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

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

May 1, 2023

Abstract

One of the methods of increased privacy protection is the creation of synthetic data sets. In this paper we consider the differences that emerge between synthetic and non-synthetic data in one of the very few attempts to create a synthetic data file for a major household survey data set, the Survey of Income and Program Participation. The data we use are non-synthetic and synthetic versions of the SIPP linked to administrative earnings histories, known as the SIPP Gold Standard File (GSF) and the SIPP Synthetic Beta (SSB), respectively. We present a set of results on short-run earnings dynamics estimated on both the SSB and the GSF, focusing on volatility – the variance of short-run earnings growth rates – and an error components decomposition of inequality into the permanent and transitory components. We find that short-run instability – both volatility and the transitory component of earnings inequality – is higher in the SSB than the GSF though the differences are somewhat dependent on modeling choices and the treatment of low earnings. Differences between the two data sets emerge both because cross-sectional inequality is higher in the SSB than in the GSF and because the dynamics of earnings over both shorter and longer periods appear to be different.

1 Introduction

It is by now well understood that release of data obtained from files containing potentially sensitive variables on specific individuals and households runs a risk of identification by outside parties and

*The authors would like to thank Lars Vilhuber and Gary Benedetto for their help accessing and using the SIPP Synthetic Beta data. One set of results is covered by review #CBDRB-FY21-095, other results predate the requirement that disclosure reviews go through the full DRB. The validation analysis does not imply endorsement by the Census Bureau of any methods, results, opinions, or views presented in this paper.

†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

the general public. One of the methods of increased privacy protection currently under discussion is the creation of synthetic data sets. The idea was originally proposed by Rubin (1993), who, based on his prior work on imputation for missing data, proposed that an entire data set be imputed–that is, synthesized–so that, with small probability, no record on the synthetic data file would be the same as any record on the original data file. He proposed constructing a parametric model that captured the relationships among the variables on the original file and then making draws from the Bayesian posterior fitted distribution to create synthetic data files which should capture the relationships in the original file. The field has seen much work since those early papers, with newer nonparametric methods, machine learning, and other approaches being used to construct synthetic data sets.[1]

In this paper we consider the differences that emerge between one of the very few attempts to create a synthetic data file for a major household survey data set linked to administrative earnings histories and the non-synthetic data it is based on. The file is called the SIPP Synthetic Beta (SSB). Development of this file began in 2003 and continued through 2022 (Abowd et al., 2006; Benedetto et al., 2018, 2013). While the specifics of how the data are created have changed over time, the basic idea remains the same. The data are built on a set of uniform extracts of variables taken from the Survey of Income and Program Participation (SIPP), a nationally reprsesentative panel survey of 15,000 to 52,000 households that began in 1984 and draws a new nationally-representative sample every two to five years. Every individual in a SIPP household who has a valid ID (e.g., social security number), including children at the time of the SIPP panel, are linked to their administrative earnings histories in the Detailed Earnings Records (DER) and federal benefits records in the Master Benefits Records (MBR). This linked file is called the SIPP Gold Standard File (GSF). The SSB is a synthesized version of these data.

The GSF was created to improve analyses of individual earnings and benefits receipt, which are two of the most important indicators of economic well being, economic inequality, and labor

[1] See (Raghunathan, 2021) for a survey of the basics of the synthetic data approach.

force activity and success. Household survey reports of earnings are well known to be misreported as well as often not reported at all with fairly high frequency. Although methodological details are limited, broadly we know that Census fit a parametric statistical model to the variables in the GSF data which was then used to produce the fully synthetic SSB, including both the subset of SIPP survey variables included in the GSF and the variables taken from administrative sources. The SSB was then made available to all researchers through a virtual RDC, with researchers having the option to disclose results from the GSF after having first produced the same results on the SSB. The creation of the SSB, therefore, provides access to one of the most commonly used sources of administrative earnings histories (the DER), without having to go to an FSRDC.

Our study concerns one of the key issues with synthetic data files, which is how faithfully they capture the relationships among the variables in the original data. There is a significant literature on how to measure accuracy by the use of different types of norms and, in the newer machine learning literature, synthetic data sets are trained to be as accurate as possible according to those norms. But because most substantive researchers are less interested in general accuracy than in accuracy for specific research questions, much work on synthetic data sets simply examines how accurate the data are for such questions. Our study is of this type. We present a set of results on short-run earnings dynamics estimated on both the SSB and the GSF, focusing on volatility – the variance of short-run earnings growth rates – and an error components decomposition of inequality into the permanent and transitory components. There is a very large literature on this question in economics that has utilized survey data, pure administrative data, and survey-linked administrative data, including several papers that use the GSF.[2] For example, we conducted a study of male earnings volatility which was recently published (Carr et al., 2023) where we first estimated volatility on the SSB then submitted the code to Census to disclose GSF results. All results in this paper follow this same procedure: first run analyses on the SSB, then submit the analysis to Census for validation and disclosure on the GSF. Note that, unlike researchers who

[2]Carr and Hardy (2022); Carr and Wiemers (2018, 2021); Carr et al. (2023); Moffitt et al. (2023)

might use the GSF inside an FSRDC, users of the SSB cannot know with certainty what the results will be prior to disclosure review. They are completely dependent on how well the SSB recreates the GSF for whatever research question and sample they have chosen.

We find that short-run instability – both volatility and the transitory component of earnings inequality – is higher in the SSB than the GSF except in one instance. These differences, however, do not seem to be attributable to the SSB simply having more "noise" in the individual-level differences that are used to estimate measures of instability. When we decompose volatility into cross-sectional inequality and the covariance of earnings across a two year period, we find that *both* cross-sectional inequality *and* the covariance of earnings are higher in the SSB than the GSF. Because higher inequality will increase volatility holding the covariance fixed, but a higher covariance will decrease volatility holding inequality fixed, whether volatility in the SSB is higher than the GSF depends on the magnitude of the difference in these two components between the SSB and the GSF. Our findings show that the difference in volatility between the SSB and GSF depends on what measure of volatility is used, and how earnings in the lower tail of the distribution are handled. The error components model shows a broadly similar pattern, with transitory inequality higher in the SSB than the GSF. But despite total inequality being higher in the SSB, transitory inequality is so much higher in the SSB than the GSF that permanent inequality is lower in the SSB than the GSF. We provide some intuition into why this may be the case despite the SSB having a higher covariance of earnings in the short-run. When looking at trends over time in volatility, we find that estimated trends are qualitatively similar between the SSB and the GSF, with some notable differences, but level estimates are consistently different between the two data sets.

An important study prior to ours assessing the accuracy of the SSB is that of Stanley and Totty (2021), who also examined the accuracy of the SSB for a number of relationships other than earnings volatility (but all related to earnings, since that is the main purpose of the SSB). Their study showed that the shape of the earnings distribution differs between the SSB and the GSF, with the SSB having a notably higher density of very low earnings and a somewhat longer right tail of

the earnings distribution. This is similar to what we find. Median earnings were quite close in the two data sets, though mean earnings were somewhat different as a result of differences in the tails. Stanley and Totty (2021) show that the correlation of earnings with demographic variables were roughly the same in the two data sets, and tended to follow the same trends over time. However, major differences in the SSB and GSF were found when state-level data on the minimum wage were merged onto the SSB and used in a state fixed effects model, a common empirical model used by SIPP users conducting policy evaluations. The authors attributed this finding to the synthetic data model, which they said did not capture such relationships. We build on this work examining how well the SSB captures short- and longer-run earnings dynamics.

The outline of our paper is as follows. We first briefly review the literature on the research question of interest, concerning the level and trend of earnings inequality. We then describe the SIPP and the SSB and GSF, following by a discussion of our methods. We then present our results and conclude with a summary.

1.1 Earnings Volatility

The study of earnings volatility, sometimes called instability, is a major research topic in economics. Earnings volatility is both of interest in and of itself and also as a causal factor in other economic outcomes, such as consumption, where the impact of earnings "shocks" on consumption has been an on-going question for almost 70 years. More generally, how individuals deal with earnings instability is a focus in much economic research. Earnings volatility can reflect instability in employment, as individuals move from job to job, or firm instability, as firms succeed or fail or are in industries with a high degree of firm turnover and shake-outs.

A specific question which has been studied for many years is whether earnings instability has gone up in the U.S. The literature began with Gottschalk and Moffitt (1994) who found earnings instability to have risen from the 1970s to the 1980s, particularly among the less educated, a phenomenon often associated with the decline in quality of low-wage jobs and deindustrialization.

But, while many subsequent studies similarly found increases in volatility (see Moffitt et al. (2023) for a review), some more recent studies have found flat or even declining levels of volatility (Dahl et al., 2011; Guvenen et al., 2014; Sabelhaus and Song, 2009, 2010). Although there are a number of possible reasons for the differences in findings, many of the latter use administrative data from the IRS or SSA while many of the former use household surveys, which are subject to reporting error and which may be less reliable. The analyses presented here draw primarily from Carr and Wiemers (2021), Carr and Hardy (2022), and Carr et al. (2023), all of which compared SIPP survey data estimates of volatility to those from the administrative data in the SSB but, in the process of checking with the verification server, also compared estimates of volatility on the SSB with the GSF (in addition to unpublished results that were also estimated first on the SSB then the GSF). Carr et al. (2023) is part of a larger set of papers that compares volatility estimates across multiple sources of administrative and survey data (Moffitt et al., 2023).

1.2 The SIPP, GSF, and SSB

The SIPP is a nationally representative sample of the civilian noninstitutionalized population of the U.S. that began in 1984 and consists of panels that follow individuals for between two and five years, depending on the panel. Within panels the SIPP is longitudinal, but each panel draws a new nationally representative sample of 14,000 to 52,000 households. In some periods, there have been overlapping panels (i.e., with more than one in the field at the same time), but at other times only one panel was in the field. The SIPP size and panel length changed several times since 1984, with the most recent redesign occurring in 2014, though the SSB/GSF stops with the 2008 panel. The Census Bureau takes *each individual* in a SIPP household in the 1984, and 1990 to 2008 SIPP panels and links them to their Detailed Earnings Records and Master Benefits Records presuming they have the necessary ID to be linked. The link uses SSNs to link individuals in the survey to the administrative records. The linked data are compiled by the Census Bureau, and are officially referred to as the SIPP Gold Standard File (GSF). The GSF is available in FSRDCs.

