HDSR
HARVARD DATA SCIENCE REVIEW

Volume 0 Issue 0
DOI: 00.000
ISSN: 2644-2353

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

Michael D. Carr[†], Emily E. Wiemers[†,*], Robert A. Moffitt [◇]

[†] Economics Department, University of Massachusetts Boston

[‡] Department of Public Administration and International Affairs, Syracuse University

[◇] Economics Department, Johns Hopkins University

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 earnings volatility – the variance of short-run earnings growth rates – as well as an error components decomposition of inequality into 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 although 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.

Keywords: Synthetic Data, SIPP SSB, SIPP GSF

[*] eewiemer@syr.edu

Media Summary

The United States has a rich set of government data programs that collects data on its citizens under a promise of strict confidentiality of the information they provide, yet it has been increasingly realized that outside individuals and organizations can use sophisticated computer matching algorithms combined with other publicly available data to possibly identify specific individuals in government data released to the public.

There has been much discussion of ways to guard against this threat, with one proposed method involving the creation of a "synthetic" data set. A synthetic data set is an artificial data set composed of made-up persons which is designed to preserve the patterns in the original data. The appeal of synthetic data sets is that they can, in principle, replicate all the important characteristics of the original data. But since the individuals in the data set are not real, no real individual can be identified.

This paper studies one of the few instances where synthetic data have moved from theoretical considerations to actual implementation. The U.S. Census Bureau created a synthetic data set for an original data set called the Survey of Income and Program Participation (SIPP), which is a household survey of up to 52 thousand American families. Our study asks how well the synthetic data "replicated" the original data set in one specific dimension: the volatility of individual earnings. Individuals often have earnings that fluctuate from year to year because of changes in jobs, the ups and downs of their employers, and other factors. Knowing how unstable individual earnings are has been the subject of much discussion among researchers. The analysis compares earnings volatility in the synthetic SIPP data and the original SIPP data (the data are augmented by administrative earnings records).

The study finds that the level of volatility is higher in the synthetic data than in the original data for almost all measures of volatility because of differences in estimates of cross-sectional inequality and short-term earnings dynamics. The study concludes that synthetic data sets combined with validation on the non-synthetic data can be a way to provide access to data that needs privacy protection but that more work is needed on synthetic data before they can be used alone, without the original data, to generate reliable results.

1. Introduction

Many publicly available data sources contain potentially sensitive data and the information needed to identify the individuals or households who provided the data. One of the methods designed to increase 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–or 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 (see also (Little, 1993)). 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 differences in estimates of the variability of earnings between synthetic and non-synthetic data in one of the very few attempts to create a synthetic data file for a major

[1]See Raghunathan (2021) for a basic discussion of the synthetic data approach. See National Academies of Sciences, Engineering, and Medicine (2023), Chapter 6 for a comprehensive survey. See Bowen et al. (2022) for a discussion of a recent application to tax data.

household survey data set linked to administrative earnings histories. The synthetic file is called the SIPP Synthetic Beta (SSB), and the non-synthetic data file, composed of the original records, is the SIPP Gold Standard File (GSF). 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 force activity and success. Household survey reports of earnings are well known to be misreported or not reported at all with fairly high frequency and imputation methods to address these missing data have been shown to be problematic, particularly with predicting earnings and work in the tails of the earnings distribution (Bollinger et al., 2019; Chenevert et al., 2016; Klee et al., 2019). Development of the SSB files began in 2003 and continued through 2022 (Abowd et al., 2006; Benedetto et al., 2013, 2018). The GSF data are built on a set of uniform extracts of variables taken from the Survey of Income and Program Participation (SIPP), a nationally representative 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 GSF.

The SSB was a synthesized version of these data. Although the Census Bureau has not released details of the models used to construct the SSB, broadly we know that Census fit a parametric statistical model to the variables in the GSF which was then used to produce the fully synthetic SSB, including synthesis of 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 research data center (VRDC) to conduct analyses. In addition, researchers were invited to submit their programs to the Census Bureau so that they could be run on the original, underlying GSF data, to check the accuracy of the results obtained on the synthetic data (after disclosure review). The "validation server" was the term used for the server that ran the programs on the GSF. The guidance from Census Bureau was that the SSB was designed to be used in conjunction with validation on the GSF although there was no requirement that researchers submit their programs for validation and some did not. But even with the validation option, the intent of the SSB models was to preserve the underlying covariate relationships between the variables sufficiently well that submitting programs for validation would only need to be occasionally done. Combined with validation, the SSB provides access to one of the most commonly used sources of administrative earnings histories (the DER), without having to go to a Federal Statistical Research Data Center (FSRDC) and work with the GSF data directly.

The SSB was available publicly until September 2022 when the server hosting the data was shut down. This has some implications for our work reported in this paper as will be discussed momentarily. Because the SSB is not currently available for researchers and it is not known whether it will be made available again and whether it will use the same data generating models, the goal of our work is not to assess how well this particular synthetic data set performs but rather to use analysis of the SSB and GSF to illustrate the accuracy of one particular historical synthetic data set to inform future synthetic data products.

Our study concerns one of the key issues with synthetic data files, which is how faithfully they capture the characteristics of 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 these norms National Academies of Sciences, Engineering, and Medicine (2023). 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. The accuracy of synthetic data for any given research question is important for at least two reasons. First is whether it is possible to treat estimates from synthetic data as a substitute for the non-synthetic data, thus removing the need for access to the non-synthetic data and a validation server entirely. Second is whether conclusions drawn from synthetic data are qualitatively similar enough to the non-synthetic that researchers can work with the synthetic as if it were the non-synthetic under the presumption that only the necessary results will be validated on the non-synthetic. This is critical for real-world applications because the frequency and total number of validations that synthetic data users may need depends on how well the synthetic data match their non-synthetic counterpart. Given limited privacy budgets and resources available to validate synthetic data (absent automation), the accuracy of synthetic data is indeed important even if all final results will come from the non-synthetic counterpart.

In this paper, we present a set of results on short-run earnings dynamics estimated on both the SSB and the GSF, focusing on earnings volatility – the variance of short-run earnings growth rates – and an error components decomposition of inequality into 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] The results presented here come from a collection of published and unpublished estimates from (M. D. Carr et al., 2023) and from a project that was designed to compare estimates of earnings volatility across three different survey and three administrative datasets using the same samples and methods, one of which was the SSB M. D. Carr et al. (2023) and R. A. Moffitt et al. (2023). The SSB work in those projects followed the same procedure: we first ran an analysis on the SSB, then submitted the analysis to the Census Bureau for validation and disclosure on the GSF. Validation is necessary because, unlike researchers who might use the GSF inside an FSRDC, where they can view results on the non-synthetic data in the FSRDC prior to disclosure, users of the SSB cannot know with certainty what the results will be prior to disclosure review. They are dependent on how well the SSB recreates the GSF for their specific research question and sample.

A limitation of the work reported here is that it was begun after September 2022 and hence is based on saved results from our prior projects where we retained both SSB and GSF results for our main analyses. Because the original purpose of the analysis was not to compare the SSB and validated GSF results, but rather to compare the validated results to those from other survey and administrative data sets, we were dependent on the work we saved from those projects and could not go back and conduct additional SSB analyses. For example, we did not calculate additional statistics such as confidence intervals that we would need to formally make SSB-GSF comparisons because that was not pertinent to the original projects.

We find that short-run instability – both earnings 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 earnings 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. If simple independent noise were added to the cross-sectional earnings distribution, for example, the covariance would fall, not rise. But because higher inequality will increase volatility

[2]M. Carr and Hardy (2022), M. D. Carr and Wiemers (2018, 2021), M. D. Carr et al. (2023), and R. A. Moffitt et al. (2023)

holding the covariance fixed and 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 also show that level differences in volatility between the SSB and GSF depend on what measure of volatility is used and how earnings in the lower tail of the distribution are handled, but that trends across the two datasets are largely similar. The error components model shows a broadly similar pattern, with transitory inequality higher in the SSB than the GSF but with a similar trend. 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.

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 specific research questions 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, although 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 and use the panel data on earnings in the GSF and SSB to examine how well the SSB captures short- and longer-run earnings dynamics. Taken together, Stanley and Totty (2021) and our work show that the SSB does not generally quantitatively reproduce the identical estimate from the GSF, but that many qualitative conclusions carry over thus allowing researchers to potentially reduce the number of disclosure requests a given research question requires. However, in some cases, conclusions are qualitatively different across the two datasets and the differences between results from the synthetic and non-synthetic data are not always the same across subgroups or when different decisions are made about how to treat low earnings. These results can provide guidance to data providers producing synthetic data. Our analyses suggest that more information from the data provider about the models used to produced the synthetic data and thus about the underlying covariate relationships that are likely to be preserved, would greatly enhance the usefulness of the synthetic data, even with the possibility of validation on the non-synthetic counterpart. We also suggest that disclosure processes may need to be adapted to allow for additional disclosures on synthetic data early in the research process to assess how faithfully the results from the synthetic data match those from the non-synthetic data for a given research question.

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, followed 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 R. A. 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 & 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 M. D. Carr and Wiemers (2021), M. Carr and Hardy (2022), and M. D. Carr et al. (2023), all of which estimate volatility in the GSF by using the SSB, with M. D. Carr et al. (2023) also making direct comparison to the SIPP survey. M. D. Carr et al. (2023) is part of a larger set of papers that compares volatility estimates across six sources of administrative and survey data using the same sample definitions and methods (R. A. Moffitt et al., 2023). We also present unpublished results of the transitory earnings variance that were estimated first on the SSB then the GSF.

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 panels overlapped (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, although 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 are available through 2014. Because of how individuals are matched to administrative data, the GSF only 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 need 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 sample. 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 cover only the years of the individual's SIPP panel.

Based on 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 types of models used are described in the papers cited in the introduction, but the details of the models and what variables are used have not been released by the Census Bureau. The last version of the 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 general method for the last version of the SSB is described in (Benedetto et al., 2018). Precise details on the model used for each variable are not available. 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 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. 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.

The Census Bureau encourages SSB users to validate their results against the GSF.[3] 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.[4] Until

[3]The guidance of the Census Bureau on validation for SSB v.7 is to "strongly urge" researchers to validate results using the non-synthetic data and to warn that without validation, there is no guarantee of validity of the SSB for research purposes. This guidance has changed over time. In the SSB v.5, users were "invited" to validate results on the non-synthetic data.

[4]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.

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 and could have results validated on the GSF. The SSB server at Cornell was shut down in September, 2022.

1.3. **Working with the SSB and GSF.** In this section, we describe in more detail the experience of using these data. Understanding the work flow helps explain why it is important to know how faithfully characteristics of the non-synthetic data are preserved in the synthetic data even when validation is possible.

The SSB was accessed through a Virtual RDC. Users had to apply for access to the data but the application process was minimal, geared mostly towards verifying that the stated project could be carried out on the SSB and so that the applicant could be given access to the server. The server had standard statistical and mathematical modeling software preinstalled. All code was executed on the server and all output from that code remained on the server. Users could copy text into the server but not out of the server. Auxiliary data files and code could be uploaded to the server by authorized individuals. The only additional constraints on the user were that (1) the code had to written so that it could be copied from the SSB server to the GSF server and still execute, (2) the code could not rely on web-based resources (e.g., auxiliary data had to be on the server, it could not be downloaded while executing code), and (3) the output needed to be easily interpretable by Census employees to facilitate validation.

Validation procedures changed over time, but eventually SSB users were asked to follow the same disclosure procedures as FSRDC users. After demonstrating that the code ran cleanly on the SSB server, SSB users filled out the validation request document, and Census ran the SSB code on the GSF server. If the Disclosure Review Board (DRB) approved the output, it was then sent to the SSB user. This process generally took anywhere from a few weeks to a few months depending on when the DRB was scheduled to meet, the schedule of the Census employee handling the validation, and the complexity of the validation request. Unlike FSRDC users of the GSF, who see final results and tables inside of the FSRDC prior to disclosure, SSB users only see final results and tables (from the GSF) after disclosure. Therefore SSB users do not know prior to disclosure how well the SSB results matched the GSF results. If the SSB results do not match the GSF, using the synthetic data makes it more difficult to understand results that do not conform with hypotheses or test the sensitivity of a specification. In this case users may need to do additional disclosure requests increasing the amount of time an analysis takes, resources for disclosure, and potentially consuming more of the privacy budget. This is an illustration of why it is important to assess how faithfully the synthetic data replicates the non-synthetic data for a specific application, even when validation is possible.

The results reported in this paper come from a set of published and unpublished estimates from (M. D. Carr et al., 2023) and other projects. The purpose of these projects was not to compare results from the GSF and SSB. Because of that, we did not calculate confidence intervals that we would need to formally make comparisons across the data sets. We also did not always average across all four implicates in the SSB as the Census Bureau recommends and we note the number of implicates used in each figure. These are limitations that we cannot address because the SSB is no longer available. In Appendix Figures 27 - 38 , we show that estimates of volatility are very similar across the four implicates so we think that averaging across implicates is unlikely to affect our results. In other work, we have shown that the confidence intervals around earnings mobility

estimates in the GSF are very small (see (M. D. Carr & Wiemers, 2022)) but we do not know if that carries over to estimates of volatility or how large the confidence intervals are for the SSB.

1.4. **Sample Definitions.** Our baseline sample consists of men age 25 to 59 with positive labor earnings who can be matched to the DER, where earnings are defined as above. In each year, the sample is drawn from pooled SIPP panels. When we measure earnings volatility and estimate error components models of earnings, we drop men with earnings in the bottom 1% of the earnings distribution. These are the sample definitions used in (R. A. Moffitt et al., 2023) and M. D. Carr et al. (2023). Volatility can be sensitive to choices about how to treat very low earnings (M. D. Carr & Wiemers, 2021) and there are several commonly used methods for trimming low earnings in the volatility literature so we show results where we vary the way in which we trim earnings. Finally, 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 from the baseline sample. Education, measured as the highest degree attained, cannot 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 the education 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

In this section, we lay out a model of earnings and use this to motivate the measures of earnings instability that we estimate. We show that a simple model of earnings can be used to derive a measure of earnings instability, called volatility, that is easy to calculate and depends only on the cross sectional variance of earnings in two time periods and the covariance of earnings between these two time periods. We then show a more complex model of earnings where the transitory variance of earnings is estimated using a minimum distance estimator to fit the moments implied by the model to the empirical autocovariance matrix of the data. We assess the similarity of results from the SSB and the GSF for both the relatively simple model of volatility and the more complex error components model of earnings. Since the earnings model used to generate the SSB is unknown, it is not clear, a priori that the SSB will perform equally well for these two measures of earnings instability. In both cases, we follow the literature on earnings instability and show trends in earnings instability over time and assess differences in the level and trends between the SSB and the GSF.

We use a simple permanent-transitory earnings model familiar to the economics literature but modified by R. A. Moffitt and Gottschalk (2012) to capture changes in the variance of those components over time. 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 2.1.

$$y_{it} = \lambda_t \mu_i + \nu_{it} \tag{2.1}$$

Permanent earnings is specified as the product of an underlying time-invariant individual specific component μ_i and a time-varying parameter λ_t. If permanent earnings change over time, it is due to changes in λ_t. The variance of the transitory component, ν_{it} is allowed to change over time as well. Thus changes in the variance of earnings over time can be traced to changes in the variance of the permanent and transitory components. Typically y_{it} is measured in logs, as we do here.

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

$$\sigma^2_{y_{it}} = \lambda_t^2 \sigma^2_\mu + \sigma^2_{\nu_t}. \tag{2.2}$$

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_{it} 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.

Volatility measures the variability of changes in log earnings (the left hand side of Equation 2.1) differenced over short time periods, as given in Equation 2.3.

$$\text{Var}(y_{it} - y_{it-\tau}) = (\lambda_t - \lambda_{t-\tau})^2 \sigma^2_\mu + \sigma^2_{\nu_t} + \sigma^2_{\nu_{t-\tau}} \tag{2.3}$$

$$= \sigma^2_{y_{it}} + \sigma^2_{y_{i,t-\tau}} - 2 * \text{Cov}(y_{it}, y_{i,t-\tau}) \tag{2.4}$$

where $\tau = 1$ in this case. Based on the model of the earnings process in Equations 2.1 and 2.2, Equation 2.3 shows 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 hence will track the transitory variance in an error components model well. Equation 2.4 illustrates an alternative way of thinking about volatility that uses the definition of the variance of a change to highlight the relationship between cross-sectional inequality and volatility. 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 2.3 is to measure volatility using the variance of the arc change in earnings in Equation 2.5 (Dahl et al., 2011; Ziliak et al., 2011). The arc-change is calculated as

$$\text{Var}\left\{ \frac{Y_{it} - Y_{i,t-\tau}}{\frac{|Y_{it}| + |Y_{i,t-\tau}|}{2}} \right\} \tag{2.5}$$

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 2.4, 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 $t-\tau$, 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 weight large percent changes in earnings differently.[5]

The literature that relies on formal decompositions to identify the components of Equation 2.2 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

[5]Large percent changes in earnings often come from the bottom of the earnings distribution where a small change in earnings on a very low base of earnings can still represent large percent changes. This is one reason why the volatility literature often trims the bottom of the earnings distribution.

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 2.1 by considering the distribution of short-run deviations in earnings from an individual-specific long-run mean. R. A. 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 & Solon, 2003; Debacker et al., 2013; Haider, 2001; R. Moffitt & Zhang, 2018; R. A. Moffitt & Gottschalk, 2012).

We rely on a newly developed model presented in R. 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 2.1, but extends the model to incorporate age specific permanent and transitory earnings that can both vary through time. Specifically, the model is given as

$$y_{iat} = \lambda_t \mu_{ia} + \beta_t \nu_{ia} \tag{2.6}$$

$$\mu_{ia} = \mu_{i0} + \sum_{s=1}^{a} \omega_{is} \tag{2.7}$$

$$\nu_{ia} = \epsilon_{ia} + \sum_{s=1}^{a-1} \psi_{a,a-s} \epsilon_{i,a-s} \text{ for } a \geq 2 \tag{2.8}$$

$$\nu_{i1} = \epsilon_{i1} \text{ for } a = 1 \tag{2.9}$$

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 2.8, 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 follows a unit root process.

Transitory earnings, given in Equation 2.9, 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 empirical moments of the earnings distribution are estimated for each individual i of age a between time t and $t-\tau$ going back to the first year of the data or age 20, whichever happens first. The parameters in Equations 2.7 - 2.9 can be estimated by minimizing the distance between the predicted moments of the model and the empirical moments. 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.

R. 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 truly permanent, meaning that their impact does not fade away over time. 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 the top panel we show demographic characteristics of the main sample of men age 25-59. The bottom panel shows educational attainment for the education sample. The GSF (SSB) 'All' sample column includes all men 25-59 (or all men 25-59 who were 25+ at the time of their SIPP survey for the education sample). The GSF (SSB) 'Matched' sample column includes the subset of these men who can be matched to earnings records. The 'Volatility' sample column includes the subset of these men who have positive earnings in two consecutive years. We report race-ethnicity as the proportion of the sample in each race-ethnic category along with the mean age of the sample. We report the educational attainment for the education sample. As we would expect, in each sample, the GSF and SSB are very similar in mean characteristics. For both the GSF and SSB, the matched and volatility samples are better educated and have fewer non-White men but do not differ in age from the sample of all men age 25-59.

