# A Penny Synthesized is a Penny Earned?
# An Exploratory Analysis of Accuracy in the SIPP Synthetic Beta*

Jordan Stanley, U.S. Census Bureau

Evan Totty, U.S. Census Bureau†

April 21, 2023

**ABSRACT**

The Census Bureau has expressed interest in using modern synthetic data modeling techniques for privacy and confidentiality protection in future microdata releases. In order to aid understanding of the accuracy and usability of synthetic microdata going forward, we perform an exploratory analysis comparing results generated using an early synthetic microdata release known as SIPP Synthetic Beta to results from the same analyses using the corresponding confidential microdata. We compare numerous descriptive and model-based use cases of the data and discuss explanations for how performance of the synthetic data relates to modeling decisions by the data provider and methodology choices by the data user. We also summarize differences in confidence interval overlap and statistical conclusions. There is a strong association between the GSF and SSB results in terms of both magnitudes and statistical conclusions, but the SSB is not a perfect replication of the GSF. Finally, we discuss the implication of our results for the role of modeling decisions and user feedback when creating synthetic data, validation and verification options, and the evolving science of creating synthetic data. Importantly, we consider our findings to be something of a lower bound for the accuracy of future synthetic microdata because of improvement in synthetic data modeling since the SSB was created and the fact that we do not account for other sources of survey error when comparing the confidential data to the synthetic data.

*The authors would like to thank Gary Benedetto, Jason Fields, Caleb Floyd, Robert Moffitt, Joanna Motro, Rolando Rodriguez, and seminar participants in the Census Economic Research Brown Bag Series and the National Academies Committee on National Statistics meeting for *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation* for their comments and feedback. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the Census Bureau or other organizations. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product. This research was performed under Census project P-6000562. Data from the SIPP Gold Standard File are confidential. (Disclosure clearance numbers: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025, CBDRB-FY20-CED001-B0003, CBDRB-FY21-CED002-B0003, CBDRB-FY21-195, CBDRB-FY21-285, and CBDRB-FY23-CED009-0001). This paper updates a version previously published online as a Census Bureau working paper (https://www.census.gov/library/working-papers/2021/adrm/CED-WP-2021-006.html).

†Author contact information: jordan.c.stanley@census.gov; evan.scott.totty@census.gov.

## 1. Introduction

In this paper we study how the results from socioeconomic empirical analyses differ between the U.S. Census Bureau's Survey of Income and Program Participation (SIPP) Gold Standard File (GSF) and a fully synthetic version of the same file known as the SIPP Synthetic Beta (SSB). Synthetic microdata replace observed data with imputed data based on underlying models to both increase privacy protection and release information that would not otherwise be available to the public (Little, 1993; Rubin, 1993). Synthetic datasets are often supported by a validation or verification option that allows users to receive output from the internal data after preparing their analysis on the synthetic data. Synthetic microdata are likely to play a key role in the future of data releases due to rising risks of reconstruction and reidentification attacks (Abowd et al., 2019; Abowd et al., 2020). Synthetic data are not yet widely used but are increasingly viewed as a realistic option for providing data users with access to detailed microdata while reducing privacy risk (Drechsler and Haensch, 2023). Still, little is known about strengths and weaknesses in terms of how accurately synthetic data can be expected to replicate results derived from the internal data. Our paper represents an exploratory analysis for understanding the accuracy and utility of synthetic data for a variety of applications of interest to social science researchers. Specifically, we generate descriptive statistics and perform model-based analyses using the SSB and compare those results to the results from running the same analyses using the SIPP GSF.

The purpose of creating the SSB was to provide the public with data from the SIPP linked to administrative records on earnings and benefit receipt from the Internal Revenue Service (IRS) and Social Security Administration (SSA) (Benedetto et al., 2013). Because of confidentiality issues associated with linking the SIPP to administrative records, it was not feasible to release such data directly to the public. Instead, a synthetic version of the data was created by modeling the GSF through sequential regression multivariate imputation (SRMI).[1] Extensive testing was done to ensure the analytic validity of the SSB; however, the Census Bureau cannot guarantee fitness of use for all possible use cases. Therefore, the Census Bureau offers researchers the option to have their results "validated" on the confidential SIPP GSF. The researchers submit their code to the Census Bureau whose employees then run the code on the confidential data. Traditional statistical disclosure limitation (SDL) techniques such as rounding are applied to the results which

[1] The SSB does not satisfy formal privacy, including the most well-known variant – differential privacy (Dwork et al., 2006).

are then released back to the SSB user. Feedback from these validations also helps the Census Bureau improve the development of the synthetic data models.

