not verified, resulting in records both with and without a valid SSN. Using the Census Numident we ascertain if each earnings record is associated with a valid SSN. Records not associated with a valid SSN may have an alternate valid person identifier such as an IRS-issued Individual Taxpayer Identification Number (ITIN); nevertheless, we can only distinguish between valid and invalid SSNs. If the SSN is valid, in addition to the UI employment history we have access to demographic characteristics, such as sex and date of birth, from the Census Numident.

Using both the Census Numident and the employment histories from the UI data, we create a “prime-age male eligible-workers” frame, including only workers in a given year that meet the following criteria: individual has a valid SSN on the Census Numident; gender of the individual is male; the year is between 1998 and 2017, inclusive; the modal age of the individual during the year is between 25 and 59, inclusive; the year is greater than the SSN year-of-issue and less than the year of death (if available); and the SSN is associated with fewer than 12 jobs during the year. An eligible worker is labeled as “active” in the labor market for a particular year when UI earnings are positive in that year and “inactive” otherwise. Inactive status is thus inferred based on the absence of positive earnings reports.

The purpose of the prime-age male eligible-workers frame is 2-fold. First, the Census Numident data allow us to consistently identify a set of males legally eligible to work each year, while at the same time implicitly removing earning records from our analysis sample not associated with individuals in the covered legally eligible-to-work population. However, we go a step further, removing earnings records with valid SSNs where the available data strongly suggest the SSN is not being used by the person to whom it was issued (Brown, Hotchkiss, and Quispe-Agnoli 2013, AMZ). Our criteria for excluding earnings records uses three annual exclusion rules, preserving job-level earnings records during years when none of the rules are violated. The three exclusion rules are: job years with positive UI earnings prior to the SSN being issued are removed, while later years are eligible for inclusion; job years with positive reported earnings where the individual is reported dead prior to the start of the year are removed, although earlier years are eligible for inclusion; and job years where the individual has more than 12 employers are excluded, although other years are eligible for inclusion. Appendix Table A2, supplementary materials shows the resulting number of analysis observations per year.

Our earnings measure is based on annual UI job-level earnings reports. First, we adjust nominal earnings to real earnings using the Bureau of Economic Analysis’ Personal Consumption Expenditures (PCE) index, with 2010 as the base year. Second, we calculate person-level real annual earnings $e_{it}$ for each eligible male worker $i$ in each year $t$ as the sum of real earnings $v_{ijt}$ at all firms $j$ for each worker $i$ in each year $t$.

$$e_{it} = \sum_j v_{ijt}, \quad l_{it} = \ln(e_{it+1}) - \ln(e_{it}), \quad a_{it} = \frac{(e_{it+1} - e_{it})}{(e_{it} + e_{it+1})/2}$$

To estimate earnings volatility, we create two measures of the change in annual earnings. Our first earnings volatility measure is the difference in log earnings (DLE) from the initial year $t$ to the subsequent year $t + 1$. The DLE measure, $l_{it}$, is available from 1998 to 2016 for workers with positive earnings in both years. We also analyze a second measure, $a_{it}$, the arc percentage change (APC), expressed herein as a proportion as is conventional in this literature (Dahl, DeLeire, and Schwabish 2008, 2011; Ziliak, Hardy, and Bollinger 2011). The APC is available from 1998 to 2016 for all workers in the composite sample and unlike the DLE, the APC can be used when one of the earnings observations for the year-pair is zero.