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

	GSF Samples			SSB Samples		
	All	Matched	Volatility	All	Matched	Volatility
Main Sample: Men 29-59						
White	0.72	0.75	0.77	0.72	0.75	0.77
Black	0.12	0.11	0.10	0.12	0.11	0.10
Other	0.05	0.05	0.05	0.05	0.05	0.05
Hispanic	0.11	0.09	0.08	0.11	0.09	0.09
Age	40.38	40.72	40.15	40.38	40.72	40.15
N	380000	283000	226000	380000	284000	228000
Education Sample: Men 25-59 Age 25+ at SIPP Interview						
< High School	0.18	0.17	0.14	0.18	0.16	0.14
High School	0.31	0.30	0.31	0.32	0.31	0.31
Some College	0.26	0.27	0.28	0.28	0.29	0.30
College	0.16	0.16	0.17	0.14	0.15	0.15
College+	0.09	0.10	0.10	0.08	0.09	0.09
N	186000	152000	142000	184000	152000	141000

Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Main sample is all men age 25 to 59, education sample is men age 25 to 59 who were age 25+ at the time of their SIPP interview. In the Main sample, the All GSF (SSB) column is all men age 25 to 59, the Matched GSF (SSB) column is all men age 25 to 59 who can be matched to administrative records, and the Volatility GSF (SSB) column is all men 25 to 59 with positive earnings in two consecutive years. The columns are comparable for the education sample. The SSB estimates use one implicate.

As shown in Equation 2.4, volatility in log changes is a function of the variance of log earnings. All else equal, higher earnings inequality increases earnings volatility. As such the level and trend of volatility is affected by the level and trend in inequality (R. A. Moffitt et al., 2023). Figure 1a shows that the variance of log earnings in the top panel and the difference between the SSB and the GSF ($SSB - GSF$) in the bottom panel. The variance of log earnings is always higher in the SSB than in the GSF and the difference between the two series (plotted at the bottom) grows over time in absolute terms. Figures 2 and 3 show 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 Totty (2021).

3.2. **Instability in Earnings.**

3.2.1. *Volatility.* Figures 4 and 5 show trends in volatility and differences between the GSF and SSB using log and arc changes, respectively. 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.

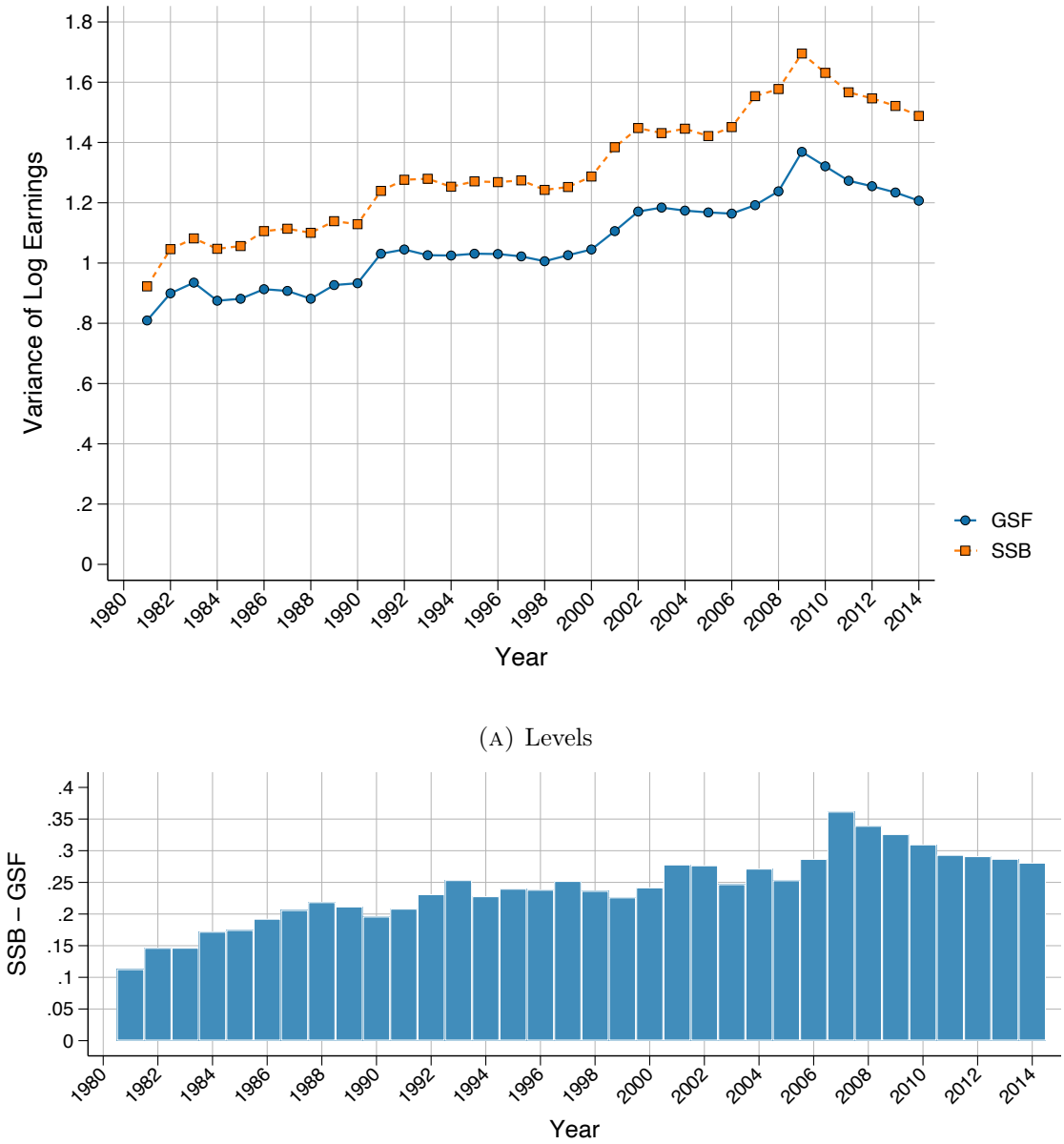
3.2.2. *Decomposing Volatility.* To understand the source of higher volatility in the SSB relative to the GSF, we examine trends in the variance of earnings and covariance of earnings over a two-year period in Figure 6 and 7 using the sample of individuals used to estimate volatility. As we saw with total inequality in Figure 1a, Figure 6 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 7 shows that the covariance of earnings is also higher in the SSB than in the GSF and the gap is also growing over time. The higher and growing level of inequality in the SSB is partially offset by a higher and growing covariance, resulting in similar trends in the two datasets but a higher level of volatility in the SSB than the GSF.[6]

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 8a and 8b 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 pattern in variance of earnings estimated directly in Figure 1a. 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 9a shows that this implies that the share of the total variance that is accounted for by the transitory 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,

[6]Depending on the synthetic data-generating model, the accuracy of the cross-sectional variance and covariance could be related. For example, if the synthetic data model generates y_{it} according $y_{it} = \alpha + \beta y_{i,t-1}$, then the covariance of y_{it} and $y_{i,t-1}$ is a linear function of the variance of $y_{i,t-1}$ and hence simulating too high a value of the latter will generate too high a value of the former. However, as we have noted, the Census Bureau has not released details of their model and it is consequently unknown if a process such as this one was used.

Figure 1. Variance of Log Earnings



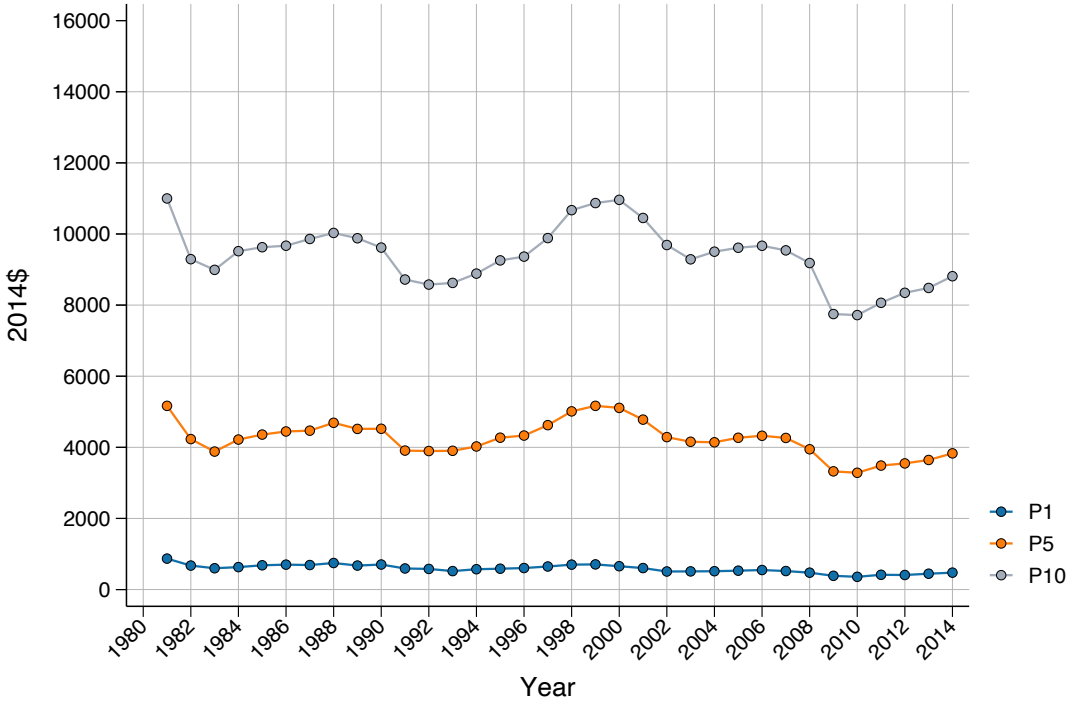
(A) Levels

(B) Differences

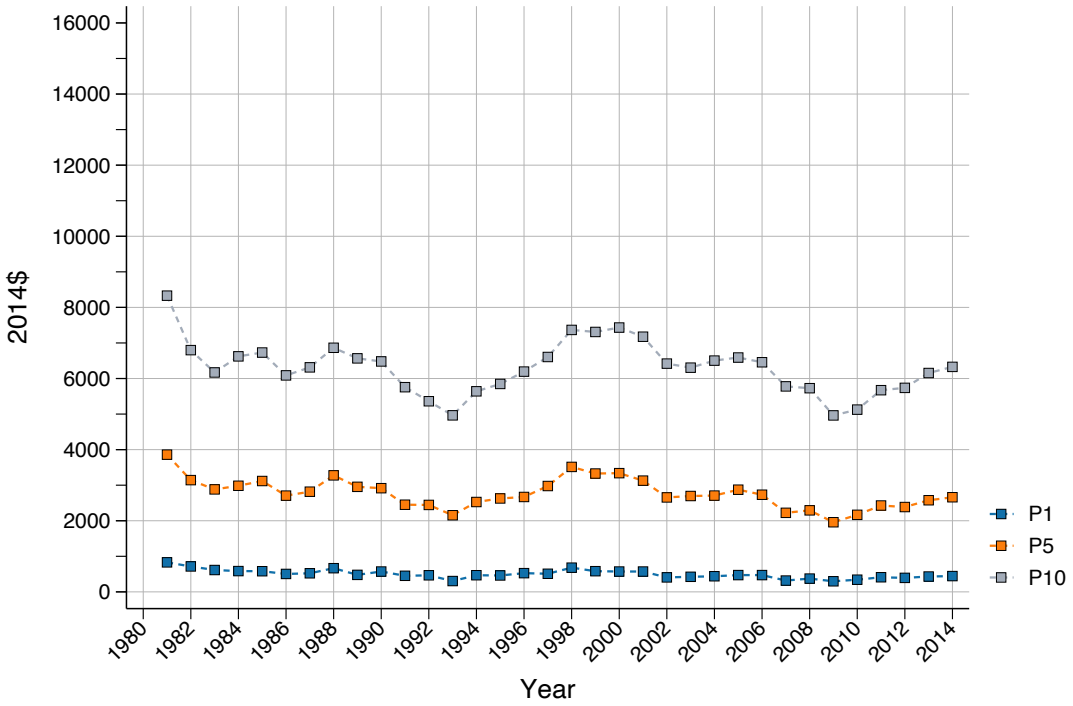
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB estimates average across four implicates.

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 $t - \tau$ 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 also 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,

Figure 2. Annual Earnings Percentiles: 1%, 5%, and 10%



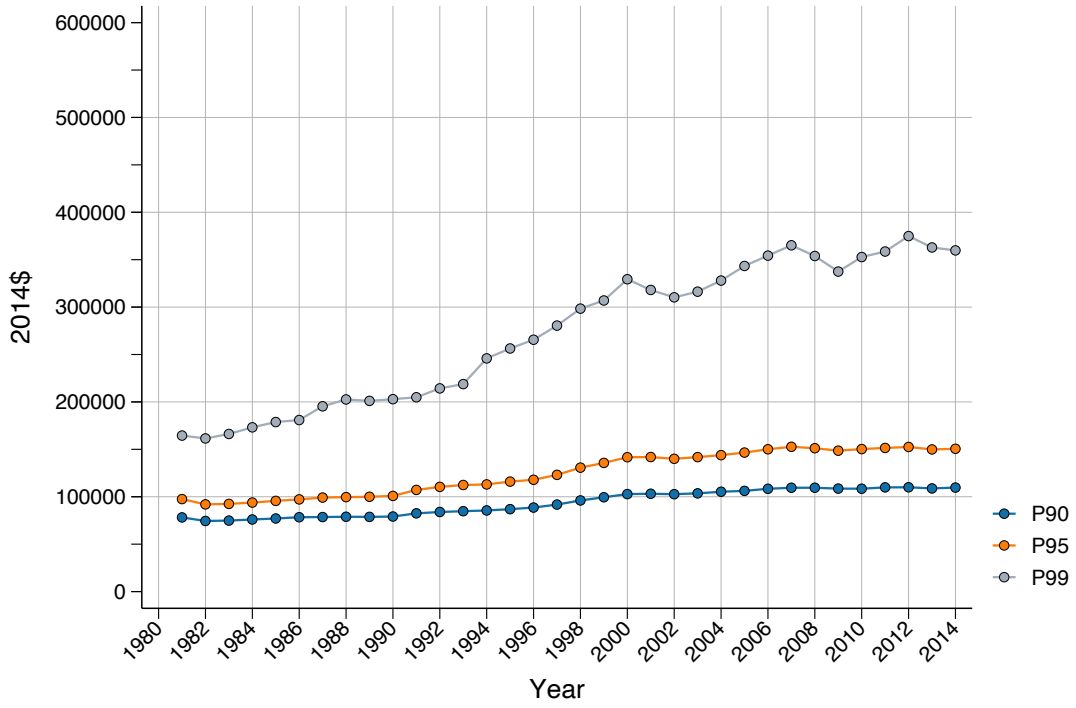
(A) GSF



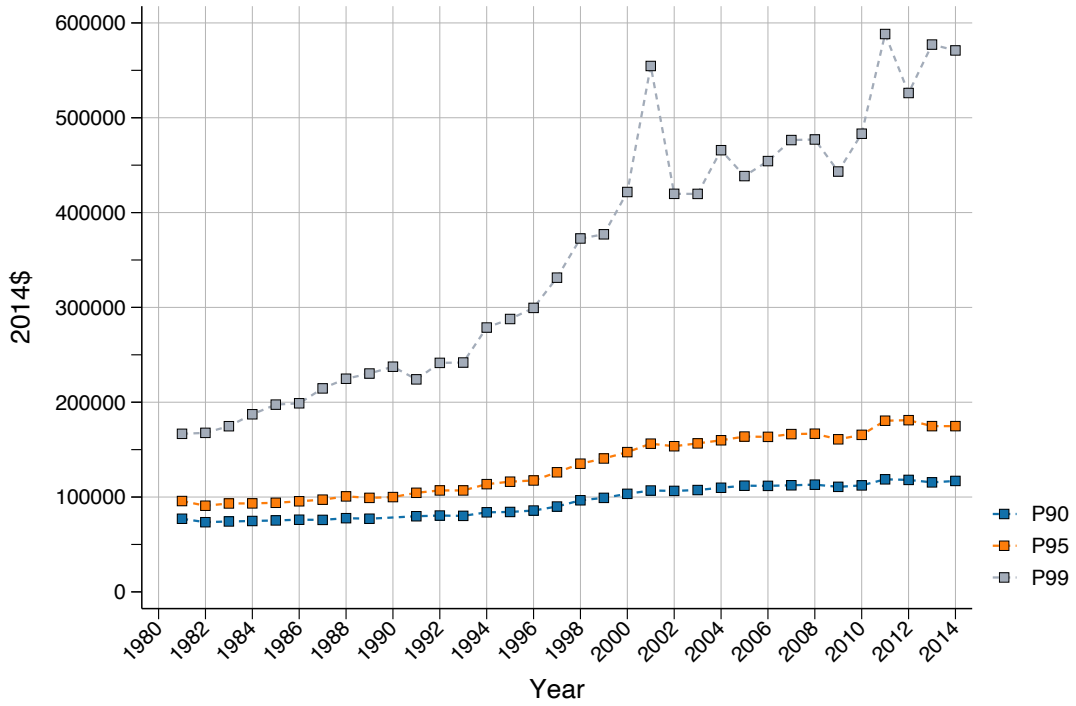
(B) SSB

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 estimates use one implicate.

Figure 3. Annual Earnings Percentiles: 90%, 95%, and 99%



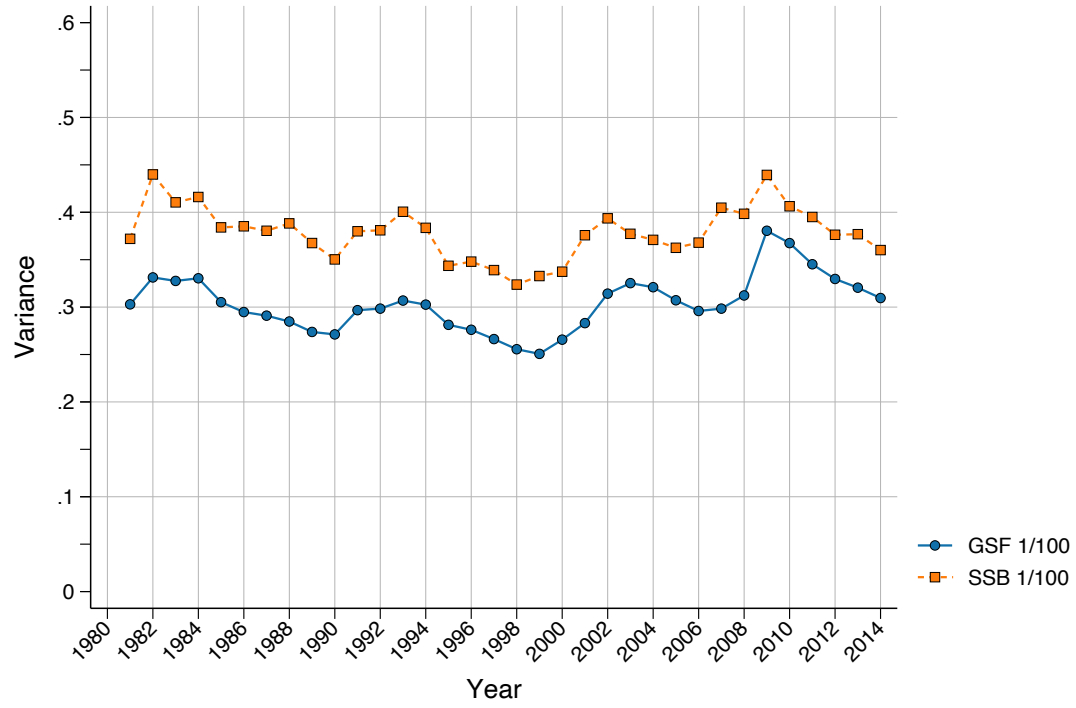
(A) GSF



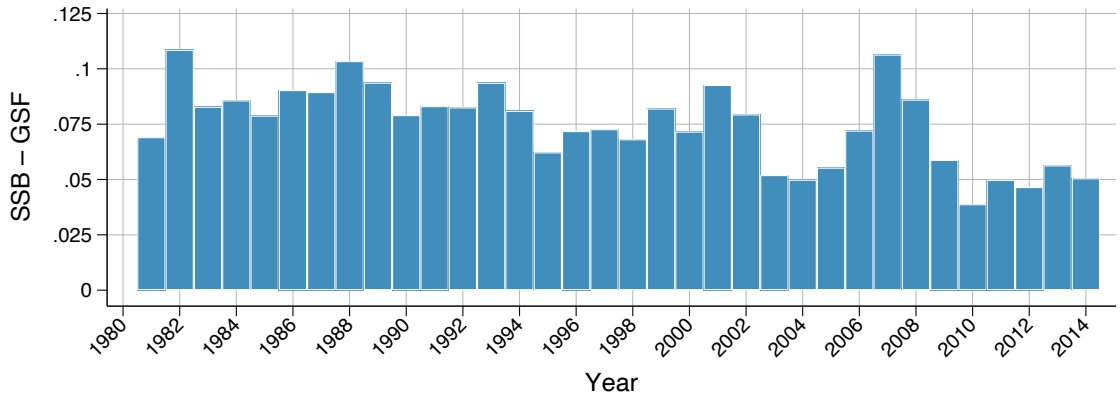
(B) SSB

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 estimates use one implicate.