The main contribution of our paper is to present new evidence on how research using a synthetic dataset compares to the same research performed using the corresponding confidential data. Little research exists that shows how empirical analyses differ between synthetic and confidential data. Many researchers have used the SSB or other similar synthetic datasets with validation to conduct research, but rarely have authors presented results from both the confidential and synthetic data in their papers.[2] With the validation option, it is not necessary that the SSB or other synthetic datasets with validation accurately replicate every statistical relationship between variables in the confidential data; however, validation is costly in terms of validation resources for the data provider, wait time and higher coding standards for the data user, and leakage of respondent privacy. It is therefore useful to assess how well the SSB replicates results from the GSF and whether there are certain types of analyses that are more likely to need validation than others. To our knowledge, our study has been the first to analyze such differences while considering a wide range of socioeconomic research questions using a household-level survey.

We focus our present analysis on applications related to person-level earnings. Earnings is an important variable in socioeconomic research, so assessing the performance of synthetic data in investigating such research questions is essential. We also have two different sources of earnings information in the GSF and SSB – self-reported earnings from the SIPP and administrative earnings information from the IRS. Having two measures of earnings provides an opportunity to investigate some unique research questions such as how well the SSB replicates relationships associated with each earnings measure separately and how well it replicates differences between the two measures.

We find that the SSB does a good job of replicating many of the empirical results we generated using the GSF, but the SSB produces different results in some of the applications.[3] For

[2] Among the few examples is Benedetto et al. (2010), which studied the effects of graduating during a recession on earnings and reported results from both the GSF and SSB. Kinney et al. (2011) and Kinney et al. (2014) described the creation of a synthetic version of the U.S. Census Bureau's Longitudinal Business Database (LBD) and in doing so compared some results between the confidential and synthetic data. Bowen et al. (2020) reported how descriptive statistics in IRS data differed between the confidential data and a synthetic version.

[3] We would like to stress that we will specifically use the term "replicate" to describe how similarly the statistical output generated using the SSB compares to the same statistical output generated using the GSF. We would also like make clear that all such statements about replication or similarity/differences are based on sign and statistical significance comparisons and/or visual evaluation (e.g., directly comparing magnitudes) rather than formal statistical testing.

example, the SSB produces median earnings by gender, race, and education that almost exactly match those from the GSF, but the mean earnings by group are less similar. The SSB also replicates median within-person ratios of IRS earnings to SIPP earnings almost exactly but does less well at replicating mean within-person earnings ratios. For our model-based results, the SSB produces similar results to the GSF for national trends over time in many regression-adjusted statistics: both show a rising college wage premium, a declining gender wage gap, and the expected life cycle earnings trajectory (i.e., individuals' earnings grow as they gain work experience but then decline as they approach retirement ages). The SSB often does not replicate the GSF results for analyses that rely on within-person variation in earnings over time such as person fixed effects models and models with person-specific coefficients. We also find that the SSB often closely aligns with the GSF on the magnitude and statistical significance of regression coefficients for variables included in the GSF and SSB, but sometimes fails to replicate regression findings that involve merging on external data not in the GSF or SSB such as state minimum wages. Most of these findings are consistent with expected and interpretable patterns related to modeling decisions, merging external data, and estimators that are sensitive to outliers. Across our model-based results, we find that the SSB confidence interval overlaps with 33% of the GSF confidence interval, on average, and covers the GSF coefficient estimate 35% of the time. In terms of sign and significance, the SSB produces the same sign 79% of the time, the same statistical conclusion 63% of the time, and the opposite statistical conclusion only 2% of the time.

