Using Synthetic Data to Estimate Earnings Dynamics: Evidence from the SIPP GSF and SIPP SSB

Michael D. Carr* Emily E. Wiemers† Robert A. Moffitt ‡

January 25, 2023

1 Introduction

The SIPP Gold Standard File (SIPP GSF), which links SIPP survey data to administrative earnings and benefits records, was used in a set of projects measuring trends in earnings volatility over time (Carr and Wiemers, 2021), comparing trends in earnings volatility over time across multiple administrative and survey datasets (Carr, Moffitt, and Wiemers, 2023; Moffitt et al., 2023), and examining trends in transitory earnings variances using the model in Moffitt and Zhang (2018). For each of these projects, the results in the SIPP GSF were obtained through access to the SIPP Synthetic Beta (SIPP SSB) data, a fully synthetic version of the SIPP GSF. Though the purpose of the projects was not to compare results from the SIPP SSB data to results from the SIPP GSF, we can use our results to make a set of comparisons on trends in volatility and transitory earnings variances in the two datasets. The current paper reports on differences across the two datasets and comments on some challenges in using fully synthetic data to estimate earnings dynamics.

Recent work has come to opposite conclusions on trends in gross male earnings volatility and transitory earnings variances, with some arguing that declines in aggregate volatility at the macro

*Department of Economics, University of Massachusetts Boston. Email: michael.carr@umb.edu
†Department of Public Administration and International Affairs, Syracuse University. Email: eewiemer@syr.edu.
‡Department of Economics, Johns Hopkins University. Email moffitt@jhu.edu.

level have been accompanied by increasingly stable earnings at the micro level while others argue that as earnings inequality has grown, so has earnings volatility for workers (see (Moffitt et al., 2023) for a discussion of this literature). Trends in earnings volatility are sensitive to differences in methods and samples (Carr and Wiemers, 2018; Celik et al., 2012), imputations (Dahl, DeLeire, and Schwabish, 2011), treatment of earnings nonresponse (Bollinger et al., 2019; Ziliak, Hokayem, and Bollinger, 2020), earnings source (Ziliak, Hokayem, and Bollinger, 2020) and how low earnings are handled (Carr and Wiemers, 2021). Each of these sources plays a role in differences in results across studies. Less work has been done reconciling trends across studies estimating transitory variances but there are differences in modeling choices and the treatment of low earnings across studies (Debacker et al., 2013; Kopczuk, Saez, and Song, 2010; Moffitt and Zhang, 2018; Moffitt and Gottschalk, 2012).

In this paper, we show estimates of trends in gross earnings volatility and transitory variances in the SIPP GSF and SIPP SSB. We estimate gross earnings volatility under several assumptions about the treatment of low earnings and using two different measures of volatility and make comparisons on the trend and level in volatility in the two datasets. We also estimate transitory and permanent variances of earnings using the model in Moffitt and Zhang (2018) on both the SIPP GSF and the SIPP SSB. We find that the level of gross volatility is higher in the SIPP SSB than in the SIPP GSF and displays slightly more cyclicality. Differences in the levels between the two datasets depend on the measure of volatility and on how low earnings are treated. However, broad conclusions on the direction of trends are robust across the datasets as are conclusions on the effects of different treatment of low earnings. Results on estimates of the transitory variance suggest that the composition of total inequality differs between datasets where transitory variances contribute more to total inequality in the SIPP SSB (Note: In what follows, we only describe the results for estimating gross volatility. We think we have a version of results for transitory earnings variances that is comparable between datasets but we are limited in what we can find at the moment because the SIPP SSB is no longer available on the Cornell Virtual RDC Synthetic Server. We are working

with the server administrators to gain access to files from the server and we are confident we will be able to retrieve the files we need. Please also excuse the screen shots in Figures 1 and 2. We hope to retrieve the data from the SSB to make better figures but we do not have access to it right now.)