Figure 4. Volatility in Log Changes, Earnings Above 1%



(A) Levels



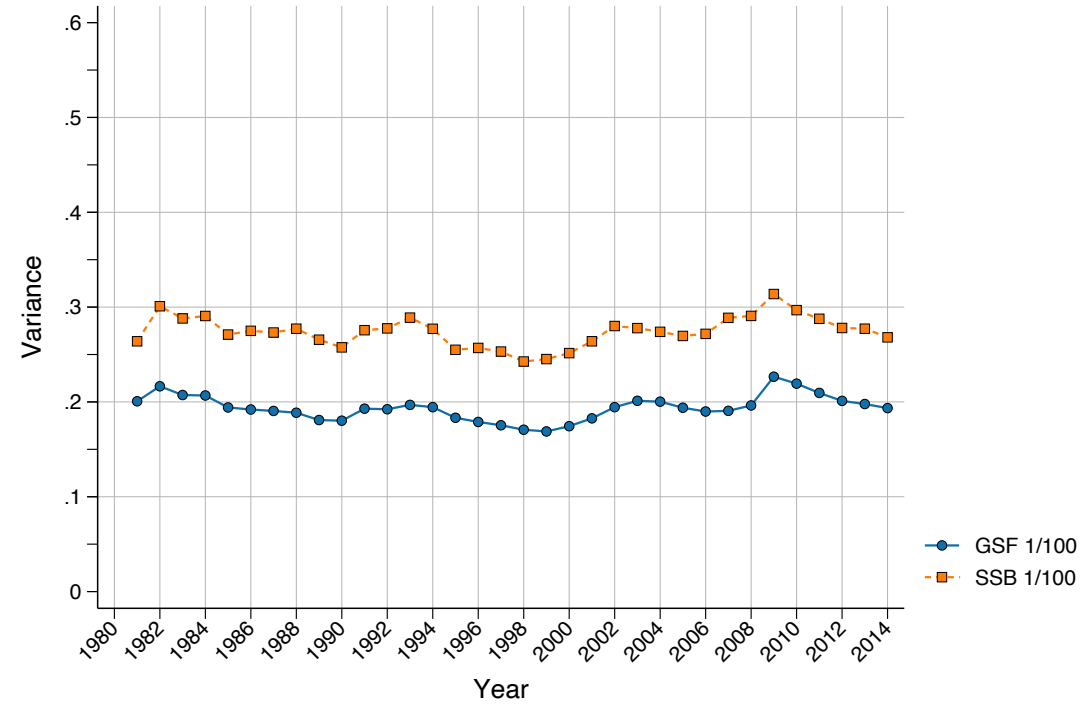
(B) Differences

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 estimates average across four implicates.

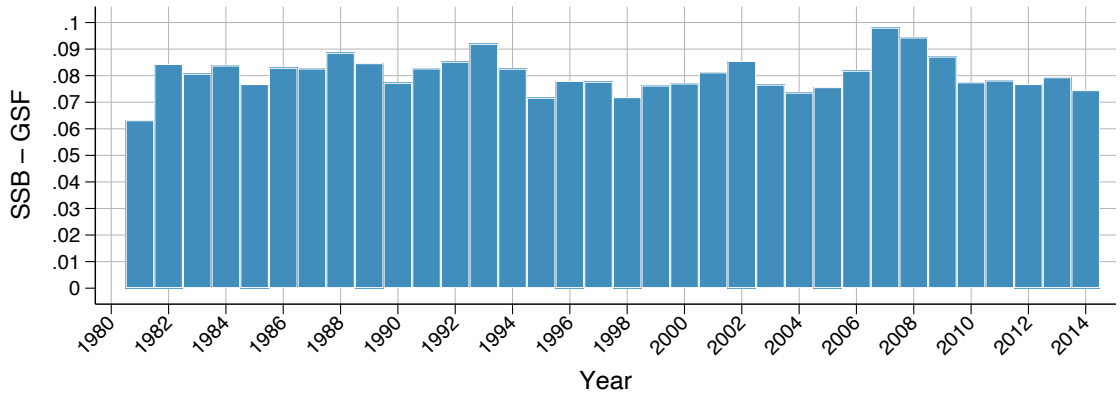
despite these caveats, one aggregate pattern is similar between the ECM model and the trends in aggregate inequality and gross volatility: total inequality is higher in the SSB than the GSF, and short-run instability is higher in the SSB than the GSF. The difference is one of degree.

3.3. **Trimming Earnings.** We have shown results with a 1% trim of low earnings. However, there are many other ways in which earnings have been trimmed in the volatility literature and the level and trend in volatility is sensitive to how low earnings are handled (M. D. Carr & Wiemers, 2021). We show the sensitivity of our results comparing the SSB and the GSF using several of the trimming methods common in the literature. This is important because it shows whether small

Figure 5. Volatility in Arc Changes, Earnings Above 1%



(A) Levels



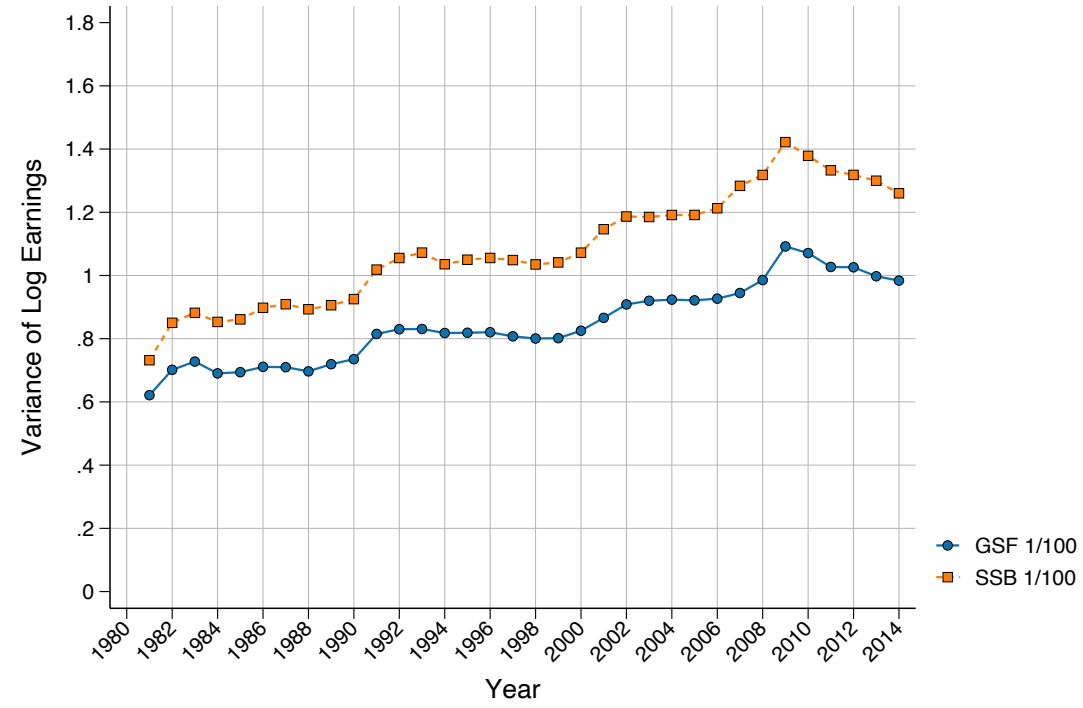
(B) Differences

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 estimates average across four implicates.

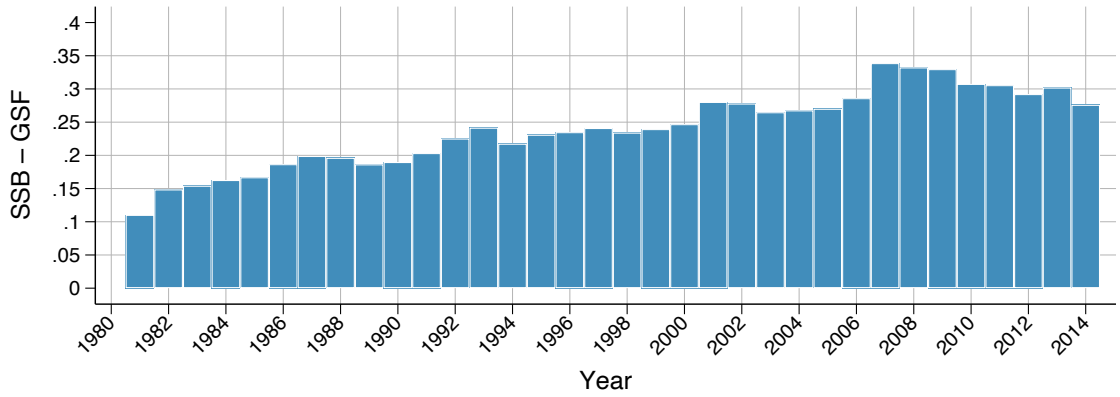
choices about the sample could affect how well the SSB replicates the GSF. We show our results using four trimming methods; (1) where we trim earnings in t and $(t-1)$ at the top and bottom 1% of the earnings distribution, (2) where we do not trim earnings at all, (3) where we exclude earnings in year t and $(t-1)$ that are below one-quarter of full-time, full-year work at one-half of the federal minimum wage; and (4) where we exclude earnings below the minimum earnings required to earn one year of credit towards Social Security eligibility. The latter two trims are in absolute dollars and do not depend on the distribution of earnings in a given year.

Figure 10 shows volatility in log and arc changes trimming the top and bottom 1% of earnings in t and $t-1$, respectively. Additionally trimming at the top, where there is a substantial increase

Figure 6. Variance of Log Earnings, Earnings Above 1%



(A) Levels



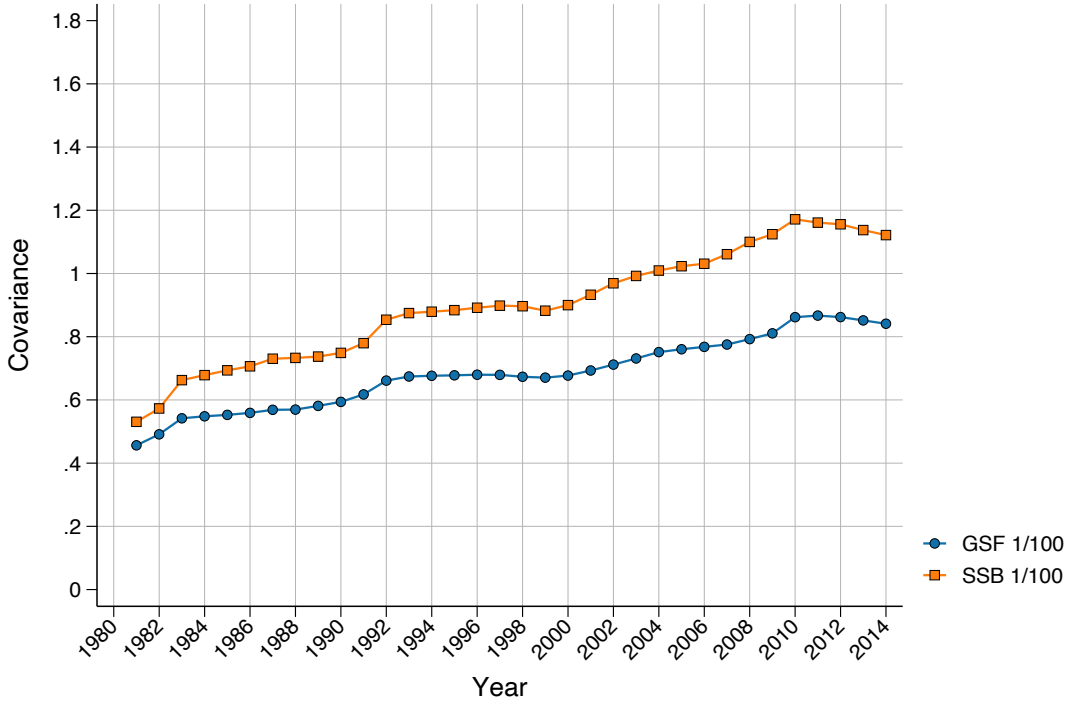
(B) Differences

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 estimates average across four implicates.

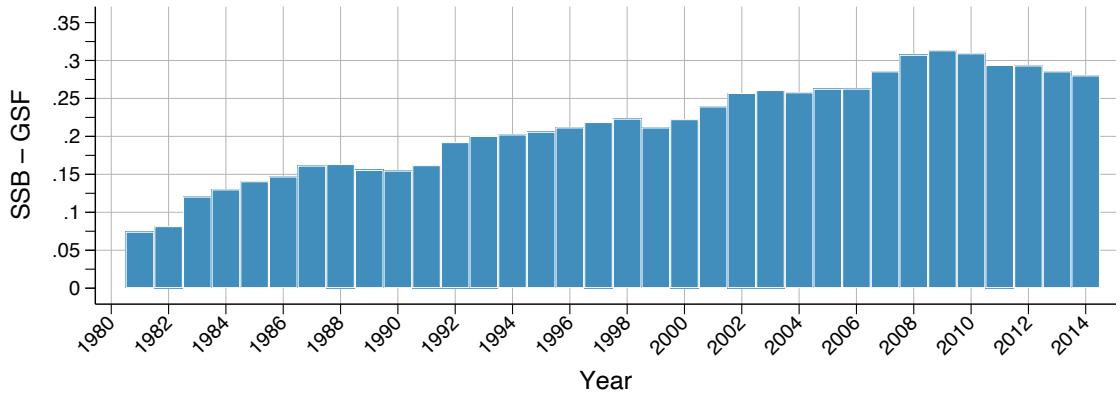
in density of high earnings in the SSB relative to the GSF, does not change the conclusions drawn above. Volatility is higher in the SSB than the GSF for both measures. In the arc change, the absolute difference is relatively stable over time, and the log change measure shows the same moderate convergence.

Figure 12 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, volatility in untrimmed log earnings 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

Figure 7. Covariance of Log Earnings, Earnings Above 1%



(A) Levels



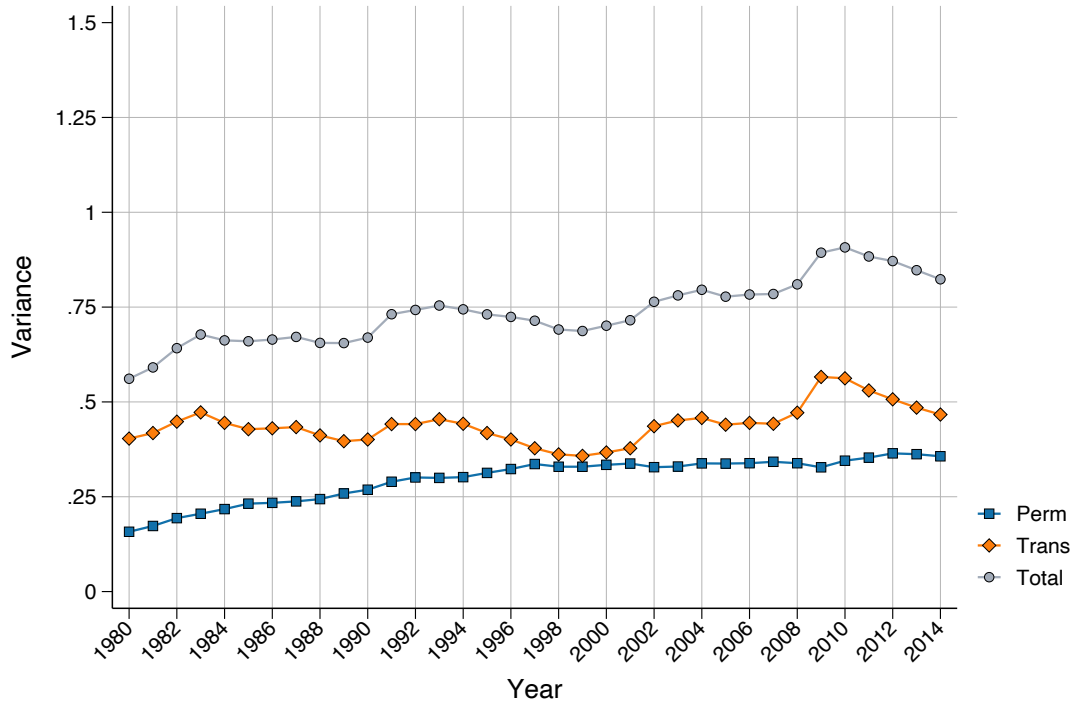
(B) Differences

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 estimates average across four implicates.

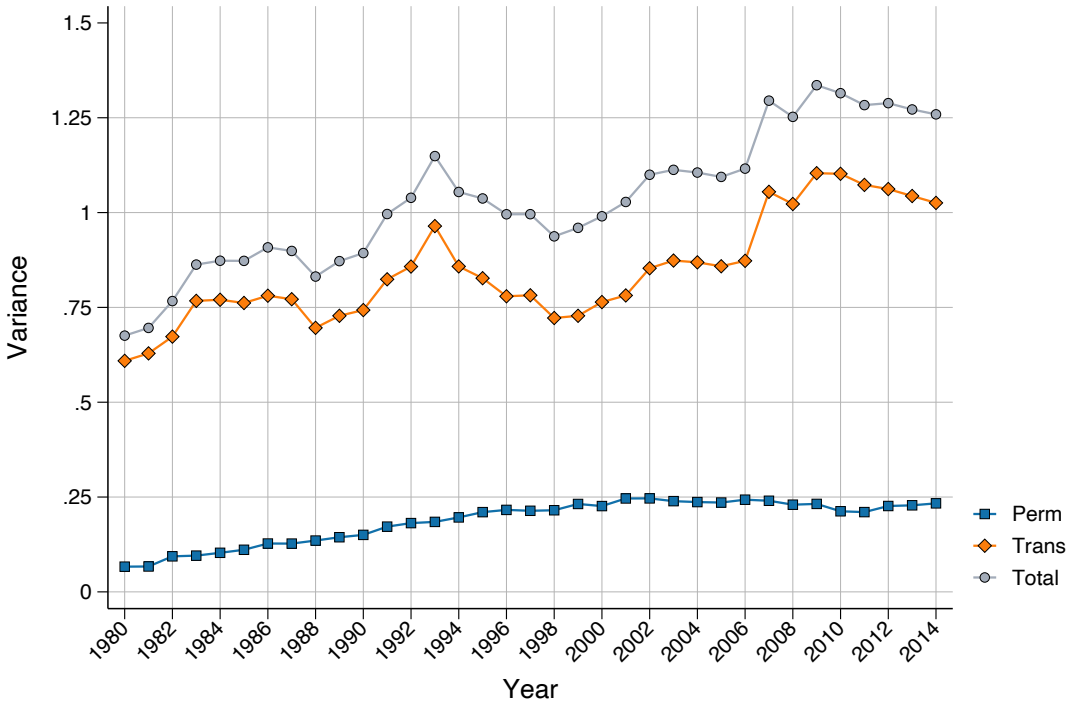
offsetting variances and covariances. Figure 1a shows that the variance of untrimmed earnings is higher in the SSB than in the GSF. Figure 14 shows that the covariance of untrimmed earnings is also higher in the SSB than in the GSF. In the case of untrimmed earnings, the higher variance in the SSB relative to the GSF is completely offset by the higher covariance so that the level of volatility is similar in the two data sets.