The measure of earnings that we use comes from the DER (not the SIPP survey) and 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). Earnings are not top coded after 1978, and run through 2014. Because of how individuals are matched to administrative data, the GSF includes complete earnings histories for anyone who can be matched, including periods of zero (taxable) earnings. 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 addition to the administrative earnings records, the Census Bureau has included basic demographic and human capital variables, marriage histories, fertility histories, as well as self-reported earnings and work hours from the SIPP survey. Variables collected in the SIPP panels that are not linked to administrative data cover only the years of the individual's SIPP panel.

From the GSF, a synthetic data file, the SSB, is created by applying sequential regression multiple imputation to the GSF (Raghunathan et al., 2001), with four implicates on the data file. The general method and the general types of models used in the creation are described in the papers cited in the introduction, but the details of the model and what variables are used have not been released by the Census Bureau. The current SSB only has 141 variables, and hence is only a small fraction of those in the SIPP survey. About a third of the variables are drawn from the administrative data, with the rest consisting of demographic characteristics of the household and a few income amounts reported on the survey. The SSB is partly synthetic in the sense that each household in the original data is represented once in the SSB (i.e., it is not intended to represent a larger

or smaller population), but is fully synthetic in the sense that all 141 variables are synthesized, including those which are on the SIPP survey public use file. This makes it difficult to link the SSB to the publicly available SIPP files.

The Census Bureau offers SSB users the option of checking their results against the GSF, using a validation server. Users submit their programs to Census, Census runs the programs, and returns the results to the user after going through a disclosure review process that is now identical to what FSRDC users go through. Our paper is based on such comparisons between our SSB estimates and those from the GSF validation server.[3] Until Fall 2022, the SSB was hosted on the VRDC server at Cornell University. It was available to any researcher after going through a nominal application process to confirm that the proposed analysis could be carried out on the GSF/SSB and to set up access to the server. Researchers could then carry out their analysis on the SSB. Once an analysis was complete, researchers could choose to either base conclusions on the SSB, or have their results validated on the GSF. The latter required going through the same disclosure avoidance protocol as FSRDC researchers. If a researcher chose to have their SSB results validated on the GSF, the only difference between accessing the GSF through an FSRDC versus the SSB is whether it was possible to know what the disclosed results would show. FSRDC researchers already know what the disclosed results will show, while SSB users were reliant on the extent to which the SSB matched the GSF. The SSB server at Cornell was shut down in September, 2022.

Publicly available methodological details on how the SSB is created are scarce. Further, the specific methods used have changed over time. However, the general method for the last version of the SSB is described in (Benedetto et al., 2018). As we understand it, the basic process is as follows. All individuals from all SIPP panels are pooled together into a single data set. Then each

[3]This analysis was first performed using the SIPP Synthetic Beta (SSB) on the Synthetic Data Server housed at Cornell University which was funded by NSF Grant #SES-1042181. Final results for this paper were obtained from a validation analysis conducted by Census Bureau staff using the SIPP Completed Gold Standard Files and the programs written by these authors and originally run on the SSB. One set of results is covered by review #CBDRB-FY21-095, other results predate the requirement that disclosure reviews go through the full DRB. The validation analysis does not imply endorsement by the Census Bureau of any methods, results, opinions, or views presented in this paper.

administrative variable for each calendar year is modeled based on the GSF, and predicted for the SSB. The model is variable specific, depending on the type of variable (e.g., continuous, categorical, binary) and the shape of the distribution (e.g., continuous variables that are not normally distributed are transformed to a normal distribution before imputing). For our purposes, there seem to be four main aspects of this process that are relevant. First, no ex-post adjustments are made to imputed variables to ensure that the distribution of the SSB variable matches the GSF. Second, to construct total earnings in the SSB and GSF users must sum two separate earnings components–total annual FICA earnings and total annual non-FICA earnings–that are imputed separately, thus total earnings is the sum of two imputed variables and is not itself imputed. Third, all observations for a given calendar year are imputed together. This could mean that, even if the distribution of a given variable is similar across the SSB and GSF for the full sample, the distribution within any given subsample may differ. This is particularly relevant when defining subsamples based on SIPP survey characteristics that may not be fixed through time, which we describe in more detail below. Finally, earnings are not normally distributed. Even after applying a transformation to make it normally distributed, the imputation process may struggle with the extreme areas of the left and right tails.

1.3 Sample Definitions

Our baseline sample consists of men who can be matched to the DER age 25 to 59 with positive labor earnings, where earnings are defined as above. In each year, the sample is drawn from pooled SIPP panels. We focus on men largely because we have more complete results from both the SSB and the GSF for men than women. We perform two sets of subgroup analyses, one by race and a second by education. Data on both race and education are taken from the SIPP survey, meaning they are not observed outside the time period of a given individual's SIPP panel. We treat race as fixed through time both prospectively and retrospectively from an individual's panel, thus the combined sample that underpins the analyses by race does not differ substantively from the

baseline sample. Education, measured as the highest degree attained, can not reasonably be treated as fixed through time. Here, we further restrict the sample to men who were at least 25 at the time of the SIPP panel. Note that this sample–men who are 25 to 59 in year t *and* who are at least 25 at the time of the SIPP panel–deviates considerably from the sample used to construct the SSB.

2 Methods

By construction, earnings of individual i in year t is the sum of permanent earnings ($\lambda_t \mu_i$) and a transitory earnings shock (ν_{it}) which, in this simple model, is assumed to be independent of μ_i, as given in Equation (1).

$$y_{it} = \lambda_t \mu_i + \nu_{it} \tag{1}$$

If permanent earnings change over time, it is due to changes in λ_t. Typically y_{it} is measured in logs, as we do here.

The variance of earnings is then the sum of the variance of the permanent and transitory components of earnings:

$$\sigma^2_{y_{it}} = \lambda_t^2 \sigma^2_\mu + \sigma^2_{\nu_t}. \tag{2}$$

The large literature on transitory instability proceeds from this basic model in two directions. The first uses changes in earnings over short time horizons to measure gross volatility or the variance of the change in y_i over one or two years. The second relies on more complex models of the earnings generating process to identify $\sigma^2_{\nu_t}$ and σ^2_μ. The former is generally referred to as volatility, while the latter is referred to as variability or error components models. We note a relationship between the two below.

The first of these two approaches, volatility, measures the variability of changes in the left hand

side of Equation (1) differenced over short time periods, as given in Equation (3).

$$\text{Var}(y_{it} - y_{it-\tau}) = (\lambda_t - \lambda_{t-\tau})^2\sigma^2_\mu + \sigma^2_{\nu_t} + \sigma^2_{\nu_{t-\tau}} \tag{3}$$

$$= \sigma^2_{y_{it}} + \sigma^2_{y_{i,t-\tau}} - 2 * \text{Cov}(y_{it}, y_{i,t-\tau}) \tag{4}$$

where $\tau = 1$ in this case. From the first line we see that volatility is the sum of the variance of the permanent component and of two transitory variances, but it is also clear that if the permanent variance is not changing, then volatility equals the sum of two transitory variances and and hence will track the transitory variance in an error components model well. The second line of the equation illustrates an alternative way of thinking about volatility, shown in Equation (4), uses the definition of the variance to highlight the relationship between cross-sectional inequality and volatility. It is straightforward to see that if volatility differs through time, across data sets, or across samples, it must either be due to differences in the cross-sectional distribution of earnings or the covariance structure of earnings over short time horizons. We will make use of this decomposition to help identify sources of differences in volatility between the SSB and the GSF.

An alternative to Equation (3) is to measure volatility using the variance of the arc change in earnings (Dahl et al., 2011; Ziliak et al., 2011). The arc-change is calculated as

$$\text{Var}\left\{ \frac{Y_{it} - Y_{i,t-\tau}}{\frac{|Y_{it}|+|Y_{i,t-\tau}|}{2}} \right\} \tag{5}$$

where Y_{it} is the level of earnings for individual i at time t. We rely on both log changes and arc changes. The arc change can also be decomposed in a way that is analogous to Equation (4), but it is considerably more complex and thus more difficult to interpret so we do not make use of it here.

There are two primary differences between the arc- and log-change methods. The first is that the arc change allows for the inclusion of time periods with zero earnings in either t or $-\tau$, though not both. The second is that the arc change is bounded between -2 and 2. The boundedness of the

arc change and the unboundedness of the log change means that they place different weights on the distribution of earnings changes.

The literature that relies on formal decompositions to identify the components of Equation (2) argues that, in the presence of time trends in the returns to permanent characteristics, time trends in the transitory earnings variance, shocks to permanent earnings, age-specific shocks to permanent and transitory earnings, and serial correlation in transitory shocks, trends in earnings volatility and trends in the transitory variance from the error components model of earnings may not be the same. In particular, earnings volatility will include some of the variance of the permanent component of earnings both because the cross-sectional variance of earnings in t and/or $t-1$ reflect both transitory and permanent shocks and because the covariance includes serially correlated transitory shocks and permanent earnings. Shin and Solon (2011) argue that earnings volatility is still a useful measure because increases in the variance of the transitory component of earnings are likely to be accompanied by increases in earnings volatility and this measure is less dependent on specific parametric assumptions.

The simplest versions of error components models separate the variance in the permanent and the transitory component of earnings in Equation (1) by considering the distribution of short-run deviations in earnings from an individual-specific long-run mean. Moffitt and Gottschalk (2012) call this method window averaging, and estimate it using random effects. But, other approaches that are similar in spirit also exist in the literature (Debacker et al., 2013; Kopczuk et al., 2010). This technique tends to overstate the permanent component of earnings particularly in the presence of serial correlation in transitory shocks.

Over time, the literature has developed to model increasingly flexible specifications of earnings dynamics. Among the important features that have been captured are individual specific growth factors in permanent earnings, permanent earnings that evolve over the life cycle, serial correlation in transitory earnings, age-related heteroskedasticity in transitory earnings, and year-specific factor loadings for both permanent and transitory earnings (Baker and Solon, 2003; Debacker et al., 2013;

Haider, 2001; Moffitt and Gottschalk, 2012; Moffitt and Zhang, 2018).

We rely on a newly developed model presented in Moffitt and Zhang (2018). The primary advantage of this model is that it significantly relaxes the parametric assumptions underlying the earnings generating process. The model builds on the same basic approach as shown in Equation (1), but extends the model to incorporate age specific permanent and transitory earnings that can both vary through time. Specifically, the model is given as

$$y_{iat} = \lambda_t \mu_{ia} + \beta_t \nu_{ia} \tag{6}$$

$$\mu_{ia} = \mu_{i0} + \sum_{s=1}^{a} \omega_{is} \tag{7}$$

$$\nu_{ia} = \epsilon_{ia} + \sum_{s=1}^{a-1} \psi_{a,a-s} \epsilon_{i,a-s} \text{ for } a \geq 2 \tag{8}$$

$$\nu_{i1} = \epsilon_{i1} \text{ for } a = 1 \tag{9}$$

for $a = 1, ..., A$ and $t = 1, ..., T$. As before, 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, as shown in Equation (8), 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, given in Equation (9), 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. Intuitively, identification rests on the assumption that permanent shocks to earnings are permanent. Any shock that does not pass fully and permanently into earnings must therefore be transitory. Transitory shocks are estimated flexibly within the class of linear models.