If the change in earnings is moderate, $l_{it}$ and $a_{it}$ produce similar results. For example, when earnings decrease by less than 50% or increase less than 100%, the relative difference between the two measures is no more than three percentage points. However, when there is an extremely large percentage change in earnings the two measures may differ substantially. For example, if a worker earns $1000 in the first year and $50,000 in the second year of a year-pair, the difference in log earnings is over twice as large as the arc percentage change (3.91 vs. 1.92). Given the sensitivity of the variance calculation to large earnings changes, the two measures may produce very different results even when presented with the same data.
3. Results

3.1. Baseline Trends

Standard practice in the literature is to separately trim initial and subsequent-year earnings using stated minimum and maximum values, thus placing an upper bound on the absolute value of the earnings change. We call this a “by-year” trim. A secondary but perhaps more important effect of trimming is that it reduces the number of very low earning workers, for whom a relatively small absolute change in earnings can result in a large percentage change. We implement trimming by year, setting the minimum at the 1st percentile and the maximum at the 99th percentile of the real earnings distribution for the actual data year. We also implement constant or “same each year” trimming using the percentiles estimated from the combined earnings distribution across all years of the composite sample. There is merit in both constant and by-year trimming. By using the same values each year, constant trimming prioritizes a worker’s absolute position in the real earnings distribution, which can be especially important for workers at the bottom of the earnings distribution where the absolute level of earnings may be material to their existence. However, using time-varying trim values allows us to include the set of workers each year who are in the same relative position in the real earnings distribution.

The choice of trim values is especially important for trend analysis if there are large changes in the tails of the cross-sectional earnings distribution over time or the tails are estimated imprecisely due to small sample sizes. Trends estimated using either the constant trim or the by-year trim would be similarly affected by increases in relative volatility in the tails of the earnings distribution, however, trends estimated using a by-year percentile-based trim will likely differ from trends estimated using a constant trim because the share of workers in the tails is changing over time. For example, if the earnings distribution shifts to the left, similar to what might happen during a recession, the constant trim would remove a larger fraction of low earning workers from the analysis sample than the by-year trim, resulting in a lower estimate of earnings volatility.

The results for both the DLE measure of volatility and the APC measure of volatility are shown in Figure 1. Three themes are evident. First, earnings volatility is counter-cyclical with increases during recessions (2000–2001, 2007–2008) and decreases during expansions. Second, ignoring the cyclicality, earnings volatility is relatively stable until 2008, when it begins a sustained decline. Third, volatility measured using the DLE is always larger than volatility measured using the APC.

The primary effect of trimming is to reduce the level of measured volatility. For example, using the DLE the untrimmed level of volatility at the 2008 peak of the great recession is about 0.62, while the trimmed values are 0.45 for the constant trim and 0.46 for the by-year trim. The substantial reduction in volatility highlights the sensitivity of the variance to the tails of the earnings distribution. The trends for the trimmed and untrimmed series are similar, although there are small, but noticeable, differences between the constant-trim and by-year trim estimates. Volatility is larger in the by-year trim series during recessions as workers face negative earnings shocks and the earnings distribution shifts to the left, resulting in a larger share of lower earnings workers in the analysis sample and higher volatility. The trends are similar for both the DLE and APC, but the effects of trimming are smaller for the APC. The APC incorporates a smooth limit to the volatility calculation, reducing the impact of the trim. For the same reason, the increase in APC volatility during recessions is also muted, irrespective of the trim. Additional sensitivity analyses are found in the Appendix, supplementary materials; Figure A2, supplementary materials shows the impact of state entry/exit, Figure B2 shows the impact of adjusting for age, Figure C1 compares our main result with a nonparametric measure of earnings volatility, Figure D1 shows a comparison of trim value trends, Figure D2 shows earn-
ings volatility trends for selected other trimming approaches, and Figure E2 shows the impact of including zero earning years.

### 3.2. Comparisons with Survey Data

Our primary analysis dataset is based on state Unemployment Insurance (UI) records. Compared with survey data sources, LEHD UI data covers a different population of jobs and workers. In this section, we document the differences in earnings volatility between the LEHD, a sample of LEHD annual earnings for workers who appear in the American Community Survey (the LEHD-ACS sample), and the PSID.