Figures 15 and 17 show the results of applying the minimum wage trim and the trim at the threshold for a covered quarter for Social Security, with the 1% trim reported earlier. In general, the absolute dollar trims trim a larger fraction of individuals at any given time than the percentile point based trims and the fraction trimmed will increase over time because the density of low

Figure 8. ECM Decomposition, Earnings Above 1%



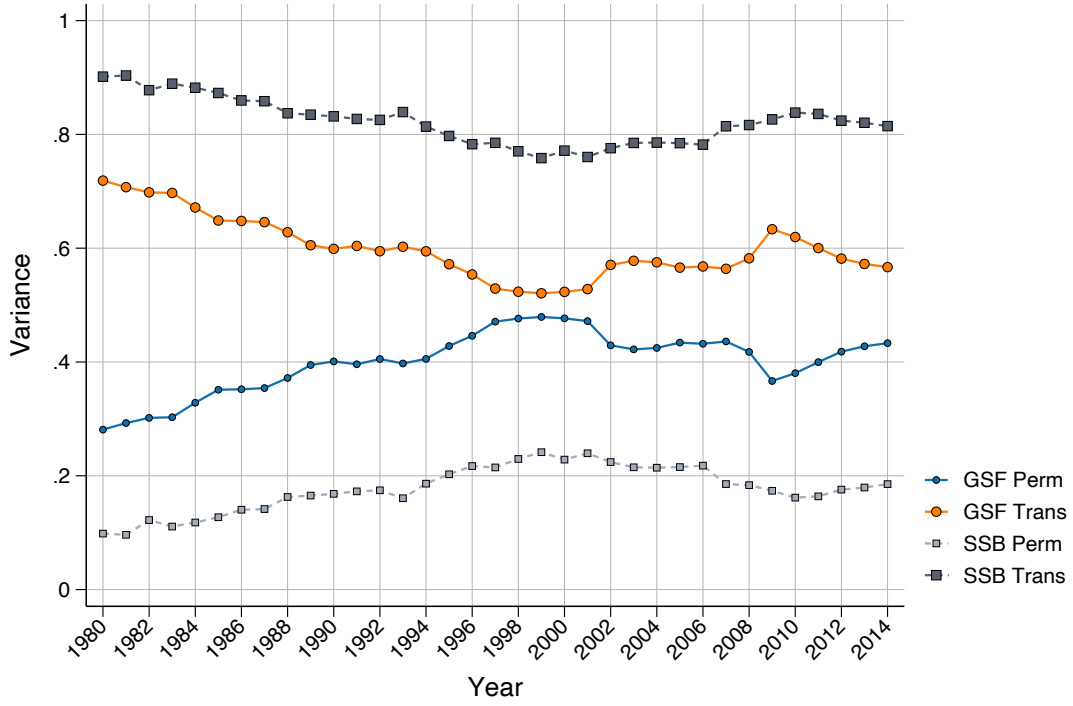
(A) GSF



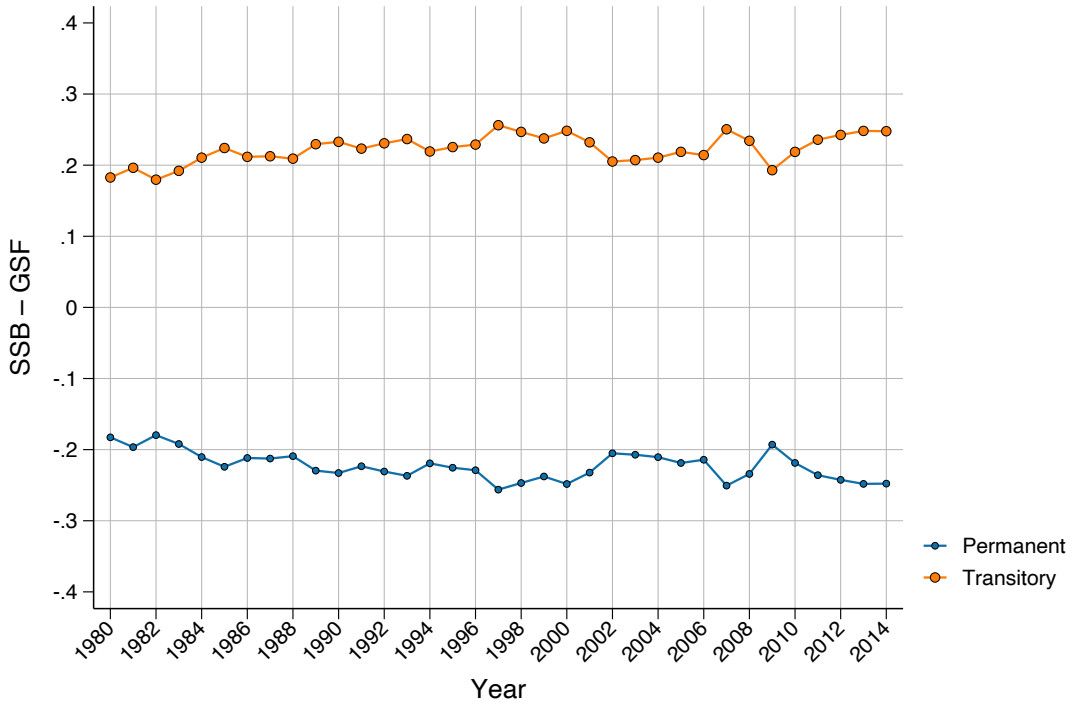
(B) SSB

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. ECM Decomposition, Earnings Above 1%, Shares and Differences



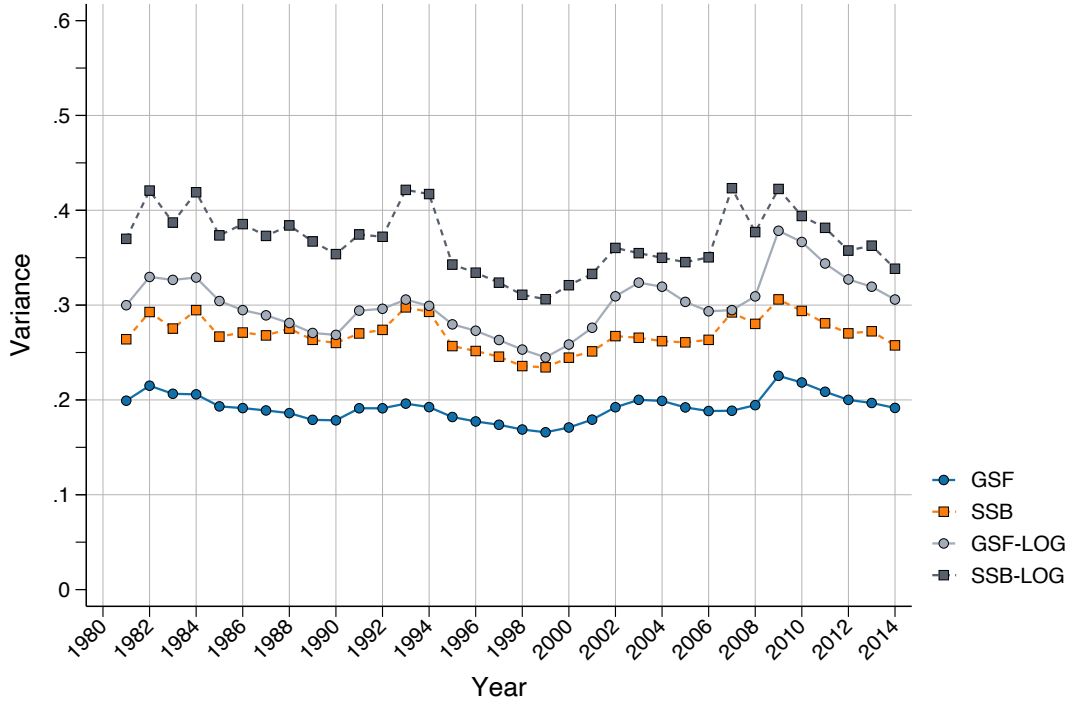
(A) Shares



(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 one implicate.

Figure 10. Volatility, Earnings Between 1% and 99%



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 estimates use one implicate.

earnings is increasing in both the SSB and the GSF.[7] However, the percent of the sample trimmed will increase more in the SSB than the GSF because the density of lower earnings is higher and rising faster, in the SSB than the GSF.

Figure 15 shows the three trims applied to the arc change. As seen in M. D. 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.

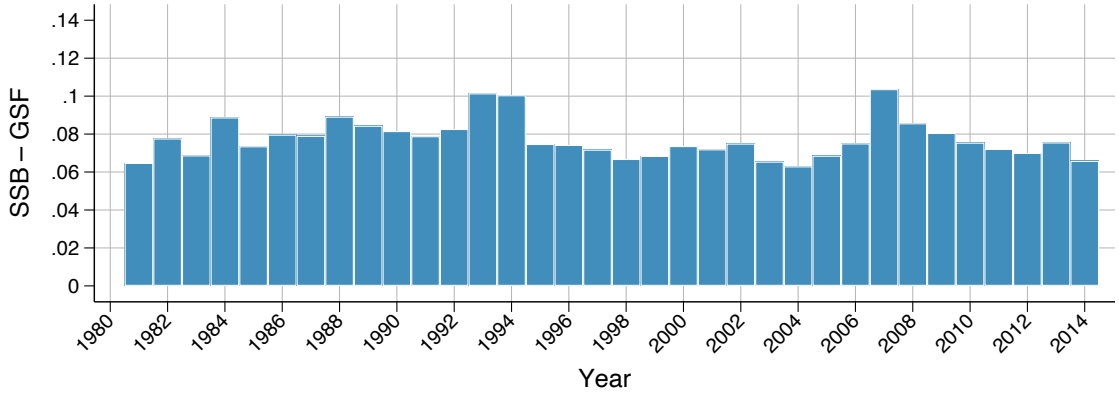
Figure 17 shows volatility for three selected trims in log changes. With the other trims, volatility is higher in the SSB than the GSF, though trends are more similar using the absolute dollar trims than the percentile point trims.

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 19 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 20 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 SSB and the GSF for each trim grows over time.

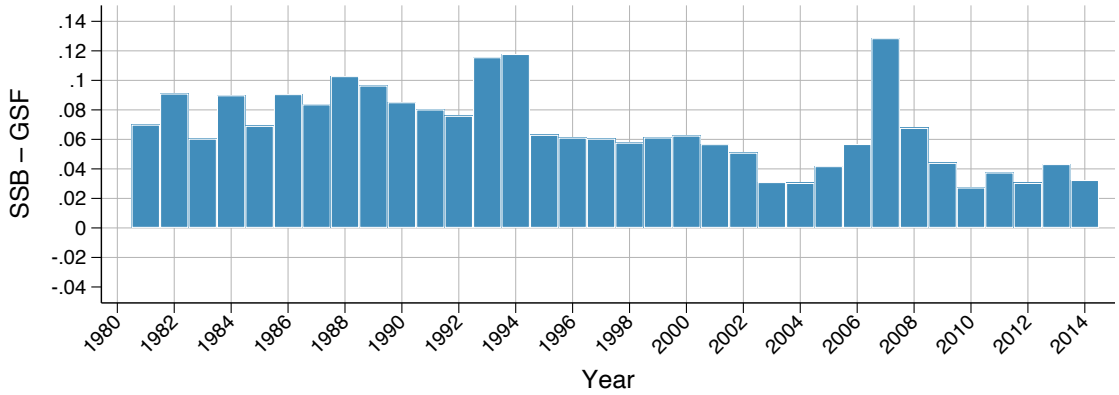
Combined, these results show that the similarity in volatility in untrimmed earnings between the SSB and the GSF is due to the two components of volatility offsetting each other. Inequality is higher in the SSB than the GSF, but the covariance is also higher, resulting in volatility that is

[7]The minimum wage trim excludes earnings below between \$1500 and \$1900 in 2014 dollars depending on the year. The SSA trim increases steadily from \$3100 to \$4550 in 2014 dollars. See M. D. Carr and Wiemers (2021) for details on the trimming thresholds in the GSF.

Figure 11. Volatility Differences, Earnings Between 1% and 99%



(A) Arc Changes



(B) 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. The SSB estimates use one implicate.

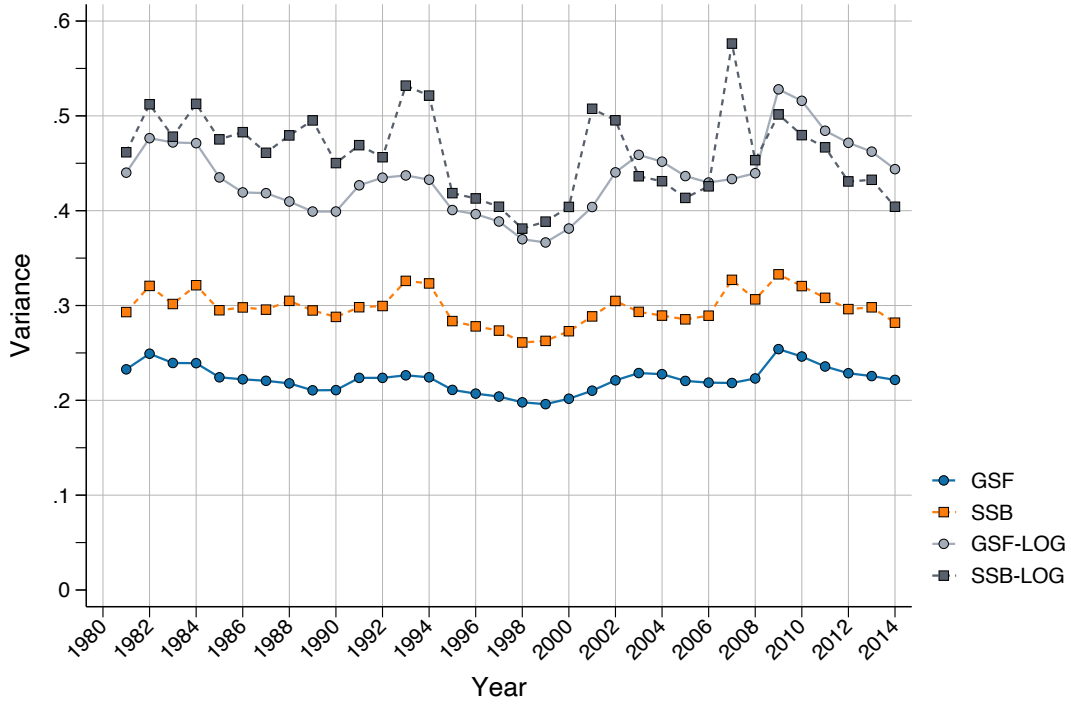
similar. For other trims, and for the arc-change measure, the higher covariance is not enough to offset the higher level of inequality seen in the SSB resulting in volatility that is higher in the SSB.

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 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 22 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 23 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

Figure 12. Volatility, Untrimmed Earnings



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 estimates use one implicate.

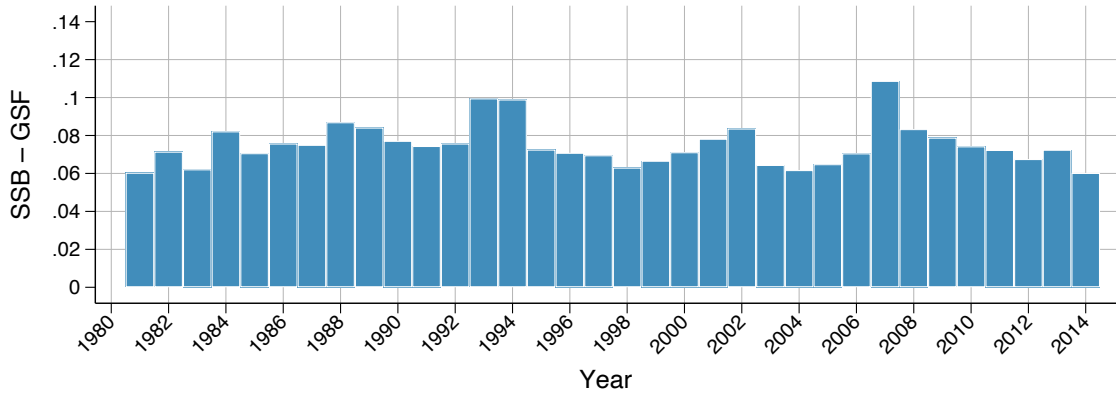
the SSB than the GSF, similar to the overall pattern for the arc change seen earlier. For both non-Hispanic White and Hispanic men, the level differences in volatility are relatively stable through time. For Black men, however, volatility in the two data sets trends in different directions. In the early 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 25 shows volatility by education. Recall 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. Because both age and education are synthesized, this introduces additional reasons why results may differ. 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.

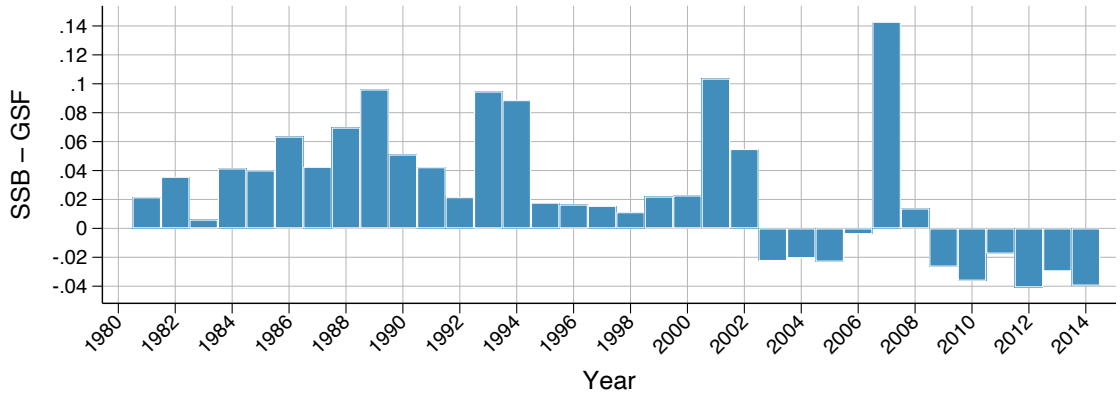
4. Conclusions

Our 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

Figure 13. Volatility Differences, Untrimmed Earnings



(A) Arc Changes



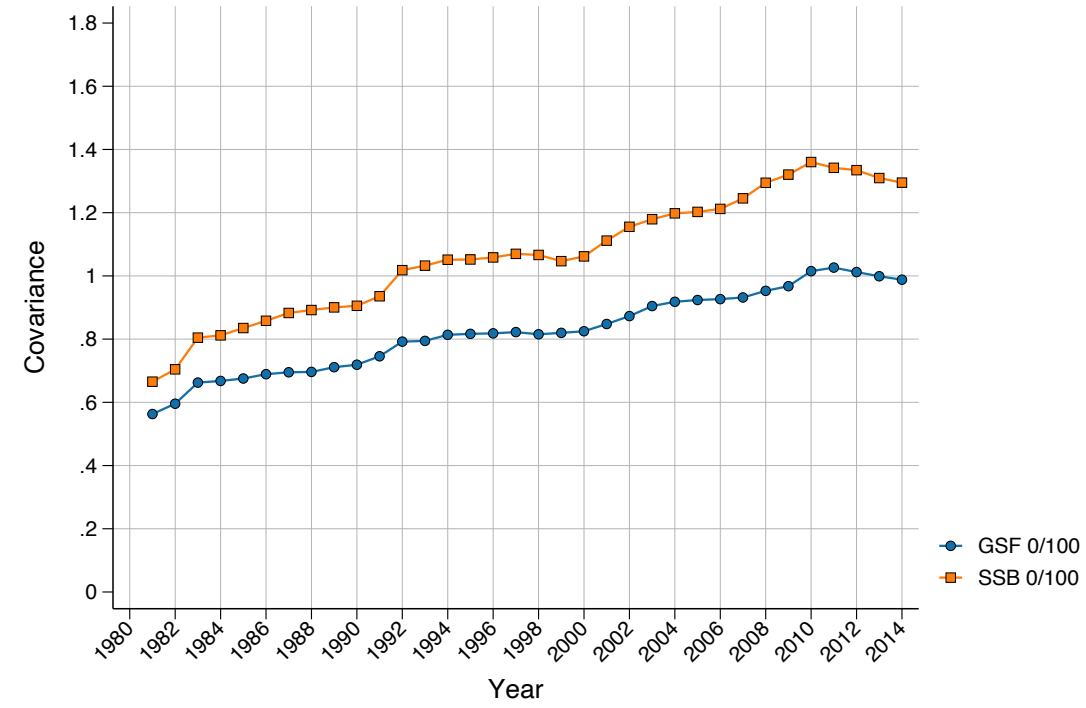
(B) Log Changes

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. The SSB estimates use one implicate.