3 Baseline Results

3.1 Descriptives

Table (1) shows the basic demographic characteristics of the GSF and SSB samples. In each case we include men 25-59. The GSF (SSB) Matched is the subset who can be matched to earnings records and GSF (SSB) Volatility are those who have positive earnings in two consecutive years. As we would expect the GSF and SSB are very similar in mean characteristics.

As shown in Equation (4), volatility in log changes is a function of the variance of log earnings. As such the level and trend of volatility is affected by the level and trend in inequality (Moffitt et al., 2023). Figure (1a) shows that the variance of log earnings is always higher in the SSB than in the GSF and that the difference between the two series (plotted at the bottom) grows over time in absolute terms. Figure (2) shows the percentile points of the earnings distribution at the top and bottom in the SSB and the GSF. The SSB has a higher density of low earnings, it also has a longer right tail which is evident at the 95th percentile of the earnings distribution and above. Differences between the SSB and GSF in the tails of the earnings distribution is consistent with Stanley and

Table 1: Demographic Characteristics of the SIPP GSF and SIPP SSB Samples

	GSF			SSB		
	All	Matched	Volatility	All	Matched	Volatility
$<$ High School	0.185	0.165	0.141	0.179	0.166	0.137
High School	0.305	0.302	0.305	0.325	0.317	0.318
Some College	0.264	0.273	0.283	0.295	0.303	0.319
College	0.155	0.164	0.172	0.128	0.134	0.141
College+	0.090	0.096	0.100	0.074	0.081	0.085
White	0.725	0.750	0.771	0.719	0.743	0.759
Black	0.116	0.108	0.099	0.122	0.114	0.108
Other	0.053	0.050	0.046	0.054	0.051	0.048
Hispanic	0.107	0.092	0.083	0.105	0.092	0.086
Age	40.380	40.720	40.150	40.525	40.840	40.260

Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Full sample is all men age 25 to 59. Matched GSF (SSB) is all men age 25 to 59 who can be matched to administrative records. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use one implicate.

Totty (2021).

3.2 Instability in Earnings

3.2.1 Volatility

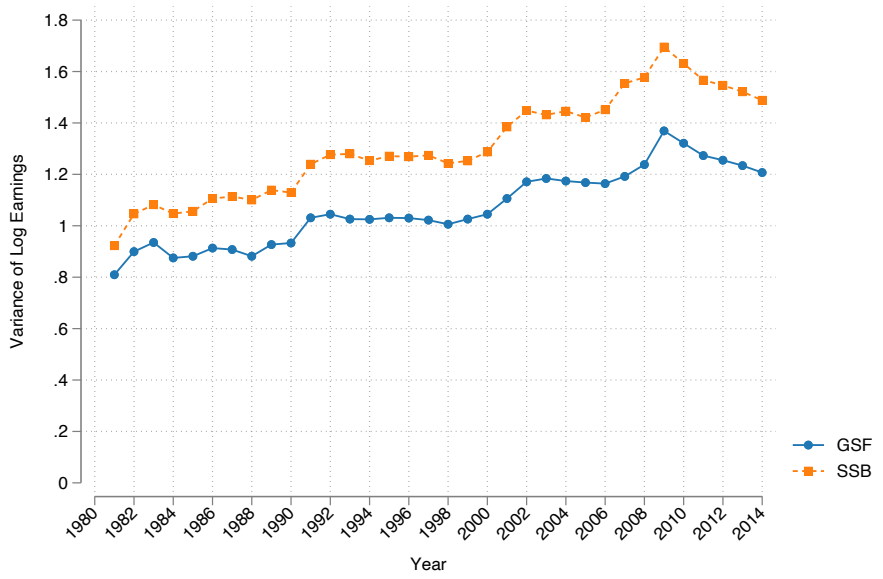
In our benchmark set of results, we further limit the sample of men age 25-59 to exclude the bottom 1% of earnings in year t ($t-1$). Figures (3) and (4) show trends in volatility in the GSF and SSB using log and arc changes, respectively, using this trimmed earnings distribution. For both measures of volatility the SSB has a higher level of volatility. When measuring volatility in log changes, the trends in the two series are slightly different though not meaningfully so. The gap in volatility between the SSB and the GSF is stable when measuring volatility in arc changes.

3.2.2 Decomposing Volatility

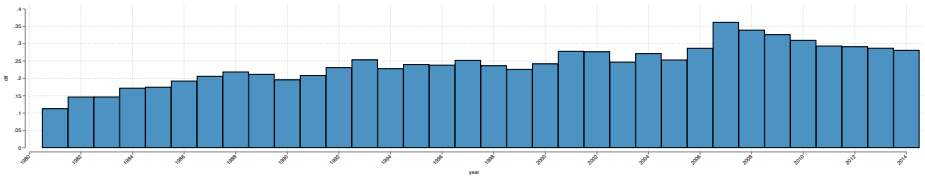
To understand the source of higher volatility in the SSB relative to the GSF, we examine trends in the variance of earnings and covariance of earnings over a two-year period in Figure (5) and (6). As

Figure 1: Variance of Log Earnings

(a) Levels

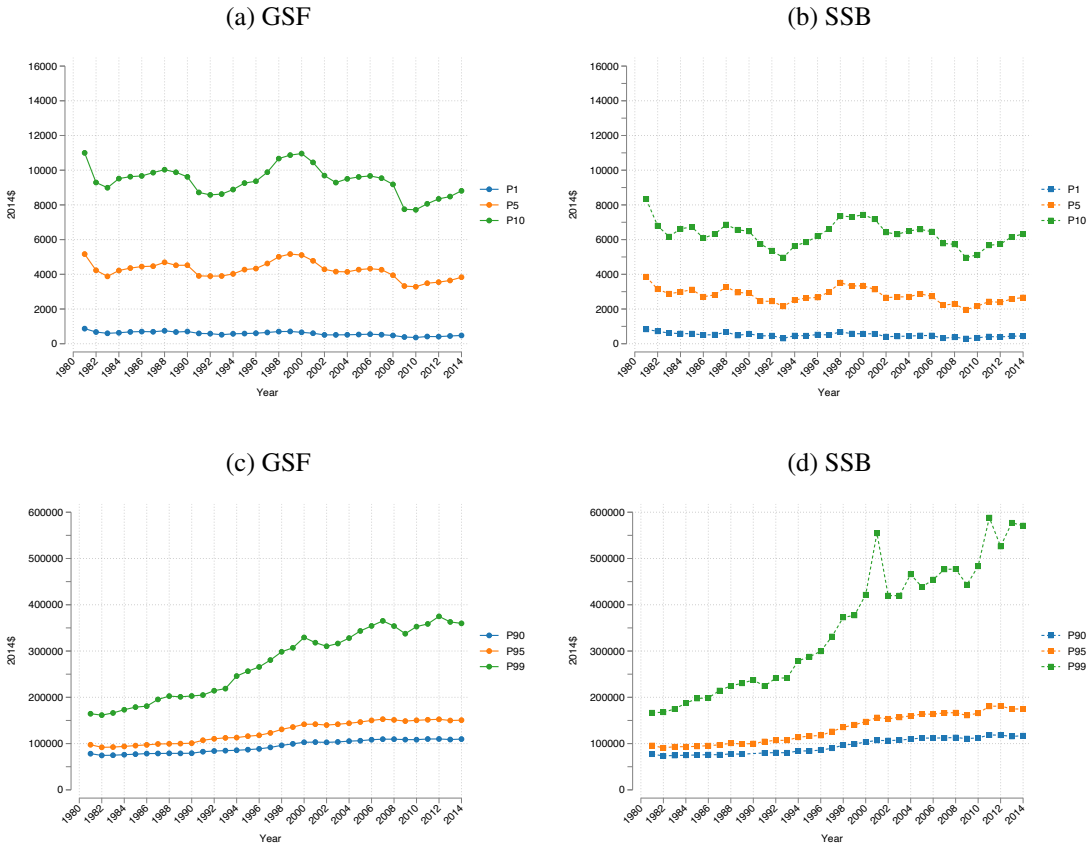


(b) Differences



Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use average across four implicates.

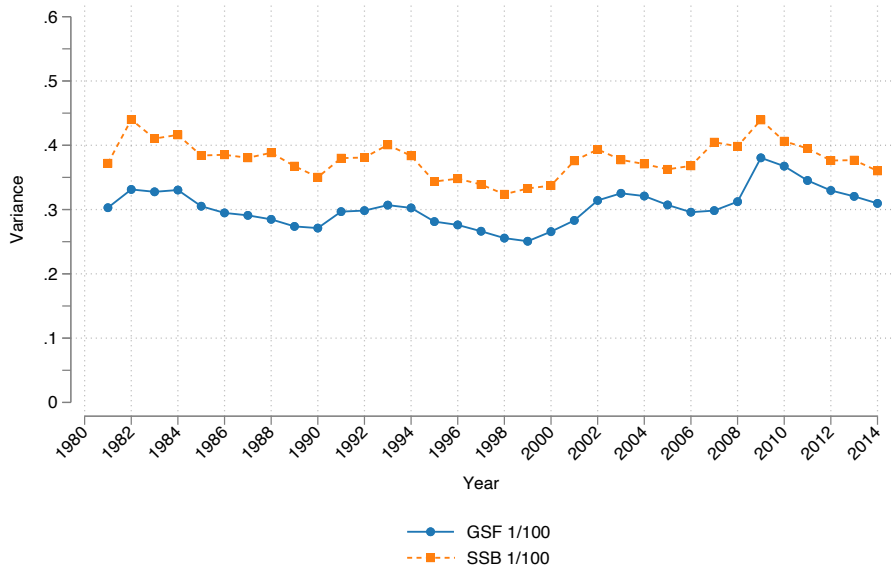
Figure 2: Selected Earnings Percentiles



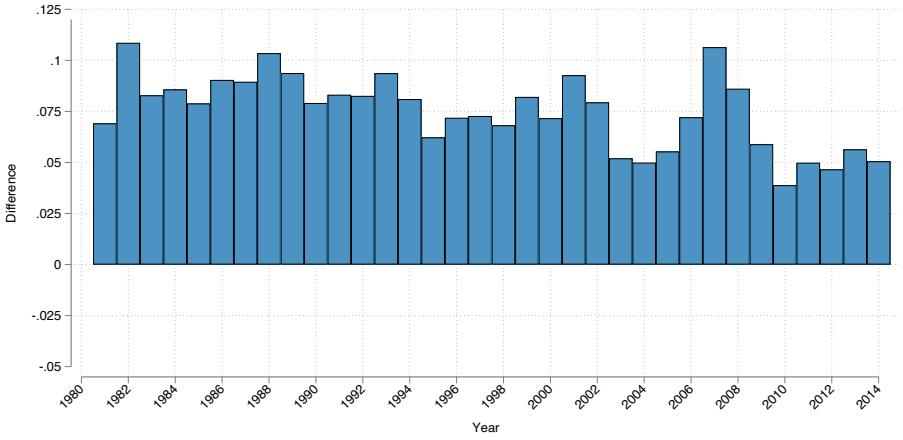
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use one implicate.