In our first exercise we use the matched LEHD-ACS sample to estimate the impact of missing earnings from federal and “self-employed in a not incorporated firm” jobs on our LEHD earnings volatility estimates. Workers that transition in and out of LEHD UI covered employment within a year-pair are especially problematic, potentially increasing our earnings volatility estimates. To create the LEHD-ACS sample, we match each LEHD worker year-pair observation with the ACS data by SSN and year, allowing us to identify records in the LEHD-ACS matched sample who are likely to have a job in employment sectors not covered in the LEHD data. See Appendix Table F1, supplementary materials for more details. The LEHD earnings volatility results for the LEHD-ACS sample are shown in Figure 2.

Compared with the baseline LEHD earnings volatility estimates shown in Figure 1, LEHD-ACS earnings volatility shown in Figure 2 is lower, although the trend is almost identical. The reduced volatility of the LEHD-ACS series relative to the LEHD series highlights the lower earnings volatility of ACS respondents in general. However, both the LEHD and the LEHD-ACS earnings volatility series are likely biased upwards by workers who move in and out of LEHD-covered employment sectors. The LEHD-ACS UI Covered Employment series shows the impact of removing these workers from the analysis sample, reducing the level of earnings volatility by about 6%, but the trend in earnings volatility is almost identical.

The PSID has been the workhorse data set in the earnings volatility literature. A more in-depth comparison of the impact of differences in cross-sectional earnings distributions in the PSID and the LEHD on estimated volatility trends is therefore warranted. Given the large differences between the PSID and LEHD data in the share of workers at the bottom of the earnings distribution, the estimated level of volatility will differ between the two data sources. However, perhaps more important, different trends in the share of workers at the bottom of the earnings distribution and different trends in volatility for these workers have the potential to noticeably affect the trend in overall earnings volatility.

For the set of LEHD workers with earnings in both years of the year-pair, we create two new earnings trims. Starting with the LEHD data trimmed at the 1st and 99th by-year percentiles of the LEHD earnings distribution, we then also trim LEHD initial and subsequent-year earnings at the PSID by-year 1st and 99th percentile annual real earnings values. Next, we trim LEHD average annual real earnings for the year-pair at the PSID 1st and 99th percentile year-pair average real earnings values. Using the PSID ventiles, we also create a set of weights designed such that the weighted LEHD annual earnings distribution is approximately equal to the corresponding PSID average annual earnings distribution.

Figure 3 shows the results of applying the various combinations of PSID trims and PSID weights to the LEHD data. All the series shown in Figure 3 use LEHD earnings data, the PSID data are only used to determine weights and trim values. As a reference point, we plot the no weight, no trim (label: No Trim) variance of the arc percentage change series shown in Figure 1. Applying just the PSID trim (label: PSID P1-P99 by Year Trim: No PSID Weight), the overall level of earnings volatility falls and the trend in volatility changes from decreasing to increasing.
Applying both the PSID trim and the PSID weights (label: PSID P1-P99 by Year Trim : PSID Weights) further reduces the level of volatility relative to the PSID-trim-only series, while also moderating the increasing trend. For comparison, we also show an additional series without PSID trims and weights where the initial and subsequent earnings values are trimmed at the LEHD 5th and 99th percentile by-year earnings values (label: LEHD P5-P99 by Year Trim). Both series show the impact of removing low earnings values, although the trend changes only when using the PSID trims and weights. Unlike in the LEHD data and other trimming strategies where the by-year trim values are relatively stable over time, the 1st percentile by-year trim values for the PSID decrease significantly over the analysis period and it is this feature of the PSID earnings distribution that generates the difference. See Appendix G, supplementary materials for additional details.