Comparisons between the SIPP SSB and SIPP GSF for estimates of gross male earnings volatility and the transitory variance of earnings are interesting for a few reasons. First, the imputation method in the SIPP SSB was not designed to preserve the variance of earnings changes so comparisons represent a somewhat "out of sample" test. Second, the SIPP SSB and GSF differ across dimensions that have been shown to matter for estimating trends in earnings volatility and transitory variances. Finally, we are applying the data to a research question where there is disagreement on the trends which suggests a need to consider how to use synthetic data alongside non-synthetic versions of data sources.

2 Data and Methods

2.1 SIPP Gold Standard File (SIPP GSF)

The SIPP GSF contains all individuals in a SIPP household in the 1984 and in the 1990 to 2008 SIPP panels linked to administrative Social Security earnings in the Detailed Earnings Record (DER) and to the Master Benefits Records. DER earnings are available from 1978 through 2014. This match is both prospective and retrospective so, for example, an individual appearing in the 1996 SIPP panel would be matched retrospectively to their administrative earnings records prior to 1996 and prospectively to their administrative earnings records from 1996 on.

Earnings histories in the SIPP GSF come from the DER, which is co-maintained by SSA and IRS. The measure of earnings that we use represents total earnings from all FICA-covered and non-FICA covered jobs with a W-2 or Schedule C (self-employment) filing. W-2 earnings are the

sum of amounts from Box 1 (Total Wages, Tips, and Bonuses) and Box 12 (earnings deferred to a 401(k) type account). The data do not allow disaggregation of this sum into its components. Earnings are not top coded.

The match rate between survey and administrative data for most panels is quite high. In the 1980's and 1990's panels, the match rate hovered around 80%. In 2001, the match rate dropped to 47% because many SIPP participants refused to provide social security numbers for matching. Beginning with the 2004 panel, the match rate increased to around 90% because the Census Bureau changed its matching procedures removing the necessity to explicitly ask for social security numbers. Aggregate annual match rates for men age 25 to 59 decline slightly over time from about 80% to 70% with a cumulative match rate of 74% across the entire period. In our analysis, we use only the sample of individuals who could be matched to their administrative earnings records.

We use a sample of men age 25 to 59 who could be matched to administrative earnings histories. Earnings are adjusted for inflation using the PCE Index. We implement five different trimming methods to estimate volatility. We trim the bottom 1% (P1) and 5% (P5) of positive earnings, respectively, separately by year, as is typically done using survey data. The three other trimming methods follow those used in the recent administrative data literature. First, we exclude in each year individuals with real earnings that are below one quarter of full-time employment at the minimum wage in 2011, or \$3,770 in 2011 dollars (\$3770) (Kopczuk, Saez, and Song, 2010); second, we exclude earnings below one quarter of full-time employment at one-half of the minimum wage in each year (Min Wage) (Debacker et al., 2013; Guvenen, Ozkan, and Song, 2014); and finally, we exclude earnings below that needed to have four quarters of credit towards Social Security coverage in that calendar year (SSA Threshold) (Sabelhaus and Song, 2009, 2010). The primary difference between the trims is that the percentile point trims change with observed changes in the earnings distribution while the absolute dollar trims do not. All trims vary by year except the \$3,770 trim. Volatility measures the dispersion of changes in earnings between t and $t-1$ so an individual is excluded from the sample if earnings fall below the trim threshold in either t or $t-1$,

or both. The majority of excluded individuals have earnings below the threshold in either t or $t-1$, not both. These are the data used in Carr and Wiemers (2021). To estimate transitory variances, we use the untrimmed earnings.

2.2 SIPP Synthetic Beta (SIPP SSB)

The SIPP SSB is an entirely synthetic version of SIPP GSF that was, until recently, available to the public via the Cornell Virtual RDC Synthetic Data Server. The SIPP SSB contains four implicates from the GSF. In each implicate, each variable is synthesized along with missing data patterns. Details on the imputation methods are available in Benedetto, Stanley, and Totty (2018). The imputation method in the SSB models the joint distribution of all of the variables as a sequence of conditional marginal distributions. It essentially builds up the synthetic data one variable at a time, each time using the previously synthesized variables as explanatory variables. At the end of the process, the whole dataset is synthesized one final time with the updated values for all of the variables. From our conversations with the Census Bureau, we understand that the SSB was designed to preserve the cross-sectional means and variances of the pooled SIPP panels in a given calendar year and because the synthesis is iterative, both past and future of all variables are used to impute any given year's values. Before we present the paper (in May 2023 or Spring 2024) we will discuss further with Census Bureau how they think the imputation procedure might affect the estimates of volatility and transitory variances specifically.