Figure 3: Volatility in Log Changes

(a) Levels



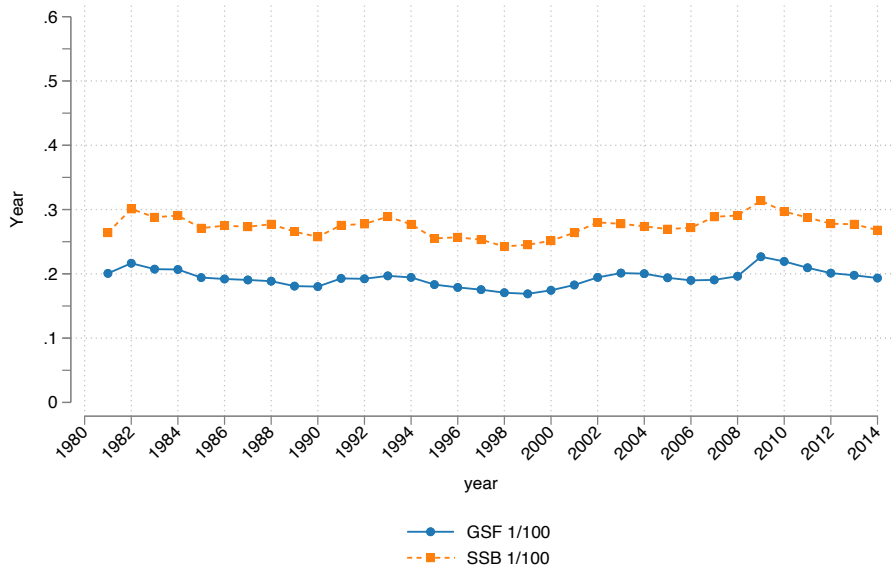
(b) Differences



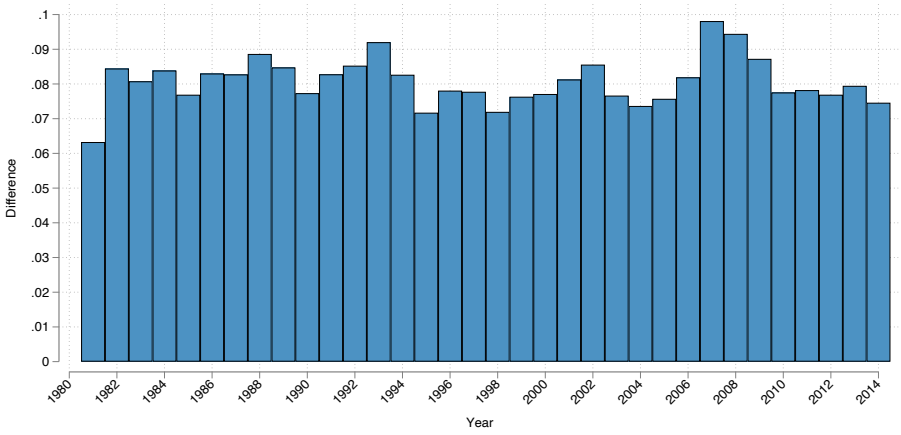
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use average across four implicates.

Figure 4: Volatility in Arc Percent Changes

(a) Levels



(b) Differences

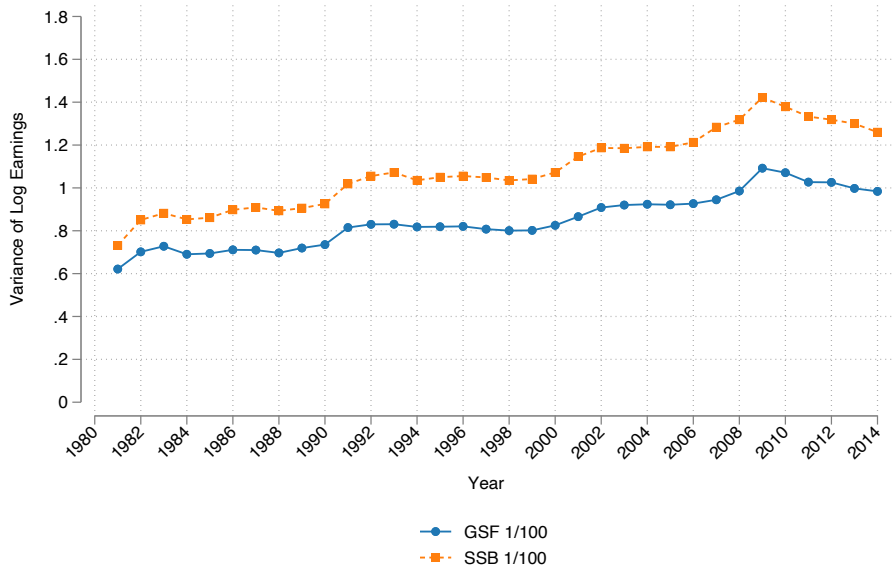


Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use average across four implicates.

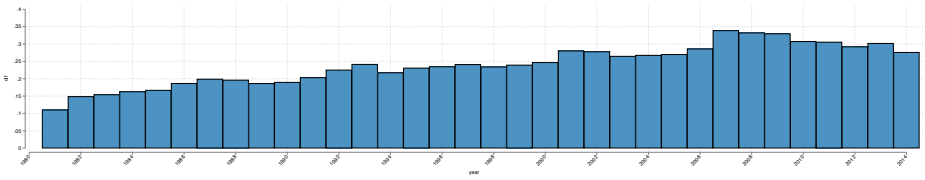
we saw with total inequality in Figure (1a), Figure (5) shows that the variance of earnings with the 1% trim is higher in the SSB than in the GSF and the gap is growing over time. Figure (6) shows that the covariance of earnings is also higher in the SSB than in the GSF and, like the variance, that gap is growing over time. These patterns offset to generate relatively stable differences in volatility measured in log changes between the SSB and GSF, with the SSB always being higher than the GSF.

Figure 5: Variance of Log Earnings 1% Trim

(a) Levels



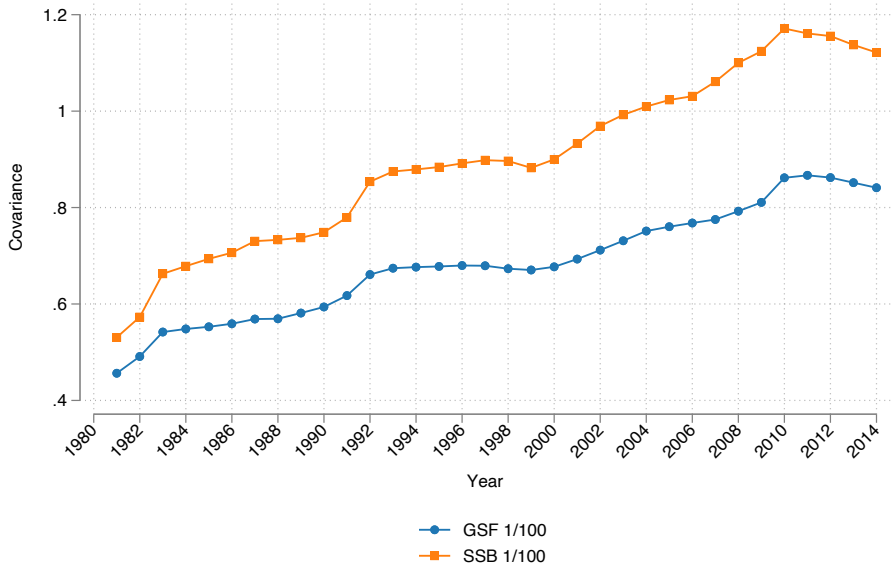
(b) Differences



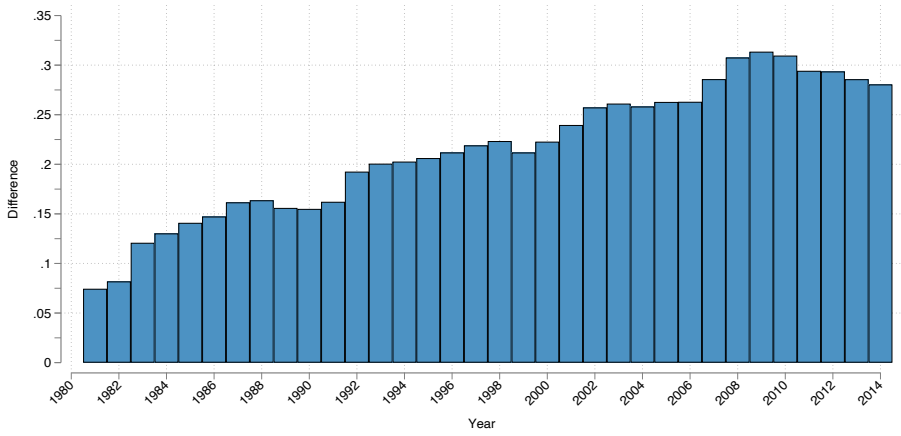
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use average across four implicates.