3.3. Volatility by Initial-Year Earnings

Above we have focused exclusively on estimating national earnings volatility trends, however our large sample size allows us to estimate separate results for subpopulations. In our final two figures we show results separately for workers at different levels of the initial-year earnings distribution. Workers in the bottom 25% of the initial-year earnings distribution are responsible for approximately 65% of earnings volatility and may have trends.
in volatility that differ substantially from workers at the middle and top of the initial-year earnings distribution. In Figure 4 we show LEHD earnings volatility estimates within initial-year earnings bin. Earnings volatility levels follow a U-shaped pattern as we move up the earnings distribution; earnings volatility is highest at the bottom, decreasing as earnings increase except at the very top where volatility increases. However, perhaps more importantly, earnings volatility trends are decreasing within each earnings bin. This result makes it extremely unlikely that any type of earnings bin reweighting would produce a positive earnings volatility trend.

Although it is hard to tell from Figure 4, there are interesting differences across earnings bins in the relative (compared with 1998) changes in earnings volatility. For example, in Figure 5 we see that the largest relative reduction in earnings volatility occurs at the top of the earnings distribution.

Earnings volatility at the top of the earnings distribution in 2016 is approximately 35% lower than it was in 1998. In contrast to Figure 4, there is an almost monotonic decrease in the relative decline in inequality as we move down the initial-year earnings distribution. Although workers at the bottom of the initial-year earnings distribution have seen a large absolute decrease in earnings volatility, the relative decline is small compared with workers higher up the initial-year earnings distribution who have low absolute levels of volatility yet also have the largest relative reductions in volatility.

4. Conclusions

Our estimates of LEHD earnings volatility trends are robust. The large sample sizes combined with our efforts to create a consistent dynamic frame result in an accurate representation of the bottom tail of the earnings distribution, a requirement when estimating earnings volatility trends. Although the relatively large number of low earning workers in LEHD data generate relatively high levels of earnings volatility, a simple by-year 5th percentile trim reduces our earnings volatility estimates to levels roughly comparable to various survey data sources such as the PSID, CPS, and SIPP.

At the national level, earnings volatility trends show a strong decline after the Great Recession. However, except for workers at the bottom of the initial-year earnings distribution (<25th percentile), the downward trend in volatility likely started earlier, after the previous recession in 2001. The downward trend in earnings volatility after 2001 is particularly strong for workers at the top of the earnings distribution.

Although volatility is declining overall, the decline is occurring most slowly for workers least able to self-insure. Future work would do well to exploit the dense nature of administrative data sources to better understand earnings volatility dynamics, keeping in mind that some level of earnings volatility is desirable.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the US Census Bureau or other sponsors. All results have been reviewed to ensure that no confidential information is disclosed (DRB clearance numbers CBDRB-FY20-279, CBDRB-FY20-288, CBDRB-FY20-420, CBDRB-FY21-CED006-0023, CBDRB-FY22-089, CBDRB-FY22-CED006-0011). This research uses data from the US Census Bureau’s Longitudinal Employer-Household Dynamics Program, which was partially supported by NSF grants SES-9978093, SES-0339191, and ITR-0427889; National Institute on Aging grant AG018854; and grants from the Alfred P. Sloan Foundation.

References

Abowd, J. M., McKinney, K. L., and Zhao, N. L. (2018), “Earnings Inequality and Mobility Trends in The United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data,” Journal of Labor Economics, 36, S183–S300. [33]

Abowd, J. M., Stephens, B. E., Vilhuber, L., Andersson, F., McKinney, K. L., Roemer, M., and Woodcock, S. (2009), “The LEHD Infrastructure Files
Brown, J. D., Hotchkiss, J. L., and Quispe-Agnoli, M. (2013), "Does Employing Undocumented Workers give Firms a Competitive Advantage?", *Journal of Regional Science*, 53, 158–170. [34]

Dahl, M., DeLeire, T., and Schwabish, J. A. (2011), "Estimates of Year-to-Year Volatility in Earnings and in Household Incomes from Administrative, Survey, and Matched Data," *Journal of Human Resources*, 46, 750–774. [34]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press. [34]

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley. [34]

Ziliak, J. P., Hardy, B., and Bollinger, C. (2011), "Earnings Volatility in America: Evidence from Matched CPS," *Labour Economics*, 18, 742–754. [34]