We make the same sample selection criteria in the SSB as we do in the GSF. We use a sample of men age 25 to 59 with positive earnings who could be matched to their administrative earnings histories. We average results across the four implicates in the SSB.

2.3 Measures of Volatility

We estimate volatility using two measures of earnings volatility typically used in the literature: the variance of log earnings changes, or Var$(y_{it} - y_{it-\tau})$ and the variance of the arc-change in earnings, or Var$[(Y_{it} - Y_{it-\tau})/(|Y_{it}| + |Y_{it-\tau}|)/2))]$ where y_{it} ($y_{it-\tau}$) is log annual earnings of individual i at time t ($t - \tau$), Y_{it} ($Y_{it-\tau}$) is annual earnings, and $\tau = 1$. In each case, growth rates are age-adjusted separately by t using a quadratic in age, and the variance of the residuals is the estimate of volatility.

2.4 Estimating Transitory and Permanent Variances

We use the error components model in Moffitt and Zhang (2018) to estimate the variance of the transitory and permanent component of earnings. The primary advantage of this model is that it significantly relaxes the parametric assumptions underlying the earnings generating process.

$$y_{iat} = \lambda_t \mu_{ia} + \beta_t \nu_{ia} \tag{1}$$

$$\mu_{ia} = \mu_{i0} + \sum_{s=1}^{a} \omega_{is} \tag{2}$$

$$\nu_{ia} = \epsilon_{ia} + \sum_{s=1}^{a-1} \psi_{a,a-s} \epsilon_{i,a-s} \text{ for } a \geq 2 \tag{3}$$

$$\nu_{i1} = \epsilon_{i1} \text{ for } a = 1 \tag{4}$$

for $a = 1, ..., A$ and $t = 1, ..., T$. μ_{ia} is permanent earnings, which now vary by age a and through time with λ_t. Transitory earnings, ν_{ia}, also vary with age a and time with β_t. Permanent earnings is assumed to have a fixed component μ_{i0}, and evolve with age according to the permanent shocks ω_{ia}. The age specific shocks to permanent earnings ω_{ia} are assumed to be independently distributed and pass fully and permanently into permanent earnings, or that $\partial \mu_{ia} / \partial \omega_{ia} = 1$. That is, permanent earnings follow a unit root process.

Transitory earnings allows for contemporaneous shocks in ϵ_{ia} and for long-lived transitory shocks in $\epsilon_{i,a-s}$, which are assumed to be independently distributed. The impact of past transitory shocks on current earnings, $\psi_{a,a-s}$ are allowed to be unconstrained as opposed to the typical ARMA family of processes typically imposed in the literature.

From this model, it is possible to derive a theoretical covariance matrix for y_{iat}, μ_{ia}, and ν_{ia}. The coefficients of the elements of the covariance matrix are estimated by minimizing the distance between the predicted moments of the model and the empirical moments. The empirical moments of the earnings distribution are estimated for each individual i of age a between time t and $t-\tau$ going back to first year of the data or age 20, whichever happens first. Individuals are pooled into three age groups: 30 to 39, 40 to 49, and 50 to 59. The model allows variances of the permanent and transitory shocks to be nonparametric functions of age and allows the ψ parameters to be nonparametric functions of age and lag length. Moffitt and Zhang (2018) provide an in depth discussion of how the moments of the model are identified. We use the estimated coefficients to graph annual predicted permanent, transitory, and total variances of earnings for 40 to 49 year olds.