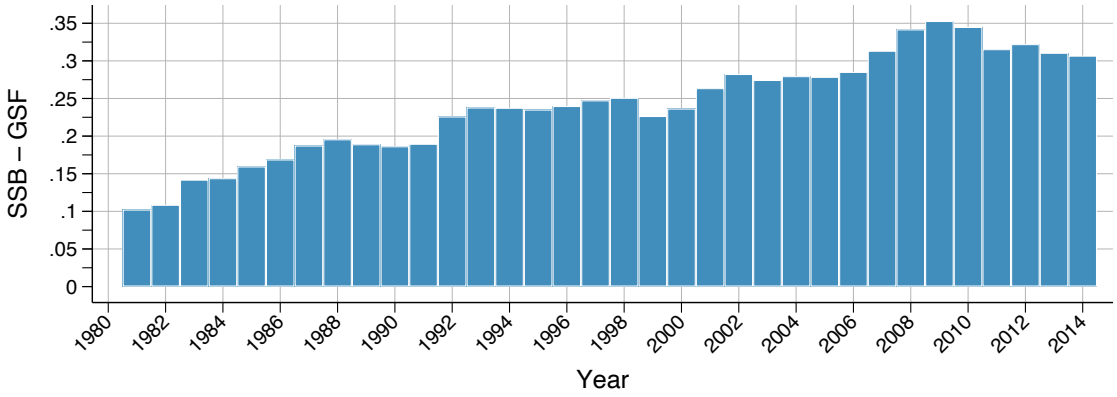
although the differences between data sets depend on the measure of volatility and the way in which low earnings are trimmed. In arc changes, the trends in volatility over time are preserved but the level of volatility is substantially higher in the SSB than in the GSF. In log changes, the difference 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 analyses 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.

Figure 14. Two-Year Covariance of Log Earnings, Untrimmed Earnings



(A) Levels



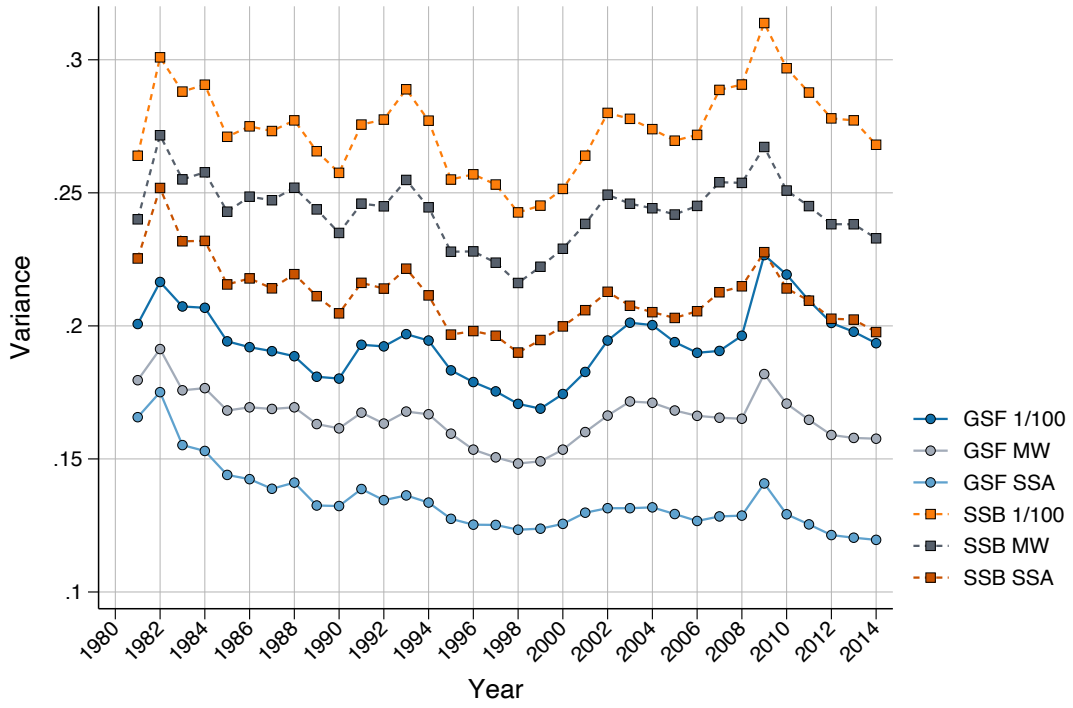
(B) Differences

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 estimates average across four implicates.

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 figure 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 Figures 27 - 38) but, because the SSB is no longer publicly available, we cannot adjust the estimates. Similarly, we are unable to construct

Figure 15. Volatility in Arc Changes, Selected Earnings Trims



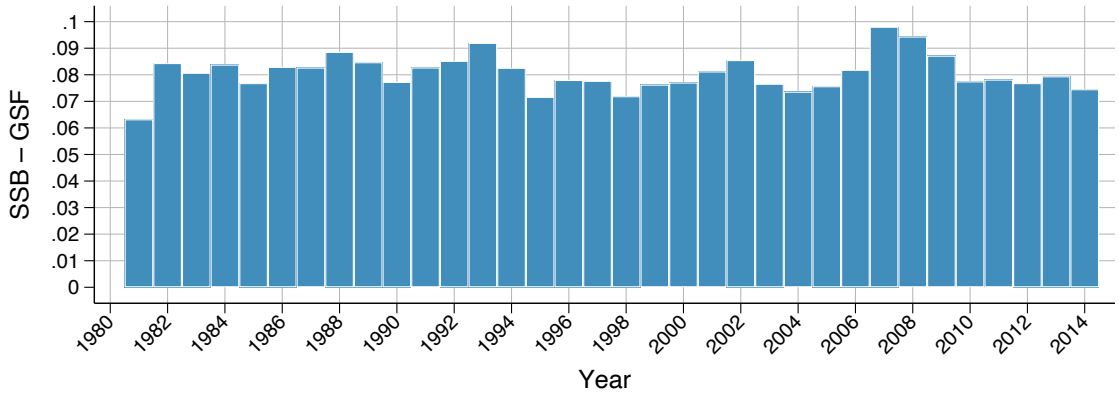
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB estimates average across four implicates.

confidence intervals and assess whether they overlap. 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 where a key component is earnings inequality which we show is larger in the SSB than in the GSF. The differences between the two datasets may not hold for other research questions that use the panel data on earnings. Fourth, we have assumed that the GSF represents the "truth" throughout but, as Stanley and Totty (2021) point out, this assumption may be flawed.

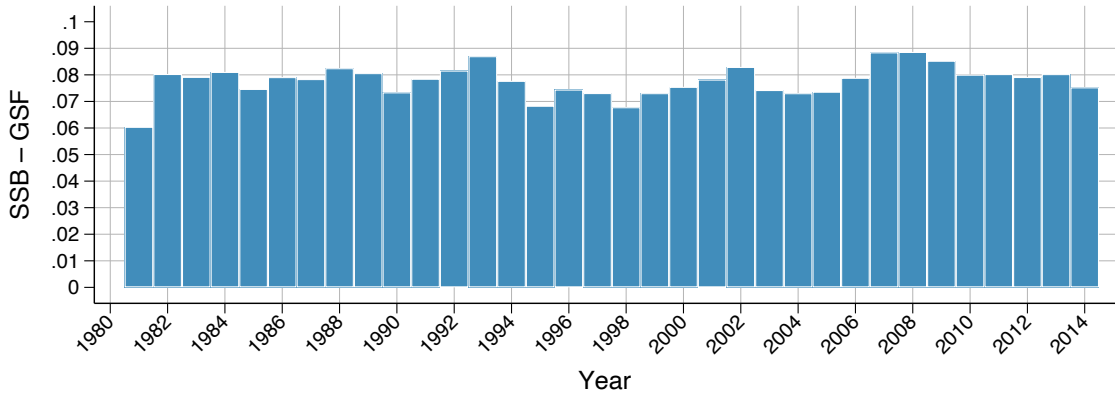
Broadly, our results suggest that the differences between the SSB and the GSF are predictable: instability is higher in the SSB than the GSF, but the levels respond similarly in the SSB and the GSF to the types of changes in sample definitions that we imposed with one notable exception. On the one hand, this is quite positive because it meant that, at least for this research question, we could be confident that qualitative results from the SSB carryover to the GSF. On the other hand, it took multiple disclosures, and thus researcher and Census Bureau resources, to learn the relationship between the SSB and the GSF and thus be able to predict how SSB results might behave relative to the GSF. And, since we know from other work (Stanley & Totty, 2021) that the SSB and the GSF do not always produce similar results, we could not always presume the results would be similar. We also show that some of the differences between the SSB and the GSF depend on the sample definition. For example, if we had first estimated volatility models using untrimmed earnings, we may have incorrectly assumed that the SSB and GSF would yield quantitatively similar estimates for volatility more generally.

This latter point about how predictable the differences are is important for the usability of synthetic data. Obviously, if the SSB, or any synthetic data more generally, are being used in place

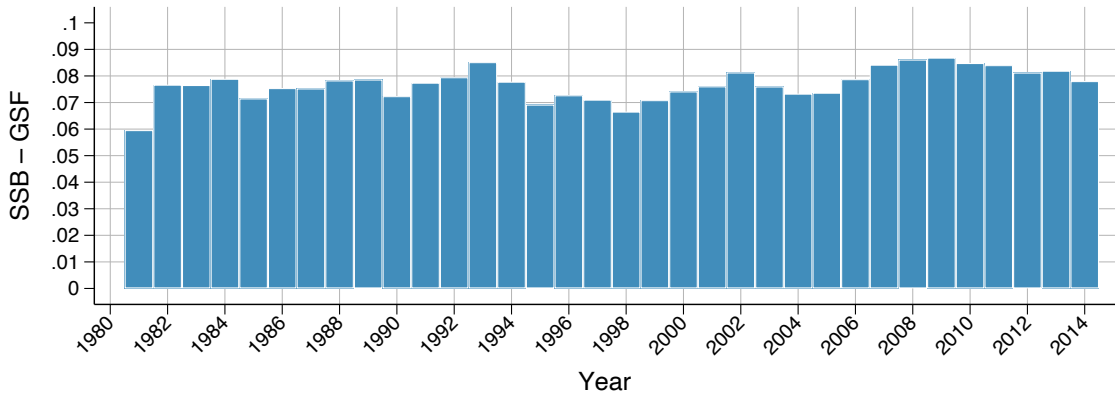
Figure 16. Differences in Volatility, Selected Earnings Trims



(A) 1%/100%



(B) MW

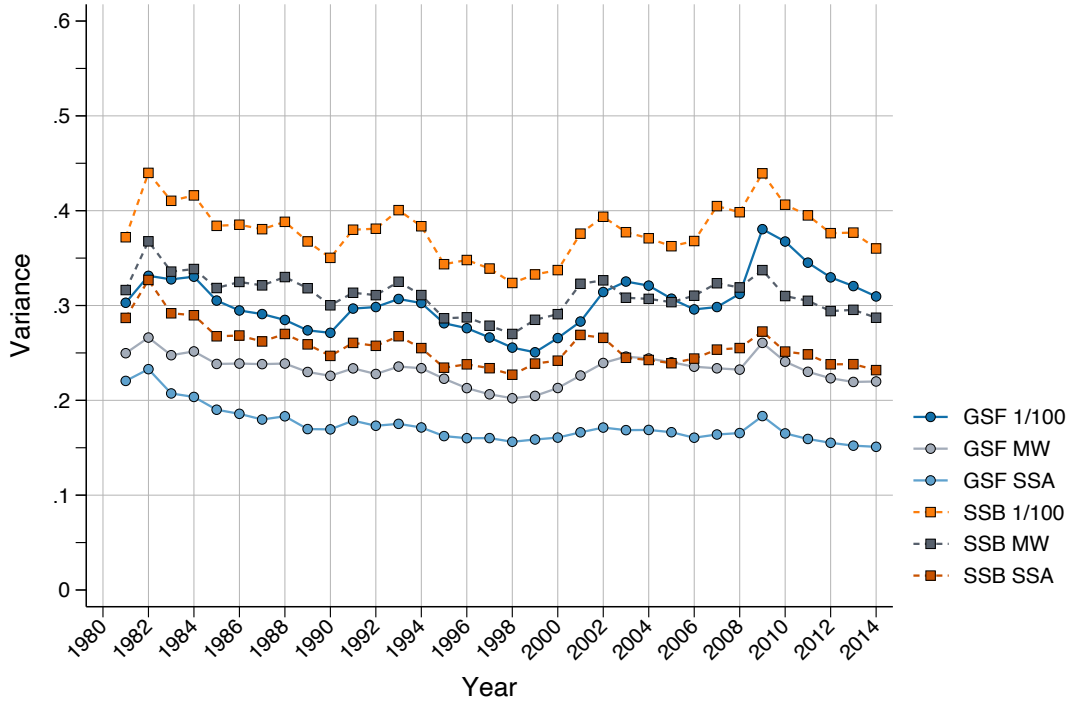


(C) SSA

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 estimates average across four implicates.

of the underlying data then they need to match more or less exactly, or at least for a classes of models specified by the data provider. At the other extreme, if the SSB is being used only to test code, and all estimates can be validated on the GSF with quick disclosure review and no privacy budget, then they need not match at all. As the SSB and GSF were set up, the typical use case is in between these two extremes: users were encouraged to validate SSB estimates on the GSF,

Figure 17. Volatility in Log Changes, Selected Earnings Trims

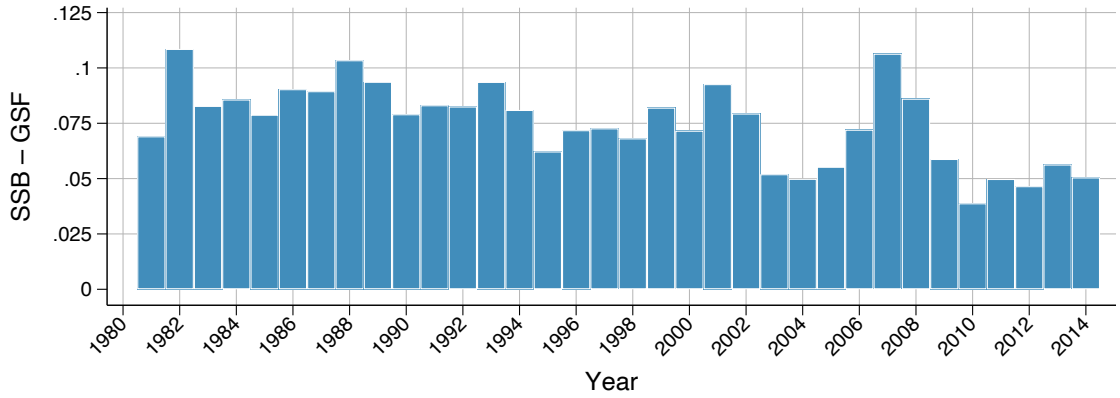


Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB estimates average across four implicates.

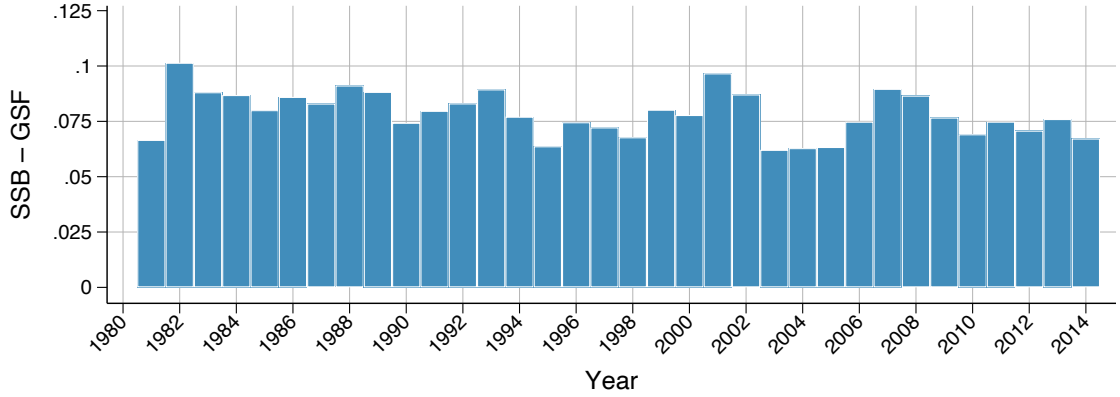
but each set of estimates were subject to a disclosure review and to some unknown total disclosure budget. Under these conditions, it is important that the SSB behave similarly to the GSF to limit the number of disclosures required. But the match between synthetic data and the non-synthetic counterpart may be research-question specific, implying that any given research question will involve some initial set of disclosures to determine how results on the synthetic data correspond to the non-synthetic. If researchers could know, for example, how earnings were modeled in the synthetic data, they could then have a better idea whether the results from the synthetic data would match those from the non-synthetic data for their particular models. For example, for volatility all a researcher would need is for the cross-sectional variance of earnings and the one-year covariance to match within any given sample. For the ECM, on the other hand, researchers would likely need the entire autocovariance matrix of earnings for all years of data to match, which is clearly a much higher bar and would be hard to determine from the model unless the model of earnings happens to generate the same moments as the ECM model.