Figure 6: Covariance of Log Earnings 1% Trim

(a) Levels



(b) Differences



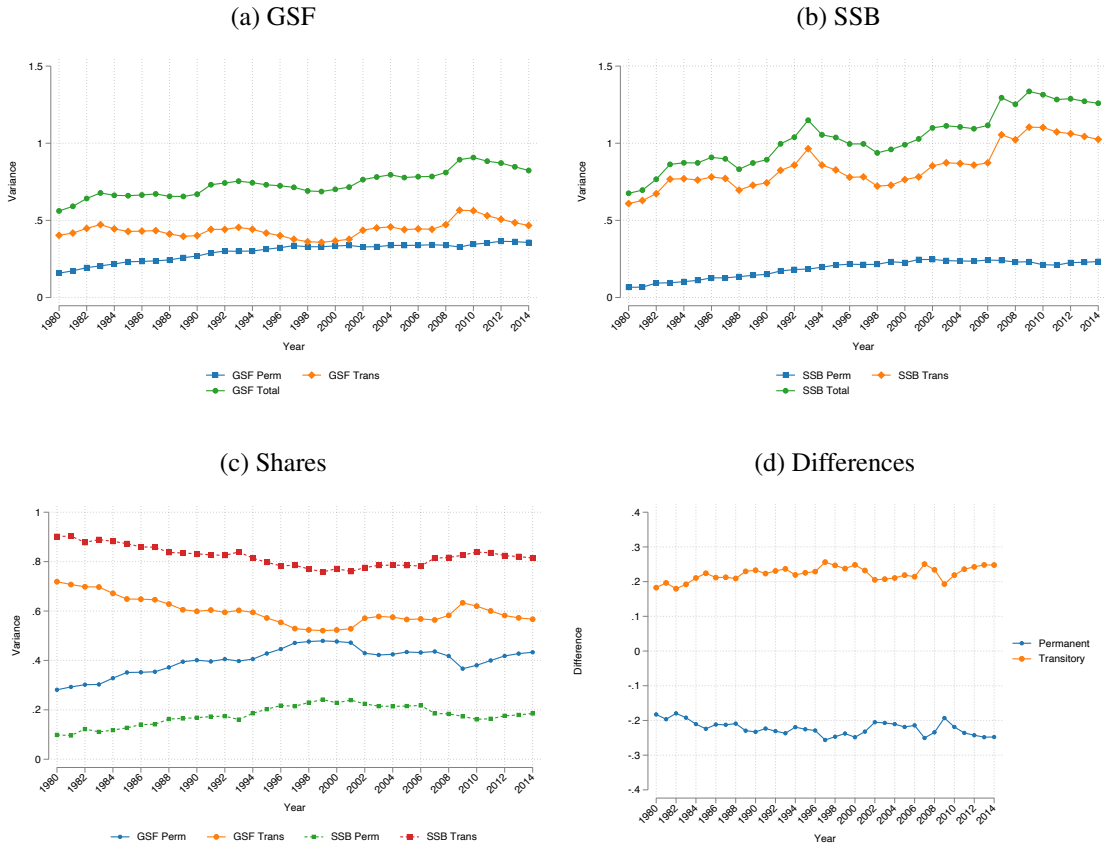
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use average across four implicates.

3.2.3 Error Components Model

We examine results of the ECM model run on the SSB and GSF, where the dependent variable is residual earnings from a regression of log earnings on controls for age and education. Figure (7a) and (7b) show the predicted values of the total variance of earnings as well as the permanent and transitory components for men age 40-49. Two clear patterns emerge. First, the variance of earnings implied by the ECM model is higher in the SSB than in the GSF. This replicates the patterns in the raw data. The decomposition into permanent and transitory components is also quite different. The level and upward trend of the transitory variance is much higher in the SSB than in the GSF. Figure (7c) shows that this implies that the share of the total variance that is accounted for by the permanent component of earnings is larger in the SSB than in the GSF.

That permanent inequality is lower in the SSB than the GSF, but the covariance of earnings is higher in the SSB than the GSF does seem to conflict with each other. However, three caveats are important. We decompose residual (log) earnings where education differences have been removed, while when we decompose volatility we use the level of (log) earnings. Presumably, the covariance would also decrease if we used residual earnings, though by how much we do not know. Second, the covariance of earnings between any $t-1$ and any t is, in the ECM model, a result both of the variance of the permanent component (σ_μ^2) and of the covariance of the transitory components across the two periods. We know the latter is higher in the SSB than in the GSF. With the higher covariance in the SSB being picked up by the transitory component, less is allocated to the permanent component. The separation between permanent and transitory components may be model dependent. This model specifies that permanent shocks follow a unit root process but other models make different assumptions. We have not tested a comprehensive set of models. However, despite these caveats, one aggregate pattern is similar between the ECM model and the trends in aggregate inequality and gross volatility: total inequality is higher in the SSB than the GSF, and short-run instability is higher in the SSB than the GSF. The difference is one of degree.

Figure 7: ECM Decomposition 1% Trim



Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use one implicate.

3.3 Trimming Earnings

For reasons outlined above, Equation (4) makes clear that both the level and trend in volatility may be sensitive to how low earnings are handled. Specifically, Carr and Wiemers (2021) show that both the level and trend in volatility change depending on small differences in how low earnings are excluded from the analytical sample. Given that Figures (1a) and (2) show that both the level and trend in inequality differ between the SSB and the GSF, and Figure (6) shows that the covariance of earnings differs, it is possible that the effect of trimming low earnings differs between the two data sets.

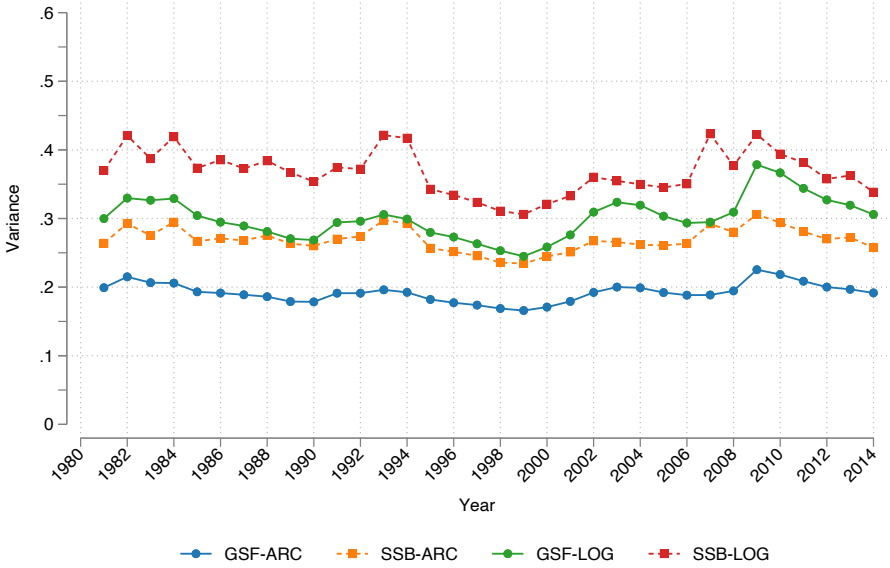
Figure (8) shows volatility in log and arc changes trimming the top and bottom 1% of earnings in t and $t-1$, respectively. Recall that in our main results, we trimmed only the bottom 1% of earnings. Additionally trimming at the top, where there is a substantial increase in density of high earnings in the SSB relative to the GSF, does not change the conclusions drawn above. Volatility is higher in the SSB than the GSF for both measures. In the arc change, the absolute difference is relatively stable over time, and the log change measure shows the same moderate convergence.

Figure (9) shows volatility in log and arc changes without any trimming. In this case, in arc changes we see the same pattern as before with volatility higher in the SSB than in the GSF and differences stable over time. In contrast to the other trims, in untrimmed earnings, volatility is similar in the SSB and the GSF though the trend is somewhat different with volatility in the SSB starting slightly higher and ending slightly lower. The similarity in levels is a consequence of offsetting variances and covariances. Figure (1a) shows that the variance of untrimmed earnings is higher in the SSB than in the GSF. Figure (10) shows that the covariance of earnings is also higher in the SSB than in the GSF and these two offset so that the level of volatility is similar in the two data sets when earnings are not trimmed.

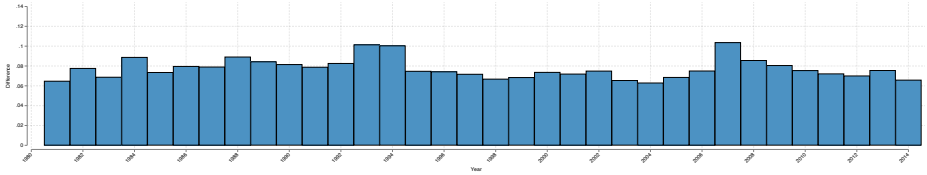
The existing volatility literature that has relied on administrative earnings data has generally only trimmed low earnings, leaving high earnings untrimmed. Carr and Wiemers (2021) investigate the effect of the different types of trims applied in this literature on the level and trend in volatility.

Figure 8: Volatility, Earnings Between 1% and 99%

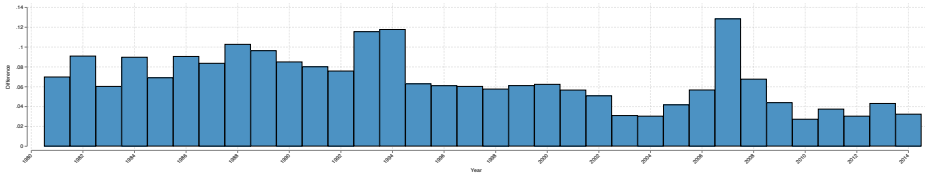
(a) Levels



(b) Arc-Differences



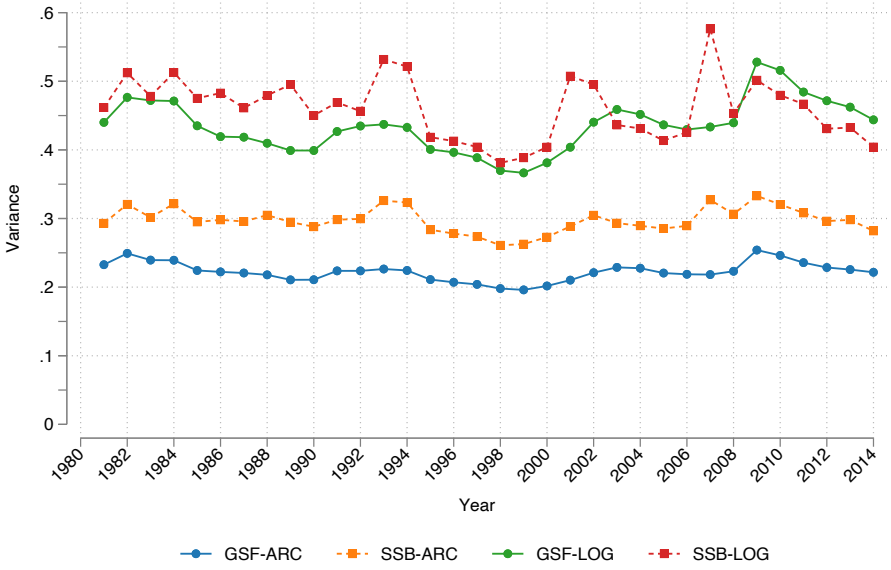
(c) Log-Differences



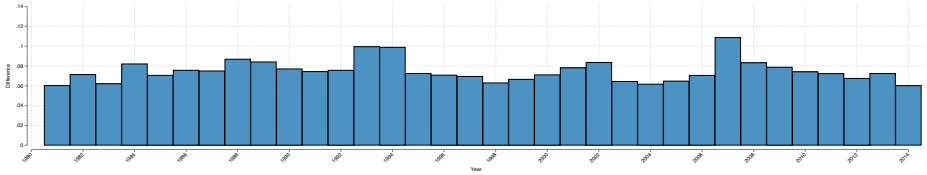
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use one implicate.