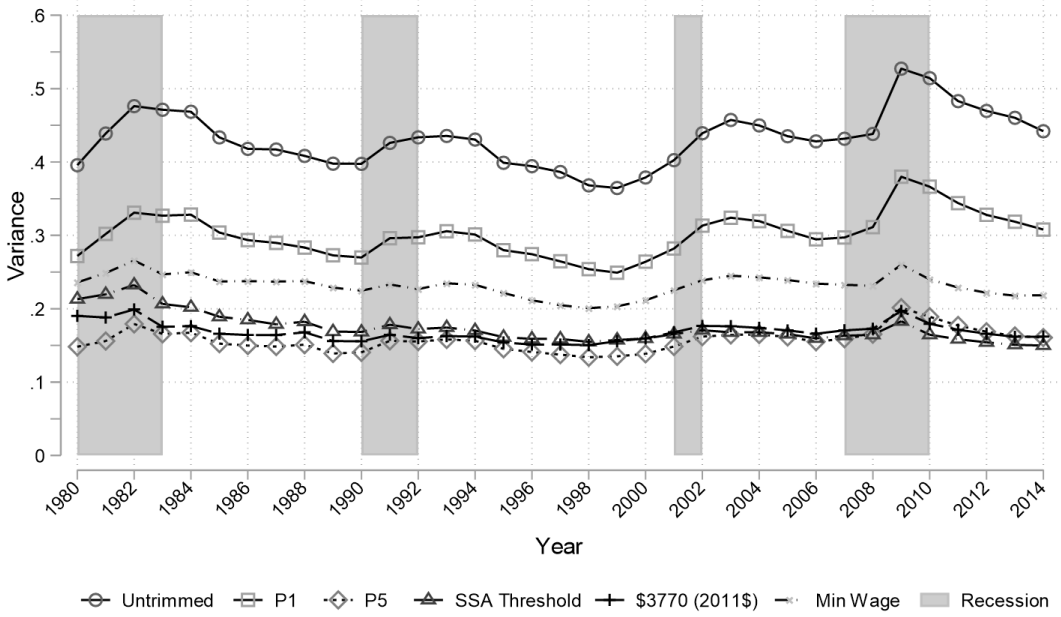
3 Results

Figures 1 and 2 show the level of volatility in the SIPP GSF and SIPP SSB for each trim using, as the measure of volatility, the variance of log changes and the variance of arc changes, respectively.

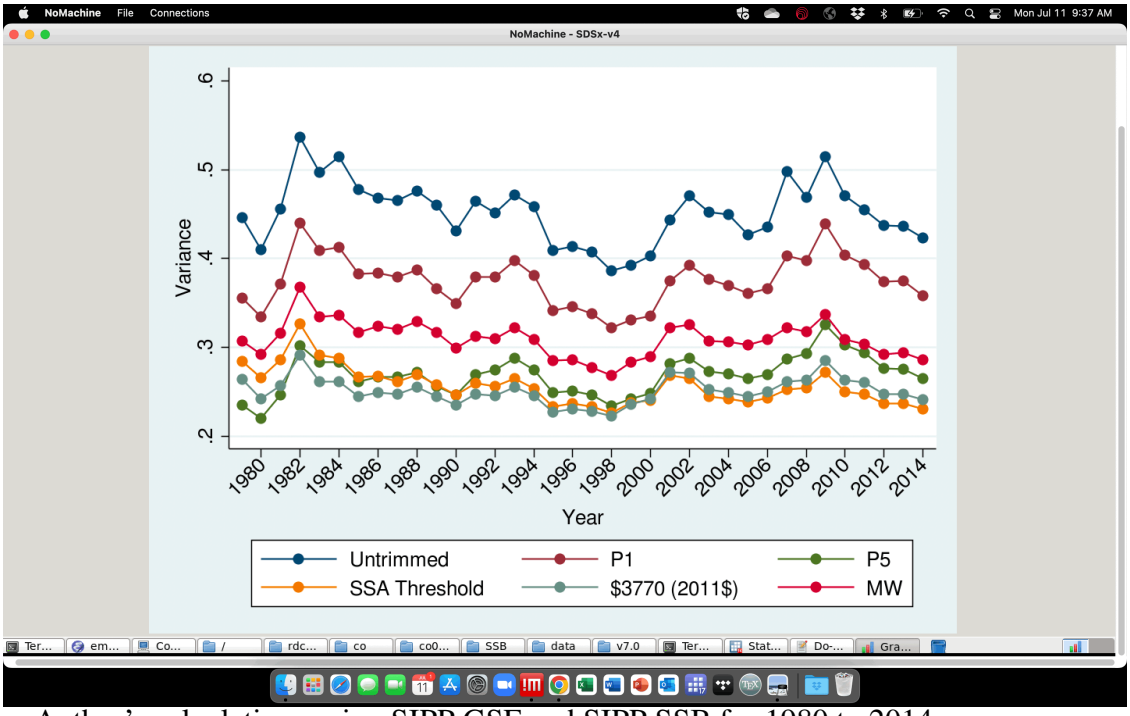
In Figure 1, the top series shows volatility in the untrimmed series. Comparing the SIPP GSF and the SIPP SSB, the level of volatility is about the same in the two series with the SIPP SSB showing more year-to-year variation and slightly higher cyclical peaks, particularly in the early years. However, trimming the data affects the correspondence between the SIPP GSF and SIPP SSB results. The second highest series in each figure shows volatility when earnings is trimmed at the first percentile. In this series, the level of volatility in the synthetic data is about 16% higher than in the true data. As with the SIPP GSF, the cyclical patterns and trends closely follow the