It is not clear how to implement a system that balances the competing demands of disclosure avoidance and usability for researchers. However, there are ways in which synthetic data providers could assist researchers using synthetic data. For example, synthetic data providers could allow for a sign and significance disclosure on the non-synthetic data to asses how it matches the synthetic data. This would help when estimating regression models or other statistics with well-defined hypothesis tests, but would not be helpful for the estimates presented here. Another possibility would be to to have a process requiring less disclosure which would allow for researchers to know whether the estimate from the synthetic data is within some confidence interval of the non-synthetic estimate (e.g., a 90% confidence interval). For datasets with large sample sizes, like the GSF, this type of disclosure might not be useful because confidence intervals are small. If the confidence overlap, the

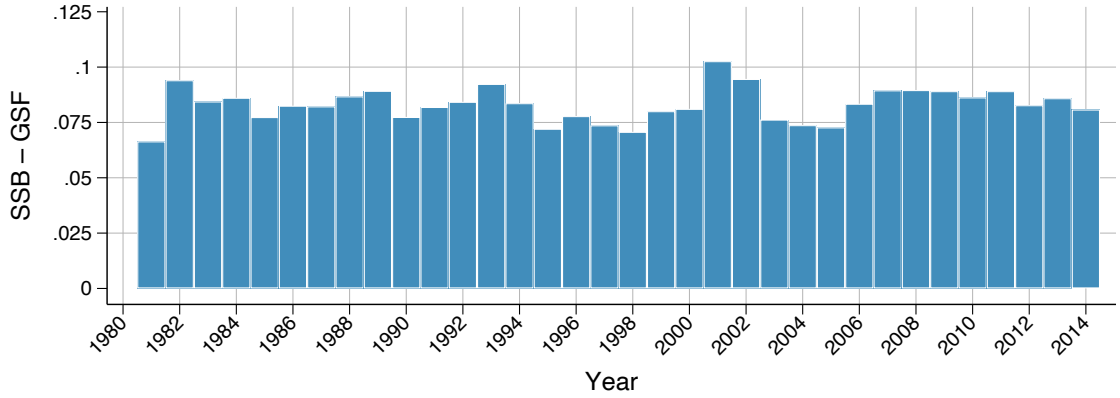
Figure 18. Differences in Volatility, Selected Earnings Trims



(A) 1%/100%



(B) MW

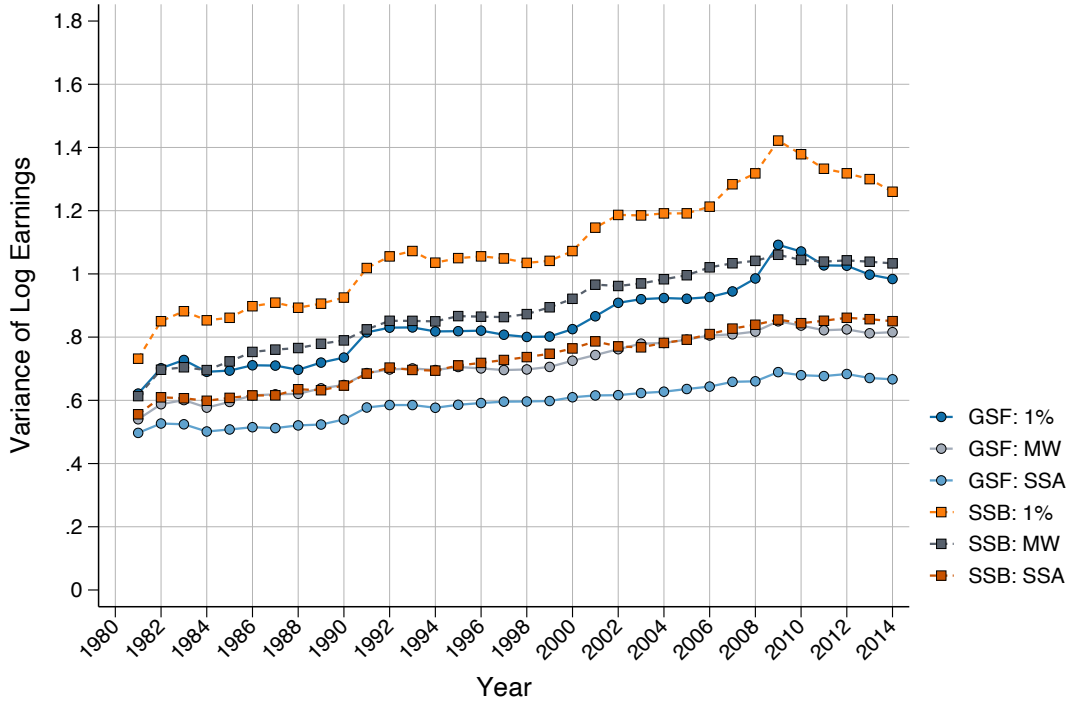


(C) SSA

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 estimates average across four implicates.

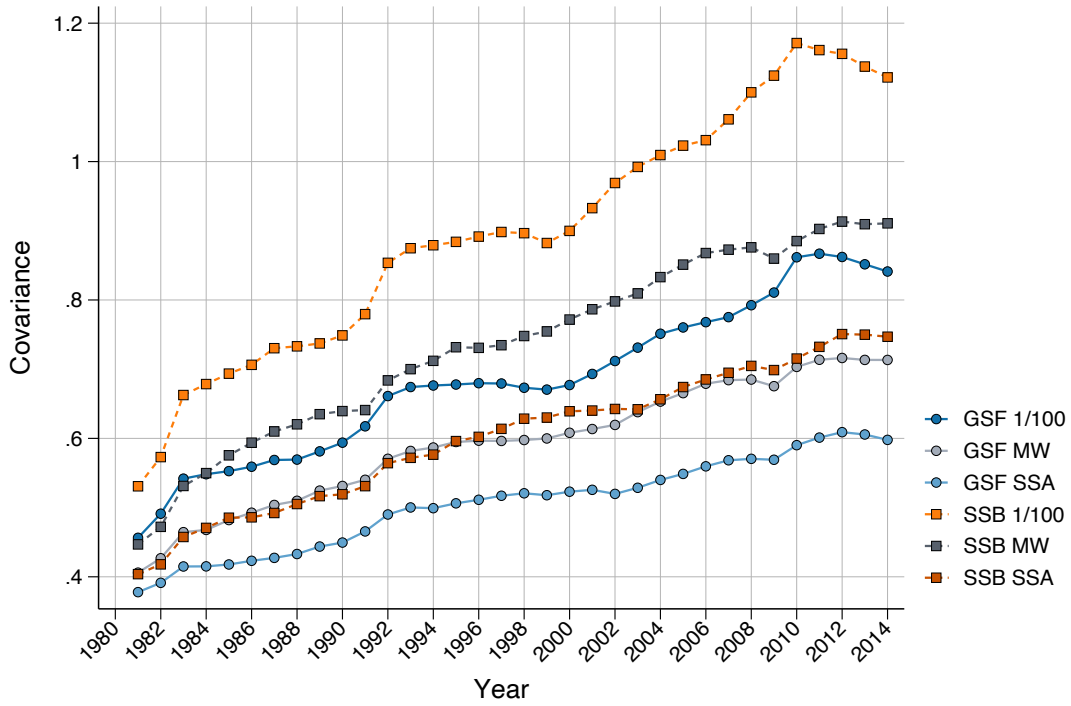
GSF estimate has essentially been disclosed without a proper disclosure review, and if they do not overlap a researcher cannot know if the difference is actually meaningful. Finally, synthetic data providers should allow for an expanded privacy budget for synthetic data users combined with the expectation that there will be frequent small disclosure requests and that these requests may come at the beginning of a research project rather than only disclosing final figures and tables. Despite

Figure 19. Variance of Log Earnings, Selected Earnings Trims



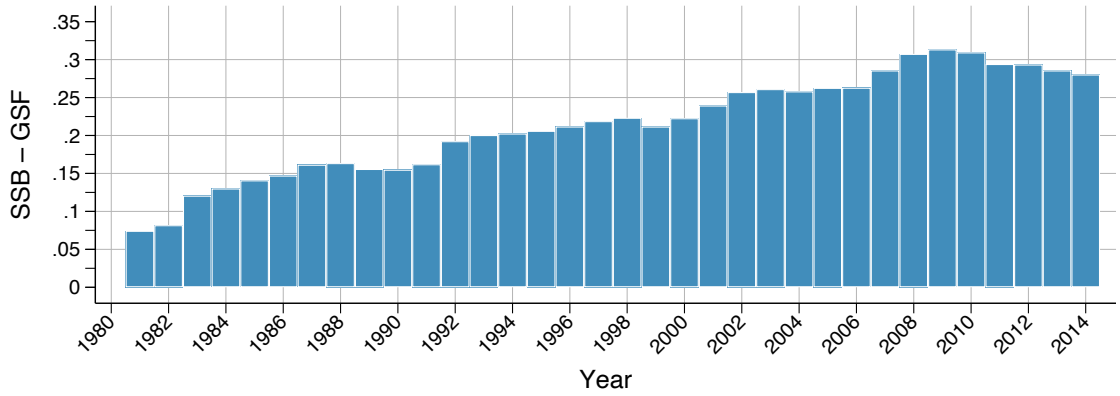
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB estimates average across four implicates.

Figure 20. Two-Year Covariance of Log Earnings, Selected Earnings Trims

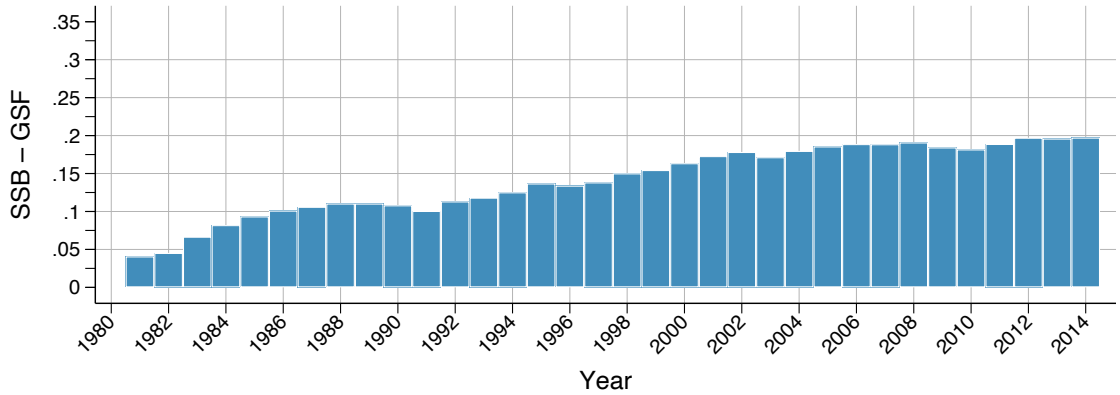


Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB estimates average across four implicates.

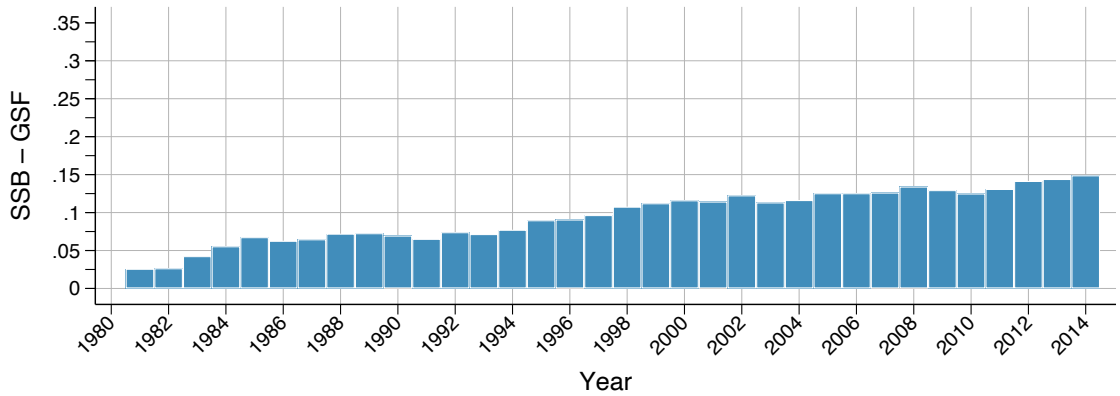
Figure 21. Differences in Two-Year Covariances, Selected Earnings Trims



(A) 1%/100%



(B) MW



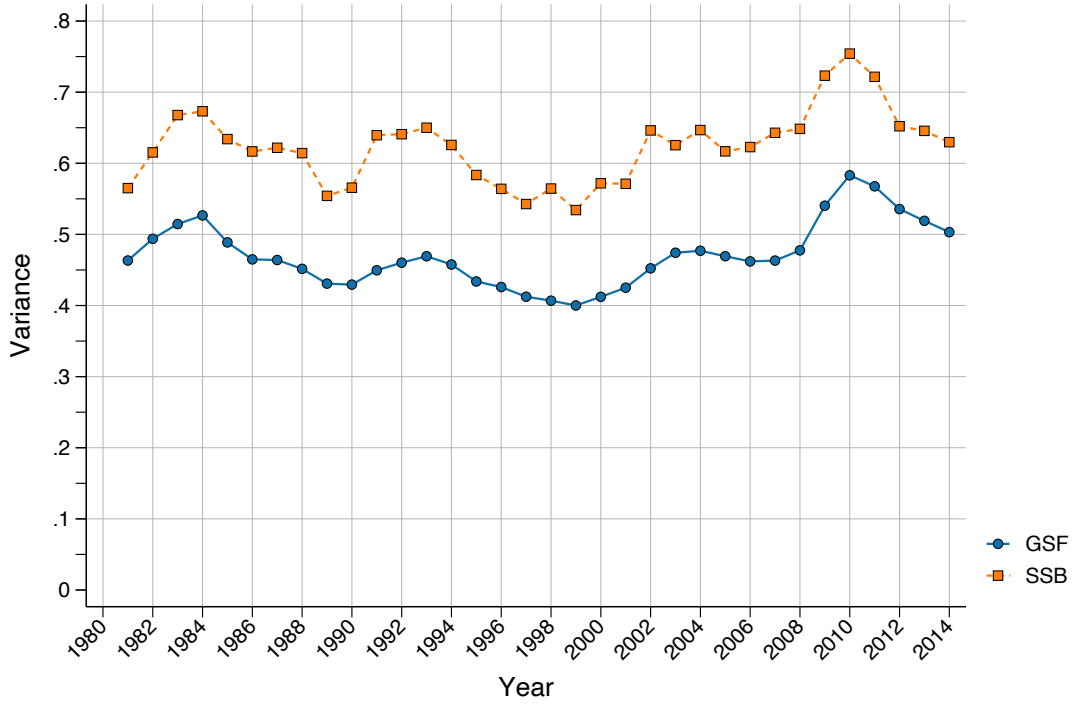
(C) SSA

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 estimates average across four implicates.

these drawbacks, these particular synthetic data served a very useful purpose of allowing access to data that would otherwise only be available at a FSRDC, thus opening access to these rich data to a broader set of researchers.

Disclosure Statement. The authors have no conflicts of interest to declare.

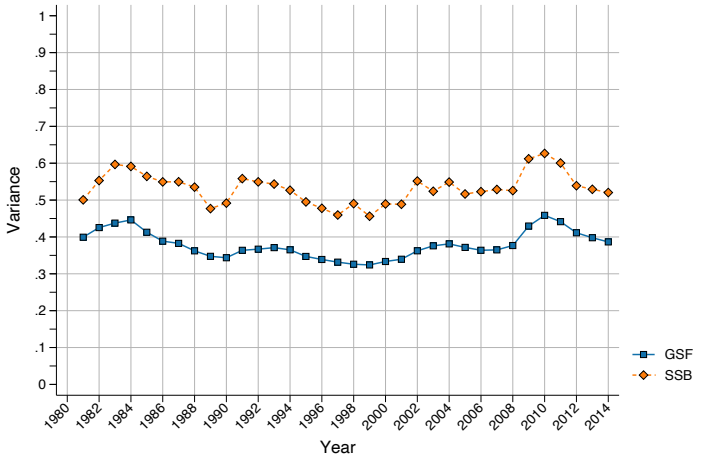
Figure 22. Volatility in Arc Changes, Including Zero Earnings



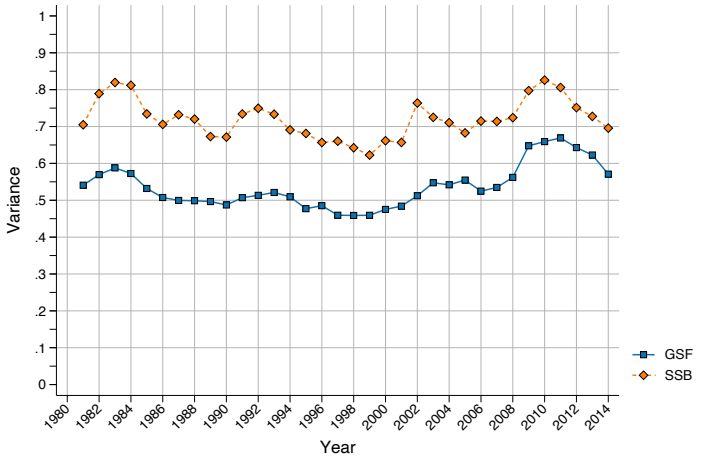
Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 with positive earnings in two consecutive years. The SSB estimates use one implicate.

Acknowledgments. The authors would like to thank Lars Vilhuber and Gary Benedetto for their help accessing and using the SIPP Synthetic Beta data, and participants at an NBER Conference on Data Privacy Protection on May 4-5, 2023. One set of results is covered by review #CBDRB-FY21-095, other results predate the requirement that all cleared output receive a Disclosure Review Board (DRB) clearance number. The validation analysis does not imply endorsement by the Census Bureau of any methods, results, opinions, or views presented in this paper.

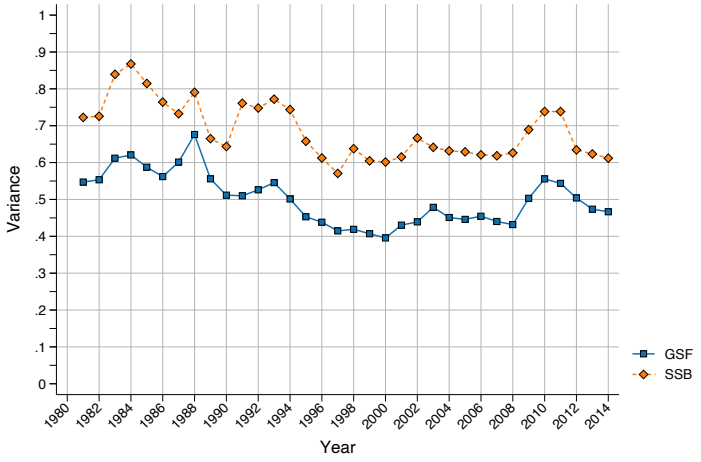
Figure 23. Volatility in Arc Changes by Race, Including Zero Earnings



(A) Levels: White



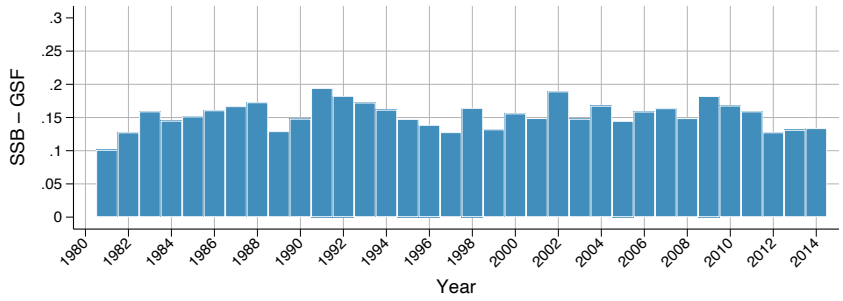
(B) Levels: Black



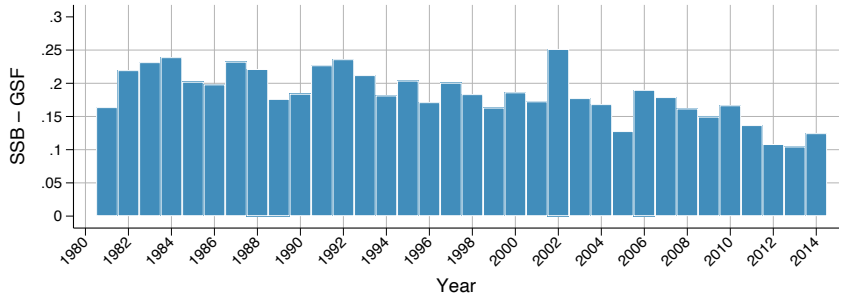
(C) Levels: Hispanic

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 estimates use one implicate.