Figure 9: Volatility, Earnings Untrimmed

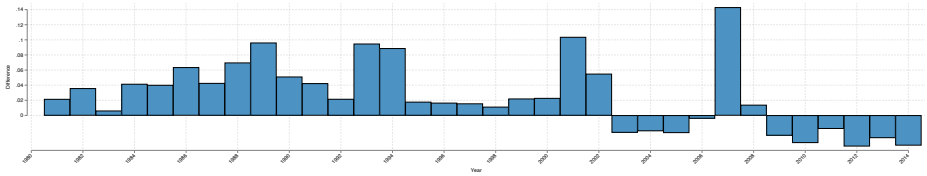
(a) Levels



(b) Arc-Differences



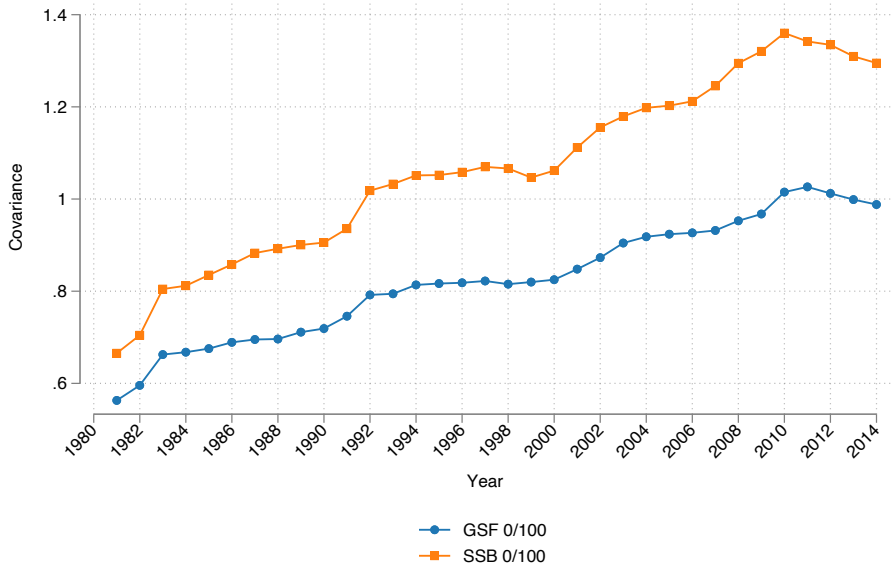
(c) Log-Differences



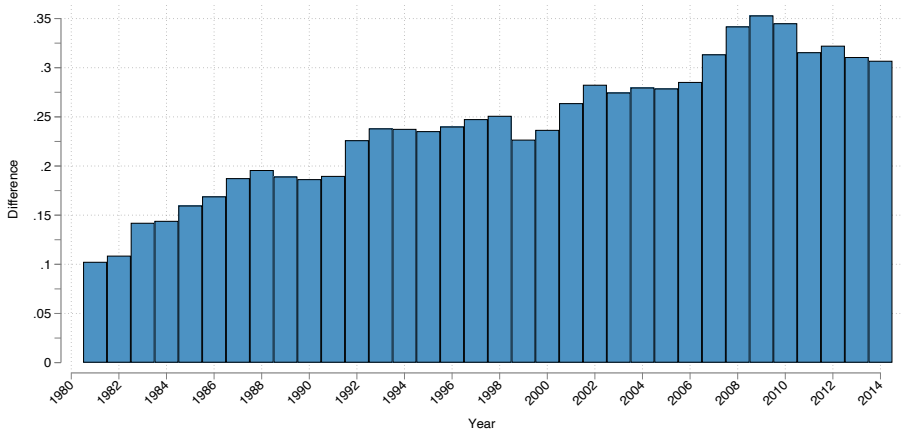
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use one implicate.

Figure 10: Covariance of Log Earnings Untrimmed

(a) Levels



(b) Differences



Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use average across four implicates.

For the most part, previous users of administrative data have trimmed on absolute dollar amounts not directly tied to the earnings distribution in any given year. Here we implement two of those: (1) we exclude earnings in year t ($t-1$) that are below one-quarter of full-time, full-year work at one-half of the federal minimum wage in t ($t-1$); and (2) earnings below the minimum earnings required to earn one year of credit towards Social Security eligibility.

Figures (11) and (12) show the results of applying these two trims, with the 1% trim reported earlier. In general, the absolute dollar trims will trim both a larger fraction of individuals at any given time and the fraction trimmed will increase over time because the density of low earnings is increasing in both the SSB and the GSF.[4] However, the percent of the sample trimmed will increase more in the SSB than the GSF because the density of lower earnings is both higher, and rising faster, in the SSB than the GSF.

Figure (11) shows the three trims applied to the arc change. As seen in Carr and Wiemers (2021) using absolute-dollar trims results in an overall downward trend in volatility, while a percentile point trim has a flat trend similar to untrimmed earnings. In arc changes, for each respective trim volatility is higher in the SSB than the GSF and trends are similar.

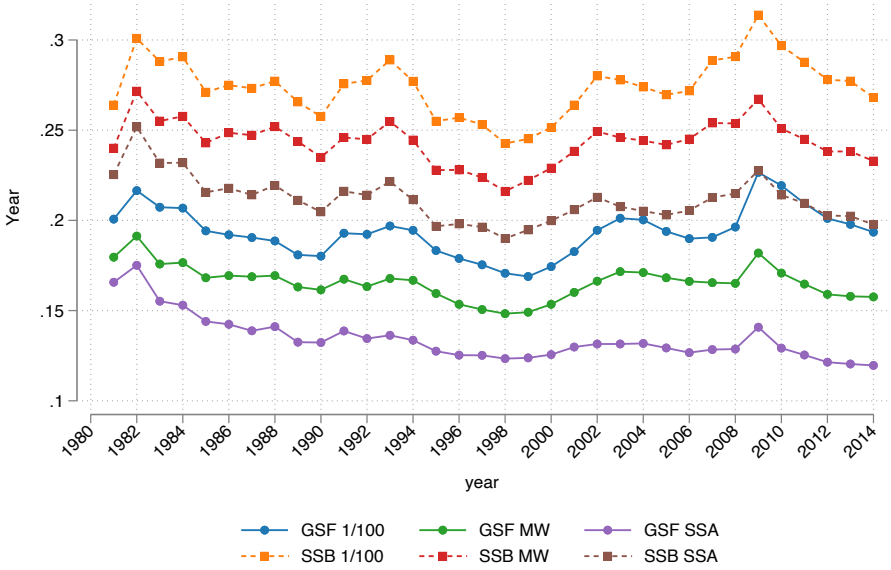
Figure (12) shows volatility for three selected trims in log changes. With the other trims, volatility is higher in the SSB than the GSF, though trends are more similar using the absolute dollar trims than the percentile point trims.

Similar to the analysis above, decomposing volatility into the variance and covariance of earnings can help identify sources of differences in volatility between the two data sets for each respective trim. Figure (13) shows the variance of earnings for each of the trims in both the SSB and the GSF. For each trim, inequality is higher in the SSB than the GSF, and increases at a faster rate. Figure (14) shows the covariance of earnings between t and $t-1$ for each respective trim. Covariances are always higher in the SSB than the GSF, and the absolute difference between the

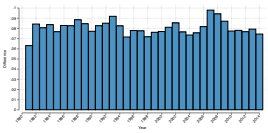
[4]The minimum wage trim excludes earnings below between $1500 and $1900 is 2014 dollars depending on the year. The SSA trim increases steadily from $3100 to $4550 in 2014 dollars.

Figure 11: Trimming: Arc Change

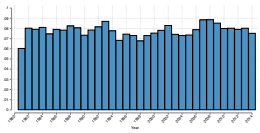
(a) Levels



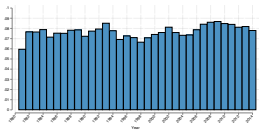
(b) Diff: 1%/100%



(c) Diff: MW



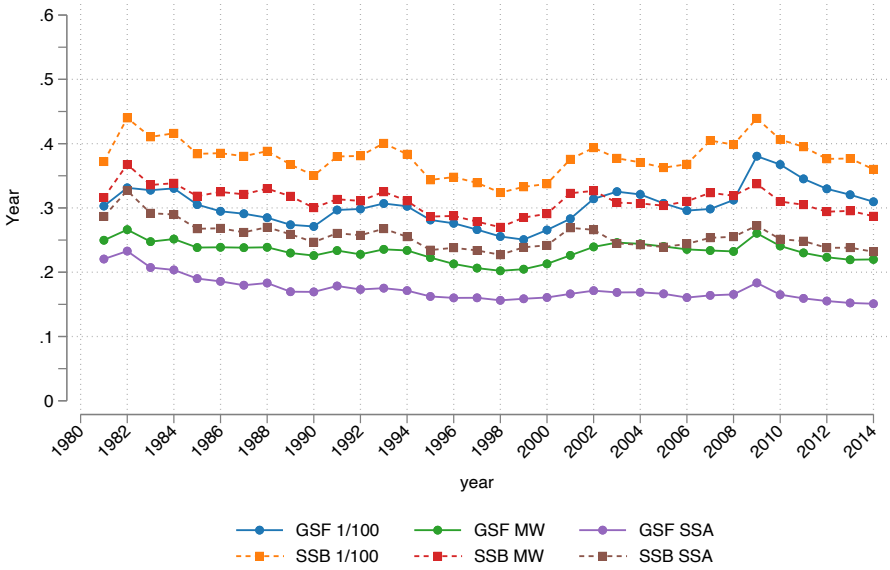
(d) Diff: SSA



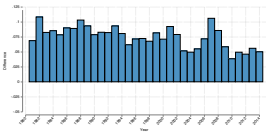
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use four implicates.