Figure 1: Volatility in Log Changes, SIPP GSF and SIPP SSB

(a) SIPP GSF



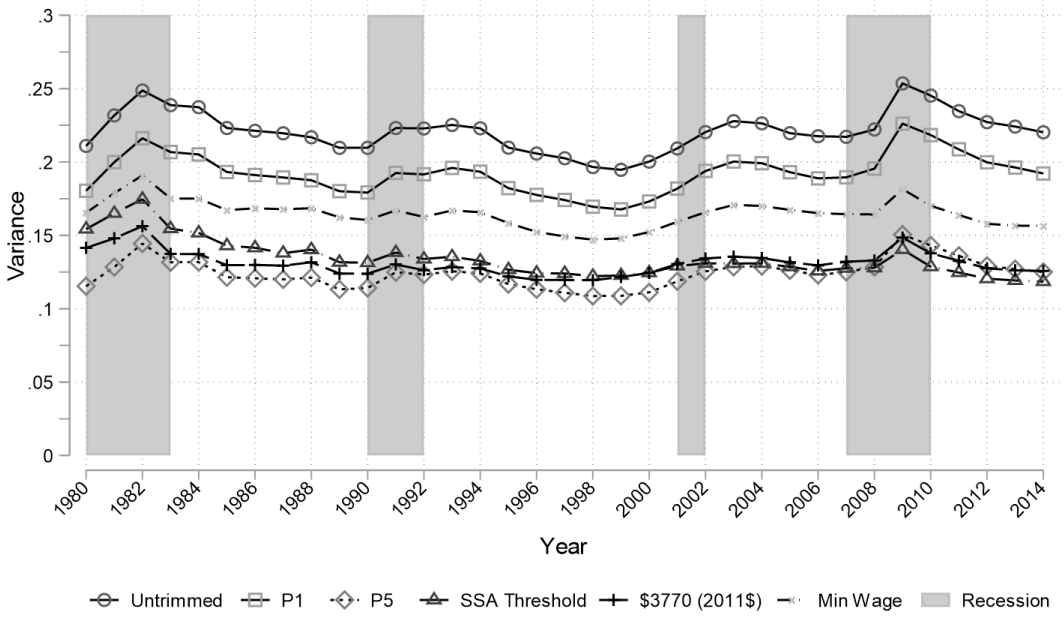
(b) SIPP SSB



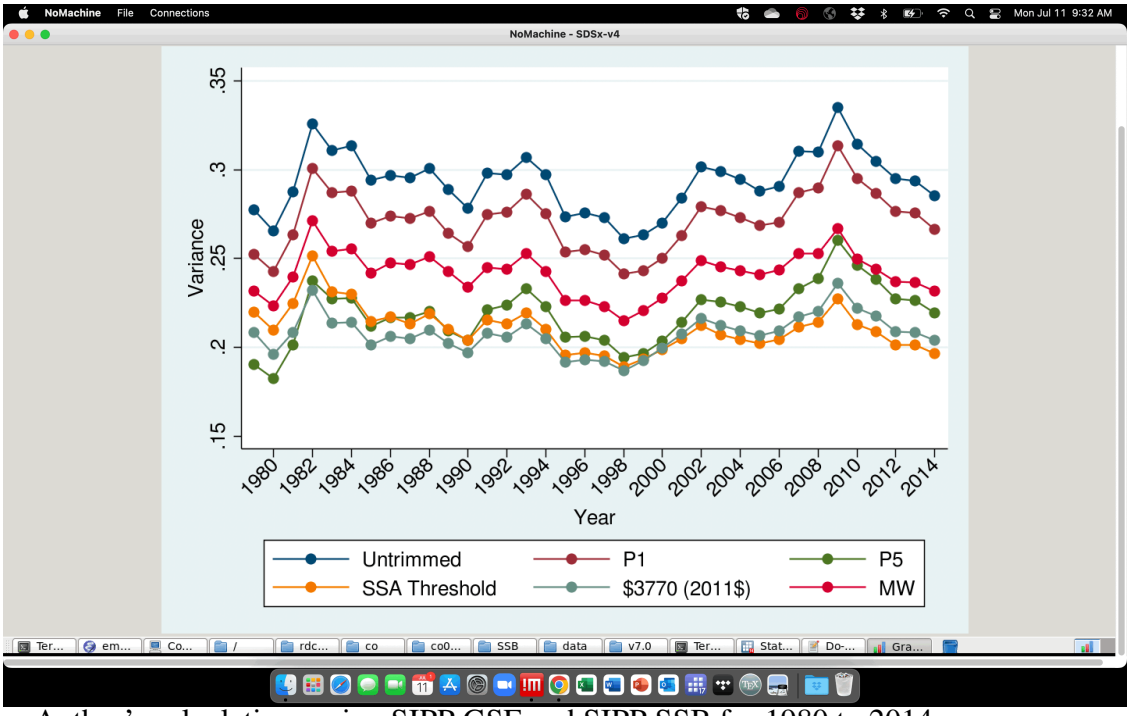
Notes: Author's calculations using SIPP GSF and SIPP SSB for 1980 to 2014.

Figure 2: Volatility in Arc Change, SIPP GSF and SIPP SSB

(a) SIPP GSF



(b) SIPP SSB



Notes: Author's calculations using SIPP GSF and SIPP SSB for 1980 to 2014.

untrimmed data with more year-to-year variation in the SSB than in the GSF. The P5, minimum wage, SSA threshold, and $3770 trims follows the same pattern as the P1, with higher levels of volatility in the SSB than in the GSF. However, we note that the level of volatility with the P5 trim is higher than with the $3770 or SSA threshold trims at the end of the period in the SSB while the level of volatility in each of these series is very similar in the GSF by 2014.

Turing to volatility estimated in arc changes in Figure 2, we see that in this case, the level of volatility is substantially higher in the SSB than in the GSF for all of the series, even the untrimmed series. Volatility in the untrimmed series in 1980 is 0.21 in the SIPP GSF compared with 0.26 in the SIPP SSB. In the P1 trim, volatility is 0.175 in GSF in 1980 while it is 37% higher (0.24) in the SSB. These level differences are substantial. As with the results in log changes, cyclicality is more pronounced and the year-to-year variation is larger in the SSB than in the GSF.

4 Preliminary Conclusions and Next Steps

Our results do not suggest that one would reach entirely different conclusions using the SIPP SSB versus SIPP GSF about the trends in volatility over time or about the effect of trimming on estimates of volatility. However, the differences between the two datasets are not consistent across method of estimating volatility or robust to choices on how to trim earnings. It would also be misleading to create a series showing the trend over time in volatility from a dataset that included some years of data from the SIPP GSF and some from the SIPP SSB because of the differences in the level of volatility and year-to-year variation in volatility between the two datasets. Even engaging in an exercise similar to Moffitt et al. (2023) and Carr, Moffitt, and Wiemers (2023) which compared trends in volatility across datasets and between the SIPP survey data and the SIPP GSF data would be more difficult using the SIPP SSB.