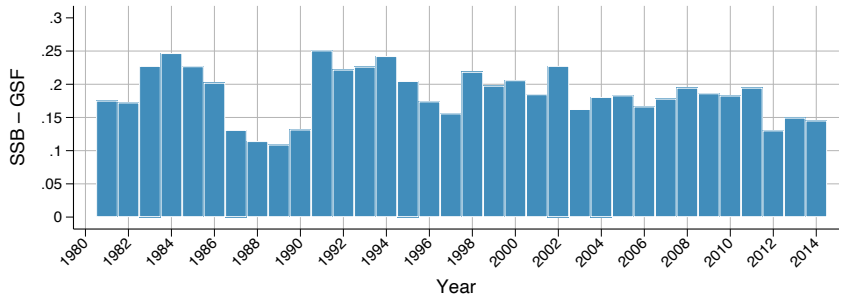
Figure 24. Differences in Volatility by Race



(A) Diff: White



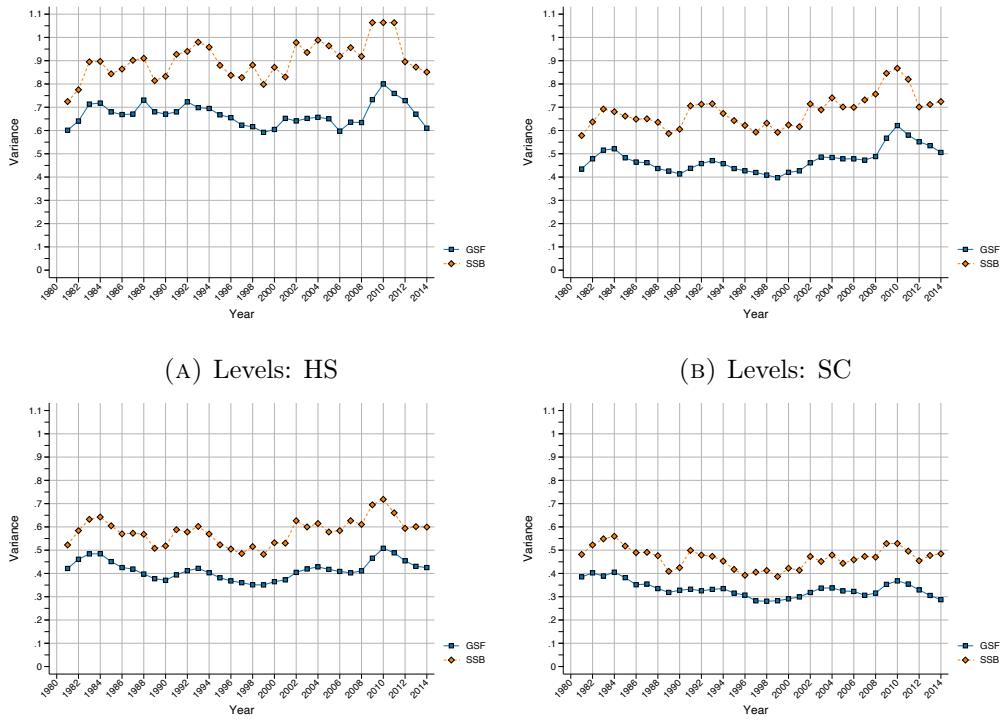
(B) Diff: Black



(C) Diff: Hispanic

Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59. The SSB estimates use one implicate.

Figure 25. Volatility in Arc Changes by Education, Including Zero Earnings

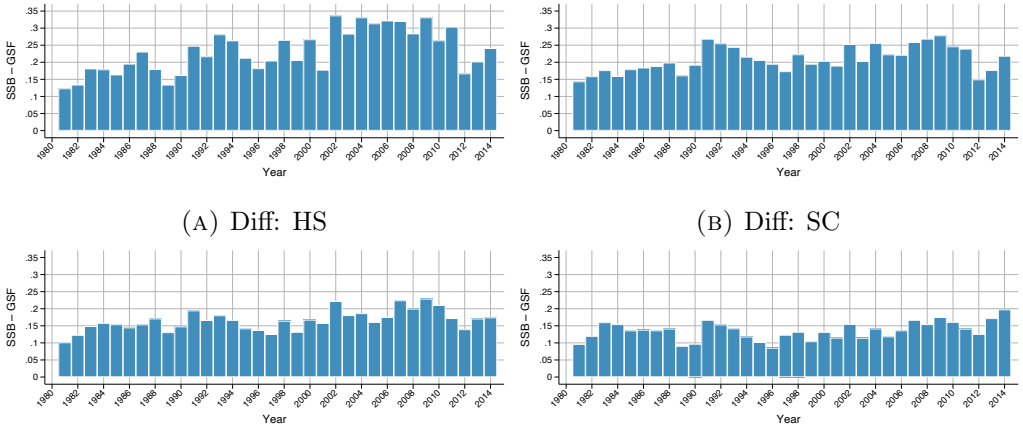


(A) Levels: HS (B) Levels: SC

(C) Levels: BS (D) Levels: Adv

Notes: Authors' calculations based on SIPP GSF and SIPP SSB. Volatility Sample GSF (SSB) is all men 25 to 59 who are 25+ when they are interviewed in the SIPP. The SSB estimates use one implicate.

Figure 26. Differences in Volatility by Education



(A) Diff: HS (B) Diff: SC

(C) Diff: BS (D) 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 and who are 25+ when they are interviewed in the SIPP. The SSB estimates use one implicate.

Appendix A. Appendix

For all estimates in the Appendix, the sample is men age 25 to 59 with positive earnings in two consecutive years.

Figure 27. Inequality, Untrimmed Earnings, 4 Implicates

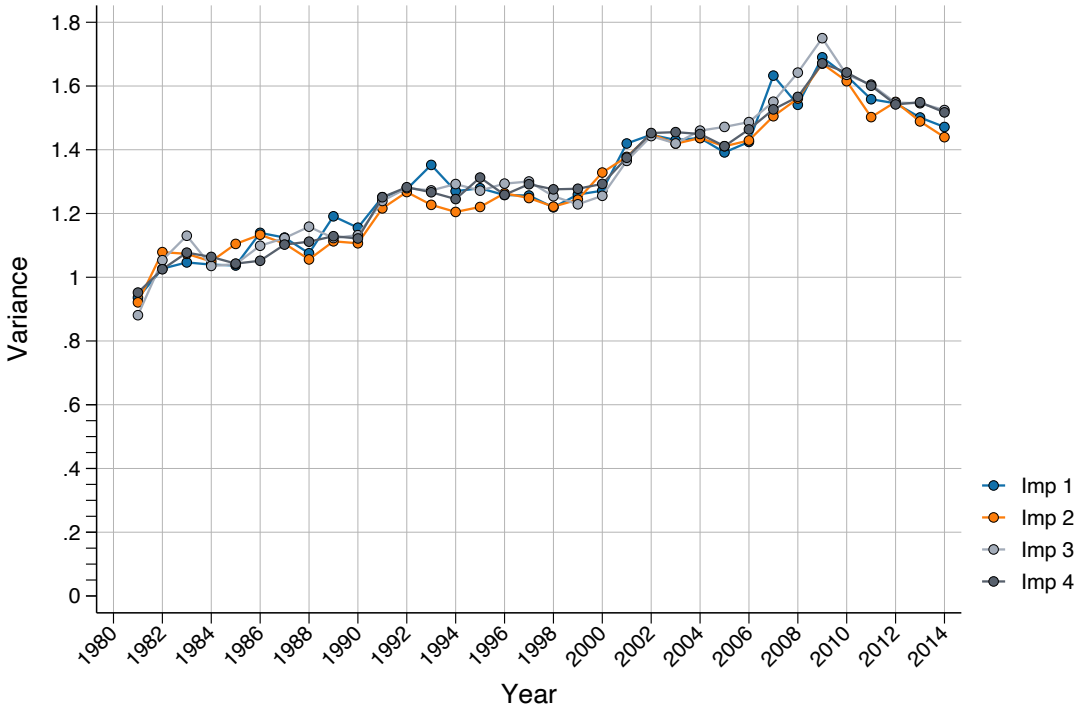


Figure 28. Inequality, Earnings Above 1%, 4 Implicates

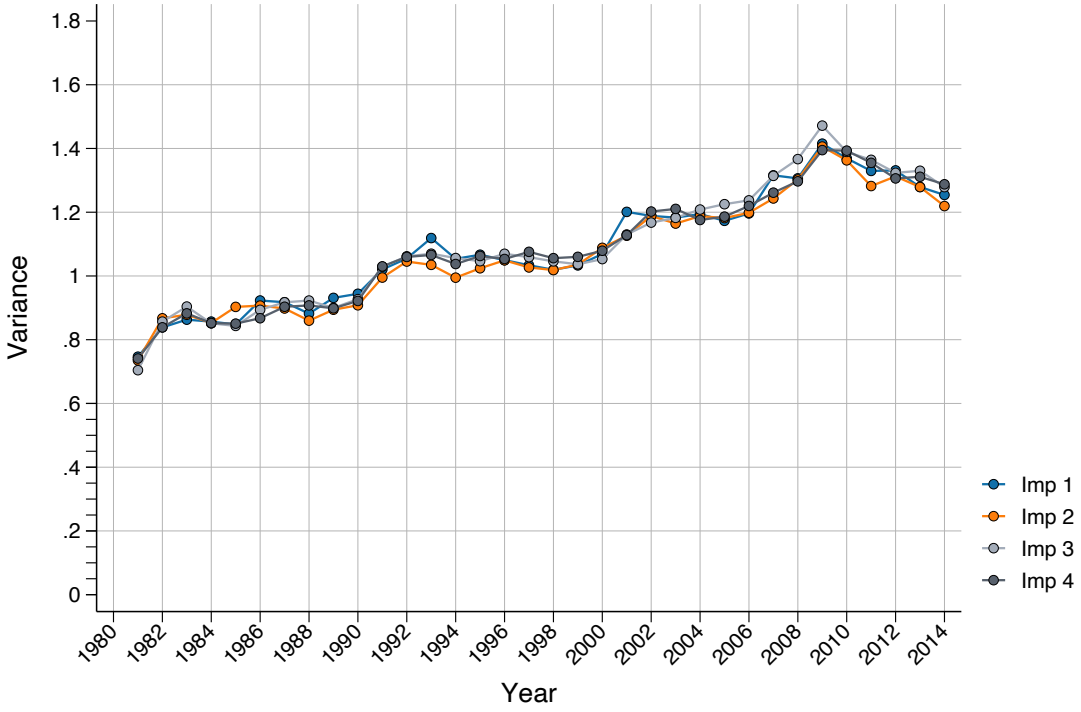


Figure 29. Volatility in Log Changes, Earnings Above 1%, 4 Implicates

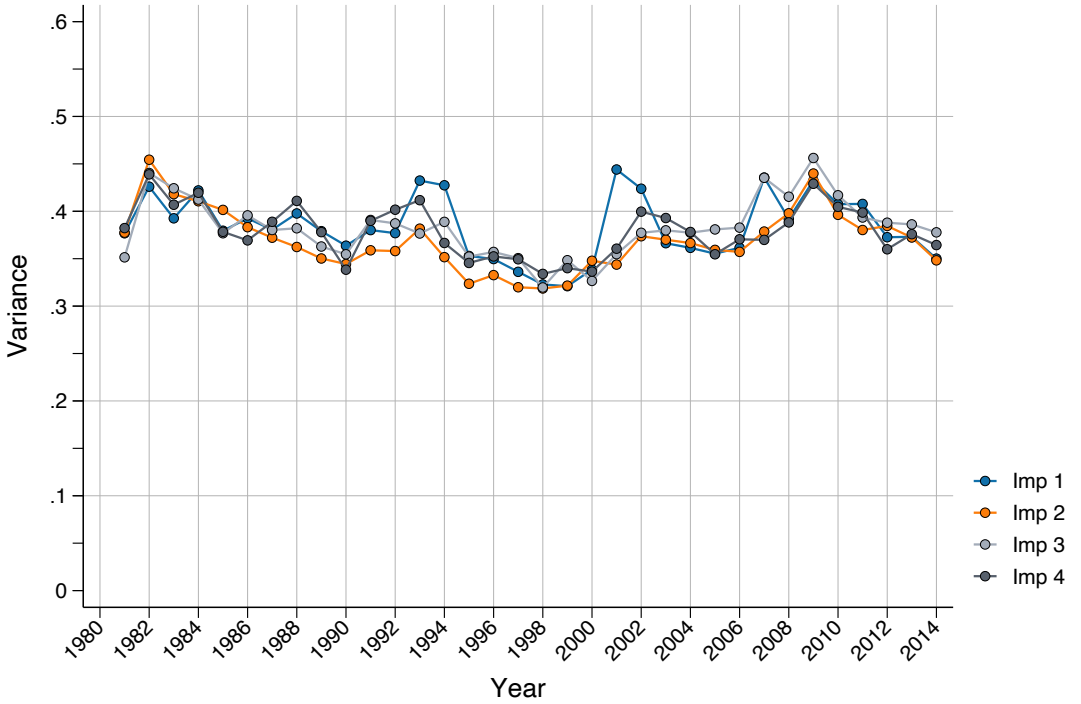


Figure 30. Volatility in Log Changes, Earnings Above 1/4 Full Time Full Year at 1/2 Minimum Wage, 4 Implicates

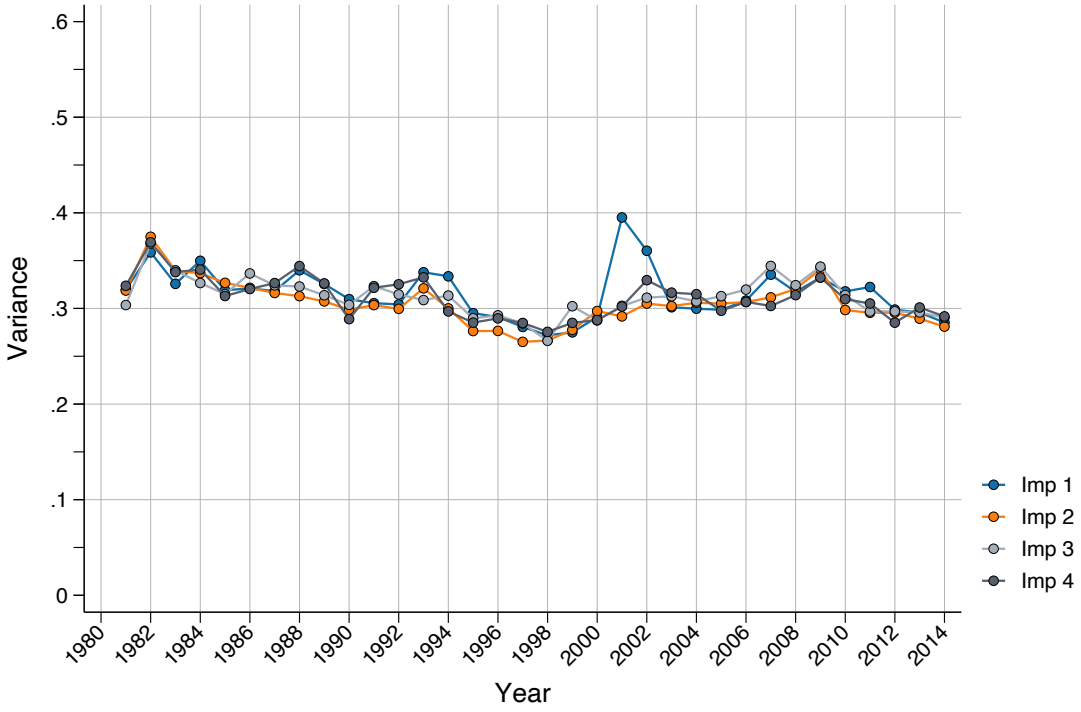


Figure 31. Volatility in Log Changes, Earnings Above 1/4 SSA Annual Earnings Threshold, 4 Implicates

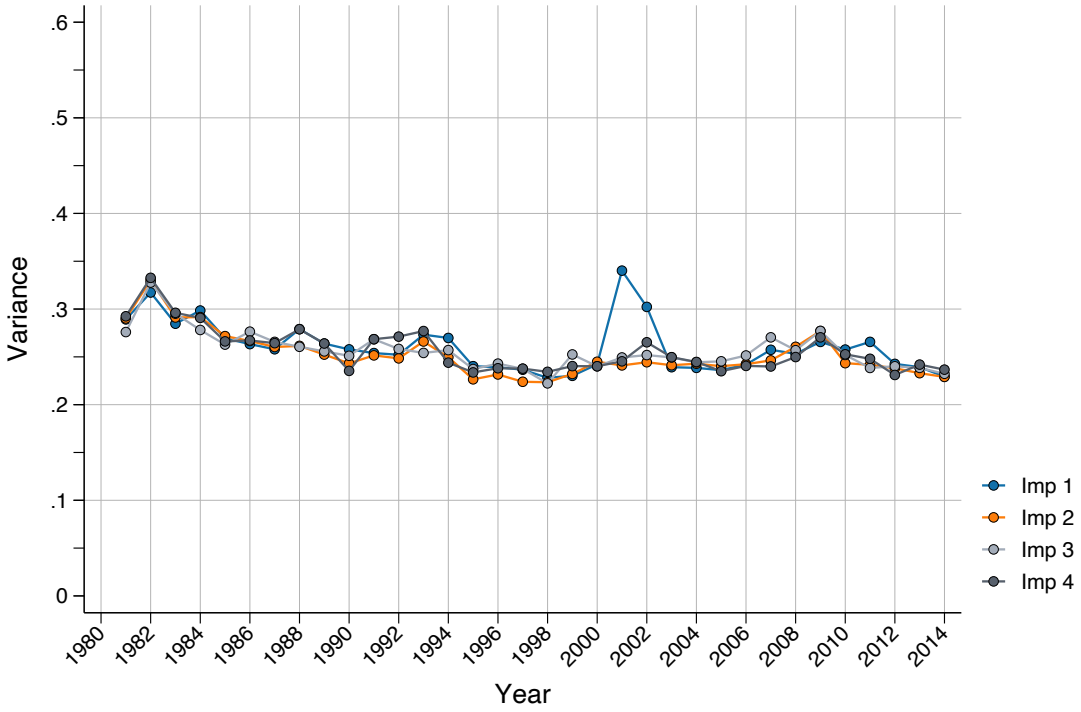


Figure 32. Volatility in Arc Changes, Earnings Above 1%, 4 Implicates

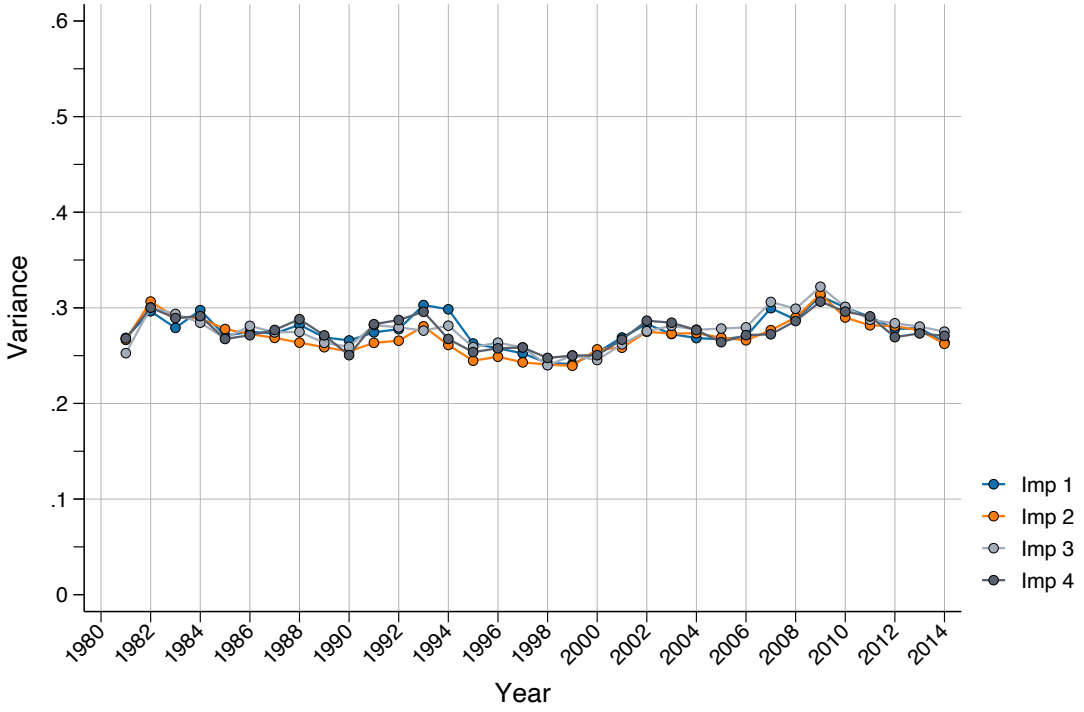


Figure 33. Volatility in Arc Changes, Earnings Above 1/4 Full Time Full Year at 1/2 Minimum Wage, 4 Implicates