Figure 12: Trimming: Log Change

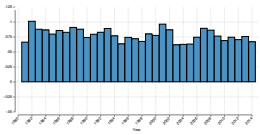
(a) Levels



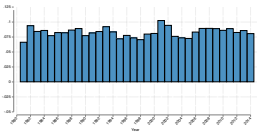
(b) Diff: 1%/100%



(c) Diff: MW

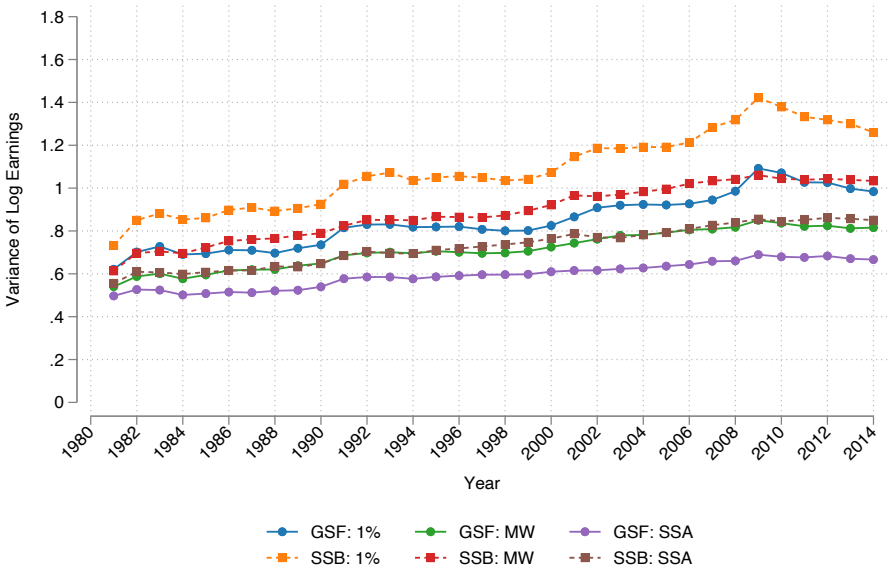


(d) Diff: SSA



Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use four implicates.

Figure 13: Variance of Log Earnings, Selected Trims



Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use four implicates.

SSB and the GSF for each trim grows over time.

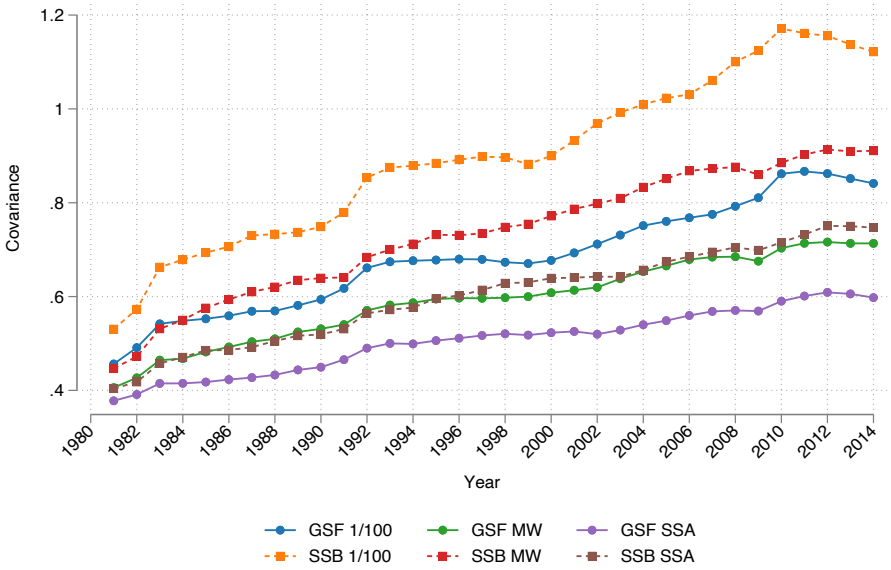
Combined, these results show that the similarity in volatility in untrimmed earnings between the SSB and the GSF is due to the two components of volatility offsetting each other. Inequality is higher in the SSB than the GSF, but the covariance is also higher, resulting in volatility that is similar. For other trims, and for the arc-change measure, the higher covariance is not enough to offset the higher level of inequality seen in the SSB resulting in volatility that is higher in the SSB.

3.4 Subgroup Volatility

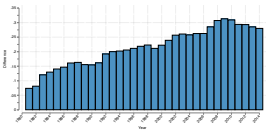
One of the primary advantages of these data is that the link to the SIPP provides demographic data that are otherwise not available with administrative earnings histories. We make use of this feature when estimating the ECM model so that we can decompose residual earnings as is typically done using survey data. Here, we make use of this feature to analyze volatility by race and education subgroups. The level and trend in volatility may differ by subgroup either because within-group

Figure 14: Volatility Decomposition: Estimated Covariances, Selected Trims

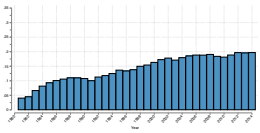
(a) Levels



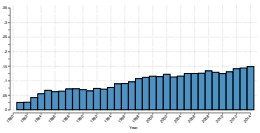
(b) Diff: 1%/100%



(c) Diff: MW



(d) Diff: SSA

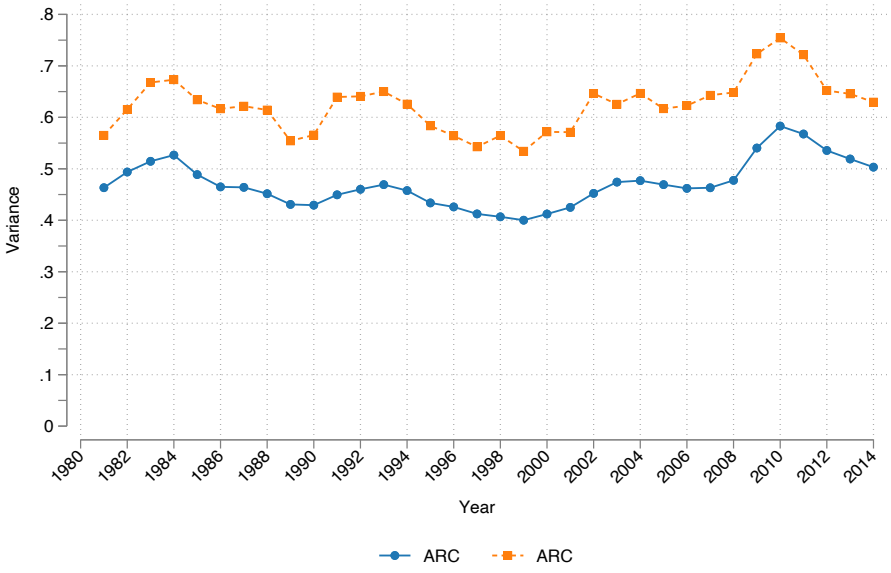


Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use four implicates.

inequality differs, or because the within-group covariance of earnings differ. Because labor supply differences at the extensive margin are particularly salient for race and education subgroups, we include zero earnings in either t or $t-1$ and use only the arc-change measure of volatility.

To put subgroup volatility in context, Figure (15) shows volatility including zeroes using the arc-change measure. As is typical, volatility is higher when including zeroes. Volatility is higher in the SSB than the GSF, but similar to when zeroes are excluded, the two trends are roughly parallel.

Figure 15: Volatility: Arc Change with Zeroes

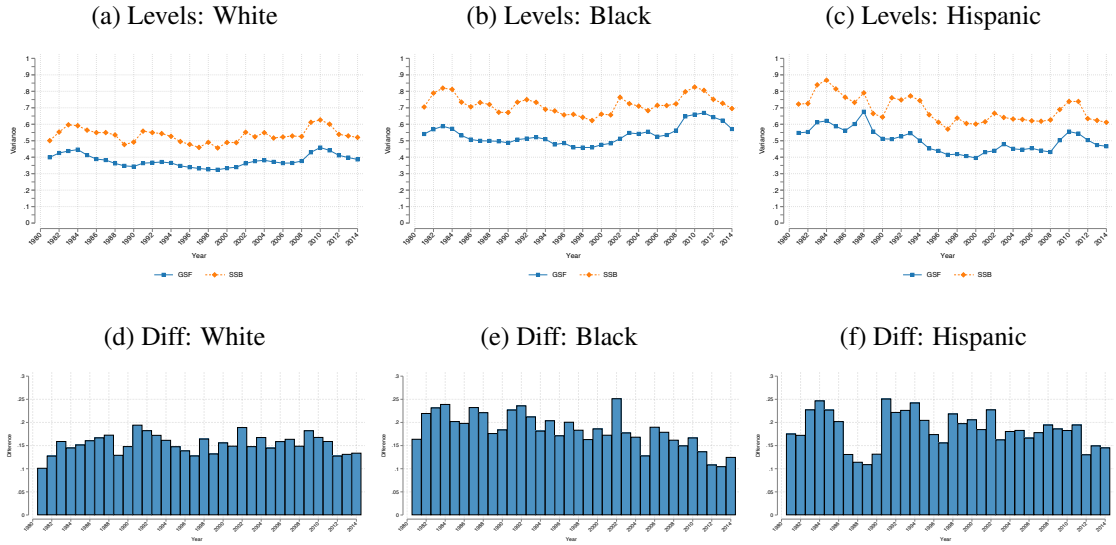


Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use one implicate.

Figure (16) shows arc-change volatility, including zeroes, for individuals who identify as non-Hispanic White, Black, and Hispanic. When we include zeroes, volatility for each group is higher in the SSB than the GSF, similar to the overall pattern for the arc change seen earlier. For both non-Hispanic White and Hispanic men, the level differences in volatility are relatively stable through time. For Black men, however, volatility in the two data sets trends in different directions. In the early to mid 1980s, volatility is 0.2 to 0.23 higher in the SSB than the GSF for black men, but by 2012/13 volatility is just over 0.1 higher in the SSB than the GSF. This is because volatility

increases between the late 1990s and 2012 for Black men in the GSF, while it is flat in the SSB.