We note that the SIPP SSB was not created as a dataset for estimating earnings dynamics so, in some sense, it is incredible that the correspondence between the two series is as close as it is.

Though we need to have further discussions with Census Bureau staff, we believe that this is likely due to the way in which the imputations use all the available data to estimate conditional marginal distributions. Finally, we note that the SIPP SSB was not intended to be a stand-alone data product. It was designed to provide access to SIPP data linked to administrative earnings records outside of the RDC environment. We used the data frequently, published several papers using the data, and think this was a wonderful data product that democratized use of the administrative earnings data.

We are confident that we can find comparable estimates in the SIPP SSB and GSF for the transitory variances from the error components model outlined above. This is likely to happen before May 2023.

References

Benedetto, Gary, Jordan C. Stanley, and Evan Totty. 2018. "The Creation and Use of the SIPP Synthetic Beta v7.0." *https://www2.census.gov/adrm/CED/Papers/CY18/2018-11-BenedettoStanleyTotty-Creation%20SIPP.pdf* .

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

Carr, Michael D., Robert A. Moffitt, and Emily E. Wiemers. 2023. "Reconciling Trends in Volatility: Evidence from the SIPP Survey and Administrative Data." *Journal of Business & Economic Statistics* 41 (1):26–32.

Carr, Michael D. and Emily E. Wiemers. 2018. "New Evidence on Earnings Volatility in Survey and Administrative Data." *American Economic Review Papers and Proceedings* 108.

———. 2021. "The Role of Low Earnings in Differing Trends in Earnings Volatility." *Economics Letters* 199.

Celik, Sule, Chinhui Juhn, Kristin McCue, and Jesse Thompson. 2012. "Recent trends in earnings volatility: evidence from survey and administrative data." *BE Journal of Economic Analysis & Policy* 12 (2):1–26.

Dahl, Molly, Thomas DeLeire, and Jonathan A Schwabish. 2011. "Estimates of Year-to-Year Volatility in Earnings and in Household Incomes from Administrative, Survey, and Matched Data." *Journal of Human Resources* 46 (4):750–774.

Debacker, Jason, Bradley Heim, Vasia Panousi, Shanthi Ramnath, and Ivan Vidangos. 2013. "Rising Inequality: Transitory or Persistent? New Evidence from a Panel of U.S. Tax Returns." *Brookings Papers on Economic Activity* Spring.

Guvenen, Fatih, Serdar Ozkan, and Jae Song. 2014. "The Nature of Countercyclical Income Risk." *Journal of Polical Economy* 22 (3):621–660.

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song. 2010. "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data Since 1937." *Quarterly Journal of Economics* 125 (1):91–128.

Moffitt, Robert and Sisi Zhang. 2018. "Income Volatility and the PSID: Past Research and New Results." *American Economic Association Papers and Proceedings* 108:277–80.

Moffitt, Robert A., John Abowd, Christopher Bollinger, Michael D. Carr, Charles Hokayen, Kevin McKinney, Emily E. Wiemers, Sisi Zhang, and James Ziliak. 2023. "Reconciling Trends in Volatility: Evidence from the SIPP Survey and Administrative Data." *Journal of Business and Economic Statistics* 41 (11):1–11.

Moffitt, Robert A and Peter Gottschalk. 2012. "Trends in the Transitory Variance of Male Earnings Methods and Evidence." *Journal of Human Resources* 47 (1):204–236.

Sabelhaus, John and Jae Song. 2009. "Earnings Volatility Across Groups and Time." *National Tax Journal* 2 (62):347–364.

———. 2010. "The Great Moderation in Micro Labor Earnings." *Journal of Monetary Economics* 57.

Ziliak, James P., Charles Hokayem, and Christoper R. Bollinger. 2020. "Trends in Earnings Volatility using Linked Administrative and Survey Data." Working Paper. http://ukcpr.org/sites/ukcpr/files/research-pdfs/DP2020-01.pdf (accessed 7/8/2020).