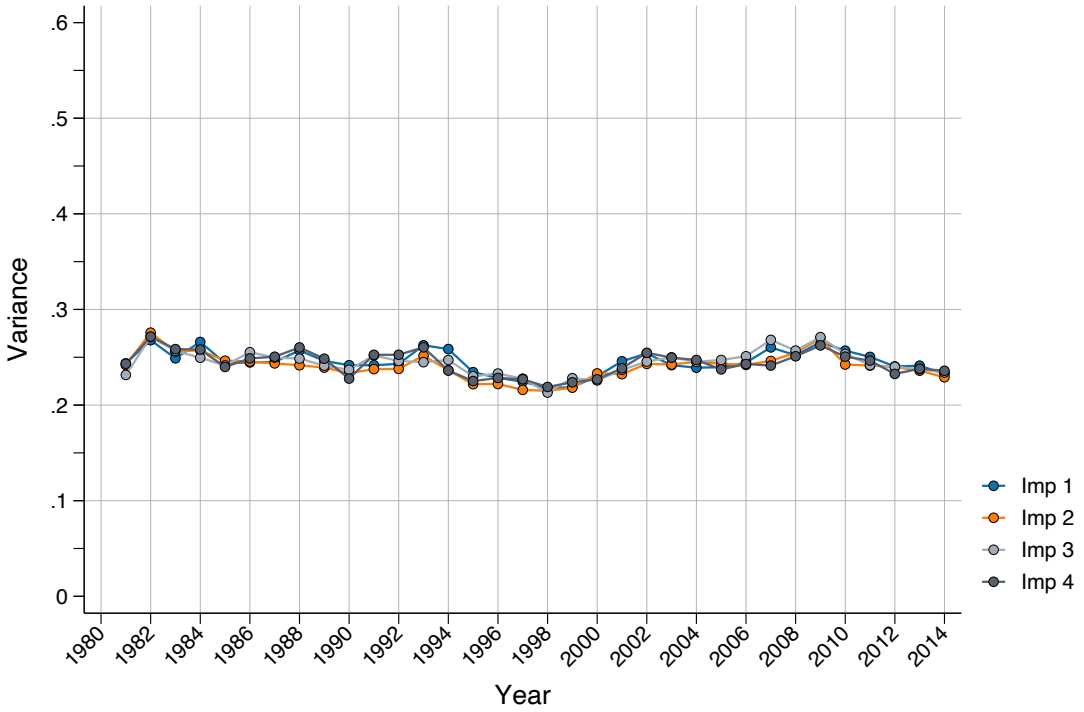


Figure 34. Volatility in Arc Changes, Earnings Above 1/4 SSA Annual Earnings Threshold, 4 Implicates

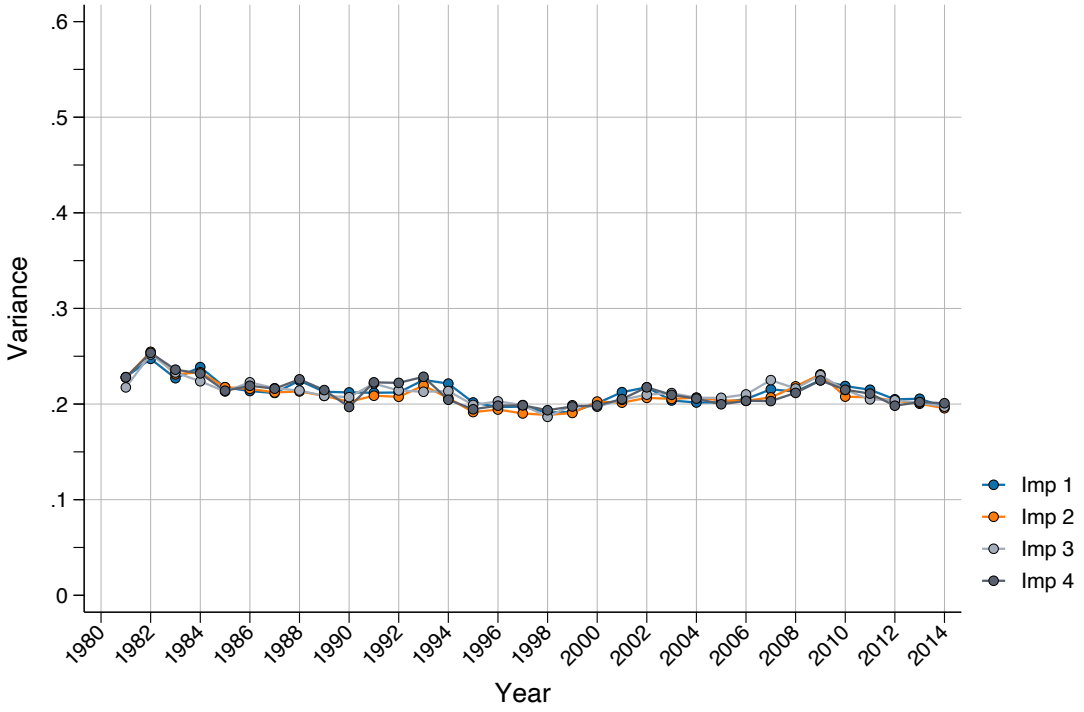


Figure 35. Two-Year Covariance, Untrimmed Earnings, 4 Implicates

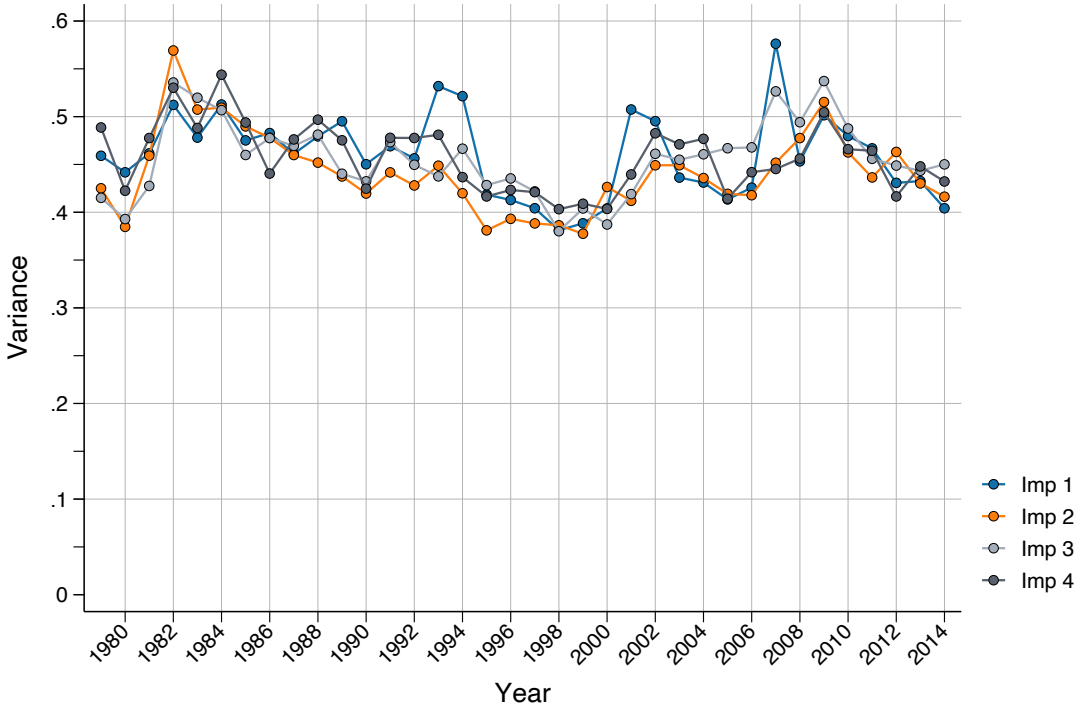


Figure 36. Two-Year Covariance, Earnings Above 1%, 4 Implicates

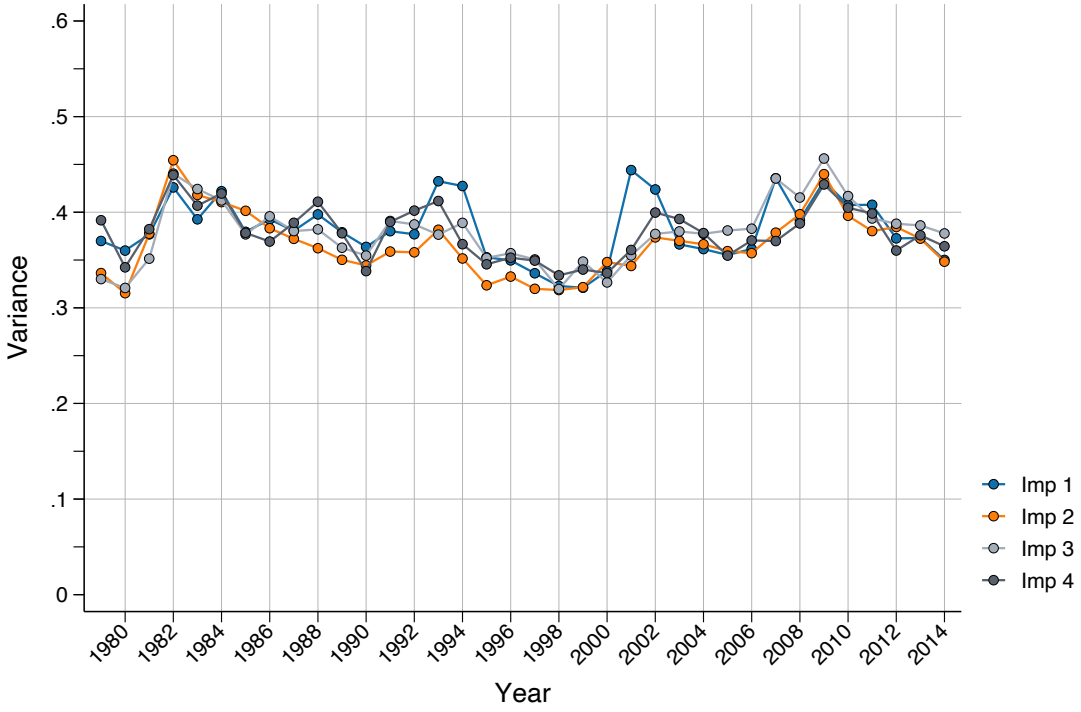


Figure 37. Two-Year Covariance, Earnings Above 1/4 Full Time Full Year at 1/2 Minimum Wage, 4 Implicates

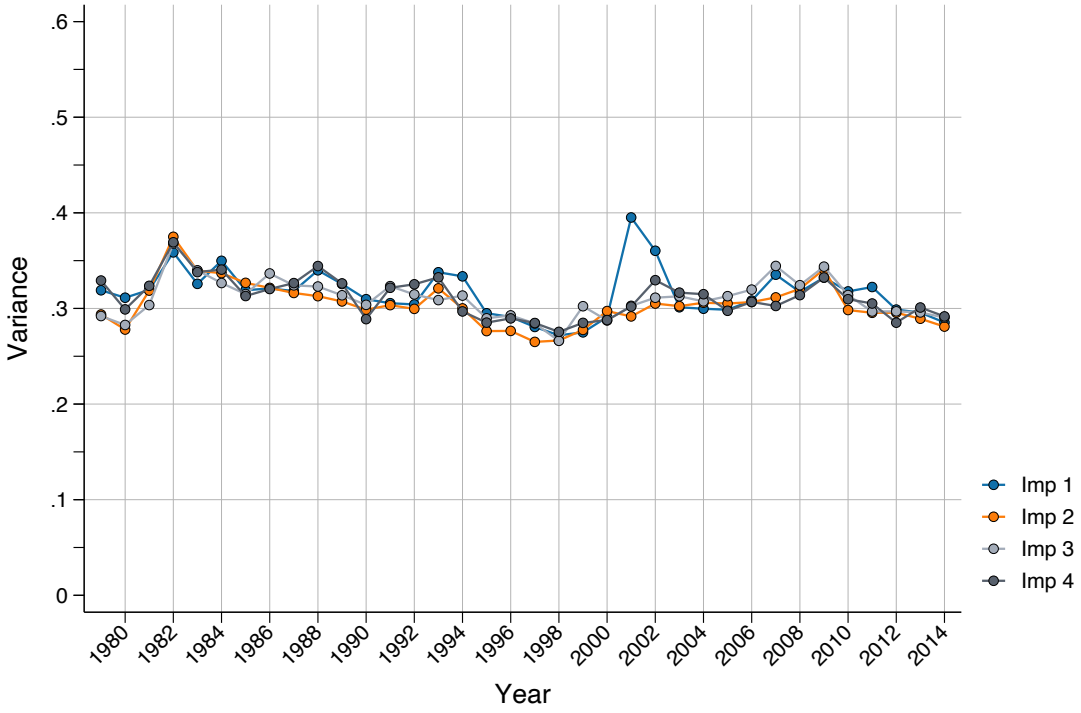
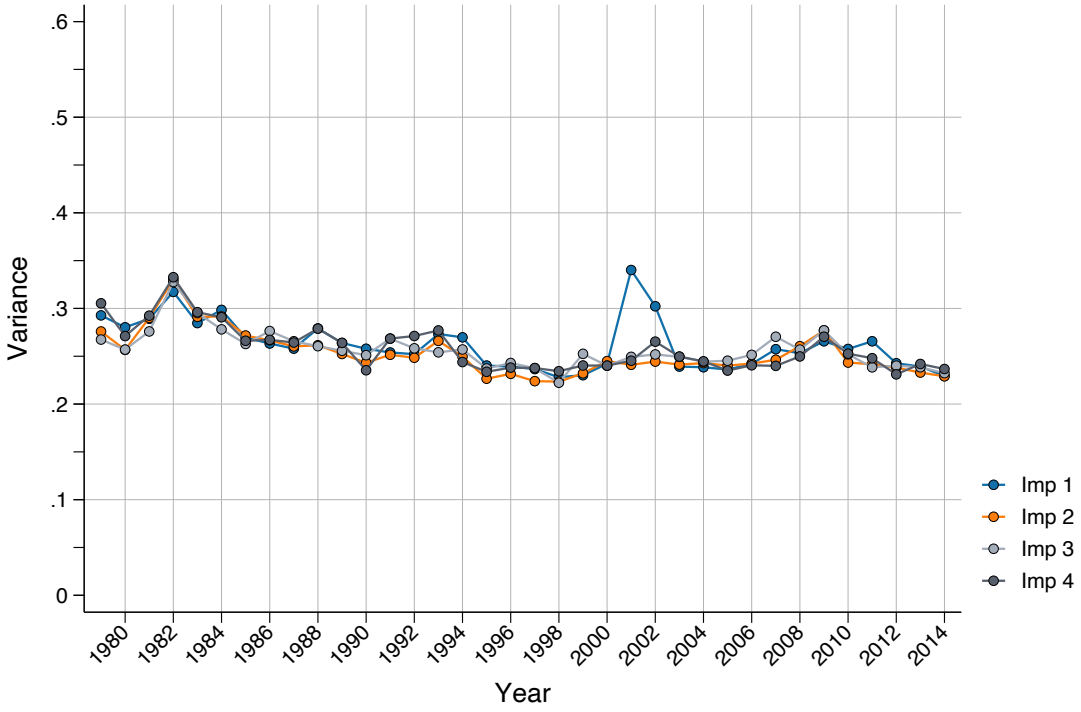


Figure 38. Two-Year Covariance, Earnings Above 1/4 SSA Annual Earnings Threshold, 4 Implicates



References

Abowd, J. M., Stinson, M. H., & Benedetto, G. (2006). *Final report to the social security administration on the sipp/ssa/irs public use file project* [US Census Bureau].

Baker, M., & Solon, G. (2003). Earnings dynamics and inequality among canadian men, 1976-1992. *Journal of Labor Economics*, *21*(2), 289–321.

Benedetto, G., Stanley, J. C., & Totty, E. (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*.

Benedetto, G., Stinson, M. H., & Abowd, J. M. (2013). *The creation and use of the sipp synthetic beta* [US Census Bureau].

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., & Ziliak, J. P. (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.

Bowen, C. M., Bryant, V. L., Burman, L., Khitatrakun, S., McClelland, R., Mucciolo, L., Pickens, M., & Williams, A. R. (2022). Synthetic individual income tax data: Promises and challenges. *National Tax Journal*, *75*(4), 767–790. https://doi.org/10.1086/722094

Carr, M., & Hardy, B. (2022). Racial inequality across income volatility and employment. *Oxford Research Encyclopedia of Economics and Finance*. https://doi.org/9780190625979.013.739

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

Carr, M. D., & Wiemers, E. E. (2018). New evidence on earnings volatility in survey and administrative data. *American Economic Review Papers and Proceedings*, *108*.

Carr, M. D., & Wiemers, E. E. (2021). The Role of Low Earnings in Differing Trends in Earnings Volatility. *Economics Letters*, *199*.

Carr, M. D., & Wiemers, E. E. (2022). The decline in long-term earnings mobility in the u.s.: Evidence from survey-linked administrative data. *Labour Economics*, *78*(78), 102170.

Chenevert, R. L., Klee, M. A., & Wilkin, K. R. (2016). Do imputed earnings earn their keep? evaluating sipp earnings and nonresponse with administrative records. *WORKING PAPER NUMBER SEHSD-WP2016-18, SIPP-WP-275, Census Bureau*, 1–74.

Dahl, M., DeLeire, T., & Schwabish, J. (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), 750–74.

Debacker, J., Heim, B., Panousi, V., Ramnath, S., & Vidangos, I. (2013). Rising inequality: Transitory or persistent? new evidence from a panel of u.s. tax returns. *Brookings Papers on Economic Activity*, *Spring*.

Gottschalk, P., & Moffitt, R. (1994). The growth of earnings instability in the us labor market. *Brookings Papers on Economic Activity*, *1994*(2), 217–272.

Guvenen, F., Ozkan, S., & Song, J. (2014). The nature of countercyclical income risk. *Journal of Polical Economy*, *22*(3), 621–660.

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

Klee, M. A., Chenevert, R. L., & Wilkin, K. R. (2019). Revisiting the shape of earnings nonresponse. *Economics Letters*, *184*, 108663. https://doi.org/https://doi.org/10.1016/j.econlet.2019.108663

Kopczuk, W., Saez, E., & Song, J. (2010). Earnings inequality and mobility in the united states: Evidence from social security data since 1937. *Quarterly Journal of Economics*, *125*(1), 91–128.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, *2*(9), 407–426.

Moffitt, R., & Zhang, S. (2018). Income Volatility and the PSID: Past Research and New Results. *American Economic Association Papers and Proceedings*, *108*, 277–80.

Moffitt, R. A., Abowd, J., Bollinger, C., Carr, M. D., Hokayen, C., McKinney, K., Wiemers, E. E., Zhang, S., & Ziliak, J. (2023). Reconciling Trends in Volatility: Evidence from the SIPP Survey and Administrative Data. *Journal of Business and Economic Statistics*, *41*(11), 1–11.

Moffitt, R. A., & Gottschalk, P. (2012). Trends in the transitory variance of male earnings methods and evidence. *Journal of Human Resources*, *47*(1), 204–236.

National Academies of Sciences, Engineering, and Medicine. (2023). *A roadmap for disclosure avoidance in the survey of income and program participation*. National Academies Press. https://doi.org/10.17226/27169

Raghunathan, T. (2021). Synthetic Data. *Annual Review of Statistics and Its Application*, *8*, 129–140.

Raghunathan, T., Lepkowski, J. M., & Stolenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, *27*(1), 85–95.

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

Sabelhaus, J., & Song, J. (2009). Earnings volatility across groups and time. *National Tax Journal*, *2*(62), 347–364.

Sabelhaus, J., & Song, J. (2010). The great moderation in micro labor earnings. *Journal of Monetary Economics*, *57*.

Shin, D., & Solon, G. (2011). Trends in men's earnings volatility: What does the panel study of income dynamics show? *Journal of Public Economics*, *95*(7), 973–982.

Stanley, J., & Totty, E. (2021). A penny synthesized is a penny earned? an exploratory analysis of accuracy in the sipp synthetic beta. *U.S. Census Bureau Working Paper*, *CED-WP-2021-006*.

Ziliak, J. P., Hardy, B., & Bollinger, C. (2011). Earnings volatility in america: Evidence from matched cps. *Labour Economics*, *18*(6), 742–754.