Figure 16: Volatility by Race

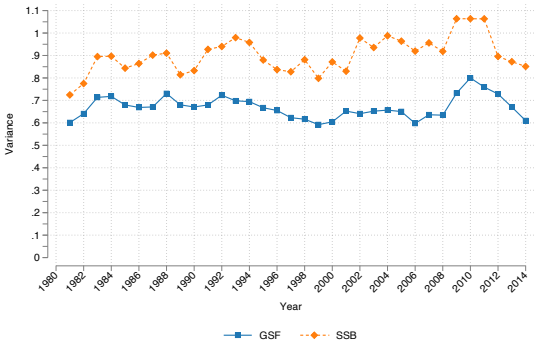


Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use one implicate.

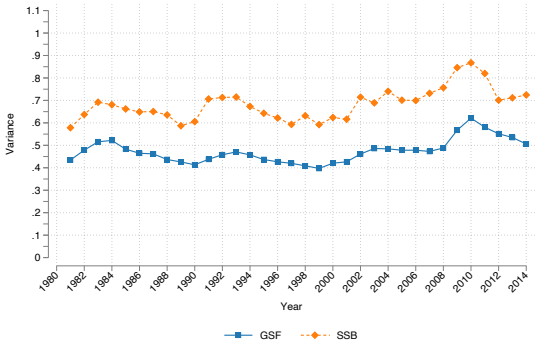
Figure (17) shows volatility by education. Recall from our discussion of the ECM results that using data on education requires making a second sample restriction based on the age at which an individual was interviewed in the SIPP so that we can plausibly treat education as fixed through time. Again, using the arc change including zeroes, volatility is generally higher in the SSB than the GSF. Both trends and levels vary between the SSB and the GSF, though levels vary more than trends. For individuals with high school or less education, volatility is between 0.1 and 0.34 higher in the SSB than the GSF, with a broadly increasing difference between the SSB and GSF. For those with some college, volatility is between 0.15 and 0.27 higher in the SSB, with little trend in the difference. For those with a college degree, volatility is between 0.1 and 0.21 higher in the SSB with what looks like a small upward trend in the difference through 2010. Finally, for those with an advanced degree volatility is mostly between 0.1 and 0.18 higher in the SSB, again with what could be a slight upward trend in the difference.

Figure 17: Volatility by Education

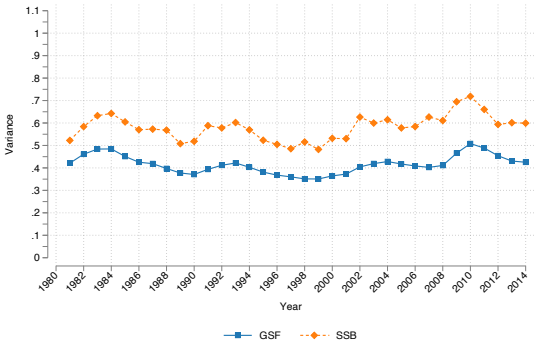
(a) Levels: HS



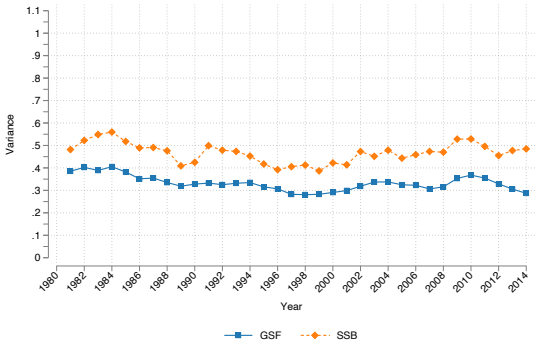
(b) Levels: SC



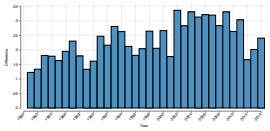
(c) Levels: BS



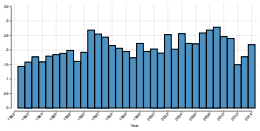
(d) Levels: Adv



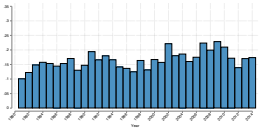
(e) Diff: HS



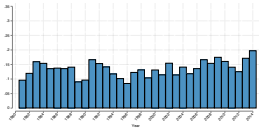
(f) Diff: SC



(g) Diff: BS



(h) Diff: Adv



Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB figures use one implicate.

4 Conclusions

These results suggest that the SSB does not capture fully the earnings dynamics in the GSF. In terms of measures of gross volatility, the SSB usually shows higher gross volatility than the GSF though the differences between data sets depend on the measure of volatility and the way in which low earnings are trimmed. In arc changes, the trends in volatility over time are preserved but the level of volatility is substantially higher in the SSB than in the GSF. In log changes, the difference in the level of volatility in the GSF and SSB is more sensitive to how earnings are trimmed and there are slight differences in the trends between the two data sets. When we decompose volatility in log changes into cross-sectional inequality and the covariance of earnings across a two year period, we find that both cross-sectional inequality and the covariance of earnings over two years is higher in the SSB than the GSF, regardless of how earnings are trimmed. These analysis show that even when volatility in log changes is similar in the SSB and GSF, as it is when earnings are untrimmed, the similarity is not the result of matching cross-sectional inequality or the short-run covariance of earnings.

The difference between the two data sets is the most pronounced when estimating the error components model with the SSB having a higher level of inequality and attributing more of the total variance of earnings to the transitory component. Despite total inequality being higher in the SSB, transitory inequality is so much higher in the SSB than the GSF that permanent inequality is lower in the SSB than the GSF. One reason for this may be that transitory earnings are more serially correlated in the SSB than in the GSF.

Finally, we show that in subgroup analysis, the differences between the two data sets overall are not always consistent across each subgroup. This appears to be the case when subgoups are split based on both time-varying and non-time-varying characteristics.

There are a few limitations to the current analyses. First, the analyses on the SSB were not created with the intention of comparisons to the GSF. Because of this, we have not averaged across

implicates for all of the results. We have noted in the table and figures notes when results are averaged across all implicates and when they are not. From what we can tell, the averaging does not affect the results we have presented (see Appendix Figure (A1)) but, because the SSB is no longer publicly available, we cannot adjust the estimates. Second, the differences between the SSB and the GSF in the results from the ECM model may reflect the specific model we chose to estimate. Other models are used in the literature and we cannot test how the two data sets perform on alternative specifications. Third, we have presented a set of results from one particular research question. The SSB was not synthesized to capture the earnings dynamics that we are estimating. In that sense, it is not clear that we would expect the results to be similar. Finally, we have assumed that the GSF represents the "truth" throughout but, as Stanley and Totty (2021) point out, this assumption may be flawed.

References

Abowd, John M, Martha H Stinson, and Gary Benedetto, "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project," 2006. US Census Bureau.

Baker, Michael and Gary Solon, "Earnings Dynamics and Inequality Among Canadian Men, 1976-1992," *Journal of Labor Economics*, 2003, *21* (2), 289–321.

Benedetto, Gary, Jordan C. Stanley, and Evan Totty, "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*, 2018.

_ **, Martha H Stinson, and John M Abowd**, "The Creation and Use of the SIPP Synthetic Beta," 2013. US Census Bureau.

Carr, Michael and Bradley Hardy, "Racial Inequality Across Income Volatility and Employment," *Oxford Research Encyclopedia of Economics and Finance*, 2022.

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

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

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

Dahl, Molly, Thomas DeLeire, and Jonathan Schwabish, "Estimates of Year-to-Year Volatility in Earnings and in Household Incomes from Administrative, Survey, and Matched Data," *Journal of Human Resources*, 2011, *46* (4), 750–74.

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

Gottschalk, Peter and Robert Moffitt, "The Growth of Earnings Instability in the US Labor Market," *Brookings Papers on Economic Activity*, 1994, *1994* (2), 217–272.

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

Haider, Steven J, "Earnings instability and earnings inequality of males in the United States: 1967–1991," *Journal of Labor Economics*, 2001, *19* (4), 799–836.

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

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

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

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

Raghunathan, Trivellore, "Synthetic Data," *Annual Review of Statistics and Its Application*, 2021, *8*, 129–140.

_ , James M. Lepkowski, and Peter Stolenberger, "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 2001, *27* (1), 85–95.

Rubin, Donald, "Statistical Disclosure Limitation," *Journal of Official Statistics*, 1993, *9* (2), 461–468.

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

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

Shin, Donggyun and Gary Solon, "Trends in Men's Earnings Volatility: What does the Panel Study of Income Dynamics Show?," *Journal of Public Economics*, 2011, *95* (7), 973–982.

Stanley, Jordan and Evan Totty, "A Penny Synthesized is a Penny Earned? An Exploratory Analysis of Accuracy in the SIPP Synthetic Beta," *U.S. Census Bureau Working Paper*, 2021, *CED-WP-2021-006.*

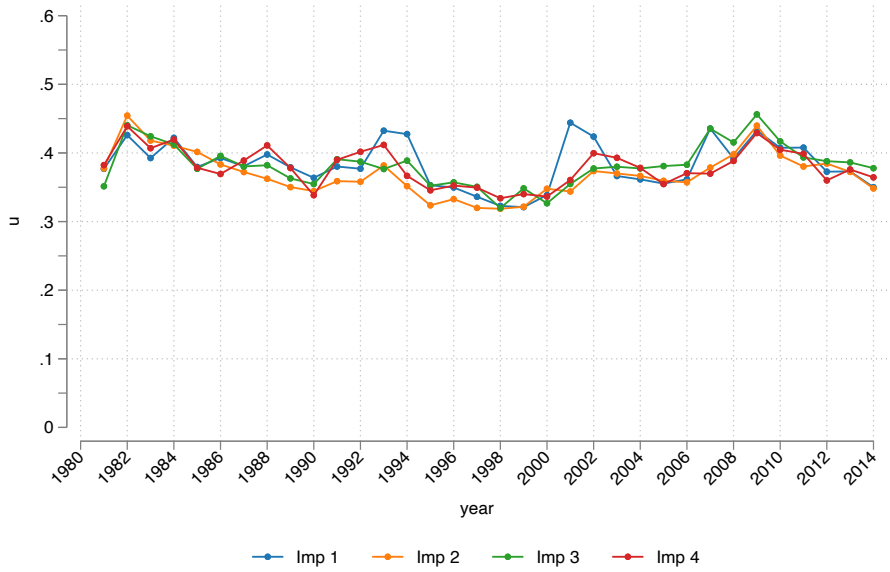
Ziliak, James P, Bradley Hardy, and Christopher Bollinger, "Earnings Volatility in America: Evidence from Matched CPS," *Labour Economics*, 2011, *18* (6), 742–754.

A Appendix

A.1 Number of Implicates

Figure A1: Volatility, Earnings Between 1% and 100%, 4 Implicates

(a) Log Changes



Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years.