Our results lead us to three key takeaways, which we will discuss in greater detail later in the paper. First, modeling decisions inherently prioritize particular use cases and feedback from data users is mutually beneficial. Second, validation and/or verification are important complements to synthetic data. Third, the science for generating and evaluating synthetic data has advanced in the years since the SSB was first developed and is still evolving. This is an evaluation of one synthetic data set (the SSB) created using a particular methodology (SRMI), and the accuracy findings should not be presumed to represent other datasets or synthetic data in general. While SRMI was the frontier of science for generating synthetic data when the SSB was first developed in 2003, non-parametric classification and regression tree (CART) methods can generate more accurate synthetic data while also being easier to implement (Drechsler and Reiter, 2011; Reiter, 2005; Reiter and Kinney, 2012). Finally, we want to emphasize that our results come with an important caveat. We characterize the tendency of results from the SSB to replicate results from

the GSF as a measure of accuracy. Implicit in this characterization is the assumption that the confidential data represents the "truth" and that the GSF therefore provides full/maximum accuracy. This assumption is correct if the confidential data have no error, but we know that survey and administrative data already contain other types of error. We ignore these other sources of error in our current paper, but we acknowledge this limitation and provide more discussion later in the paper.

We proceed with background information on synthetic data and the history of the SSB in Section 2. We describe our analyses and present the results in Section 3. We discuss takeaways and caveats related to our results in Section 4. Finally, we summarize our conclusions and discuss future research goals in Section 5.

## 2. Background on Synthetic Data

The identity and information of survey respondents have become increasingly difficult to protect. Large amounts of data are publicly available, and computers and statistical methods are more advanced than ever before. Traditional methods of disclosure avoidance are no longer sufficient in this evolving landscape. Protecting personal information is critical, but protective measures can have adverse effects of the usability of the data made available to the public. Presenting less accurate data protects respondent information but reduces the usefulness of the data set. This tradeoff is at the heart of research in the field of disclosure avoidance. The ultimate goal is to provide a high degree of protection of respondents' identity and information while also maintaining a high level of data usefulness.

The U.S. Census Bureau is striving to address these concerns. The Census Bureau is required by law to protect survey respondents' information and identity (Abowd et al., 2020). At the same time, the data made available to the public have many benefits which are inherently tied to the accuracy of the information. Concerns over the accuracy of data exist independent of privacy concerns; survey non-response and measurement error have been increasing over time (Meyer, Wallace, and Sullivan (2015)). While increasing data protections could also reduce accuracy of the final data product, some evidence suggests that not having public trust that data are protected could affect data accuracy through inaccurate responses (Gruzd and Hernández-García, 2018) or through declining response rates (Couper et al., 2008).

Surveys must balance the need to protect the information for millions of respondents while producing accurate statistics and data for public use. Synthetic data present one avenue for protecting privacy. "Synthetic" refers to a dataset where some or all of the data released are based on modeled estimates from confidential data. The models used to create the synthetic data change the original values in order to protect privacy, with the goal of also maintaining covariate relationships in order to reduce potential accuracy loss. Synthetic data can come from partial synthesis in which only some variables and/or observations are synthesized; all variables and all observations have been synthesized in fully synthetic data. The similarity of synthetic data compared to the internal data can of course vary based on the models used. A thorough survey of the history of synthetic data, including its origin, existing uses, and methodologies, can be found in Drechsler and Haensch (2023).

One mechanism for addressing the potential shortcomings of the synthetic data is an internal validation and/or verification system. These systems allow users to work with the publicly available synthetic data and then request that their analysis be run on the corresponding internal, confidential data. A validation system releases the results from the internal data to the user (with some disclosure avoidance protection applied), while a verification server returns some summary measure of similarity between the synthetic and internal results. The synthetic data protect the privacy of the underlying internal data, and the validation/verification system can address accuracy concerns by providing statistical output based on the internal data. Having synthetic output that reproduces many statistical results from the corresponding internal data could reduce the number of validation requests, thereby mitigating costs associated with validation.[4] Of course, higher similarity between the internal and synthetic data likely carries higher disclosure risk, so the overarching balance between privacy and utility must still be considered.[5]

The U.S. Census Bureau has a history of implementing a validation system. The GSF links microdata from select panels of the SIPP to several administrative data sources such as the SSA and IRS. The SSB is a synthesized version of the GSF featuring a subset of variables found on the internal data sets. The SSB has evolved over time. Version 1.0 was created in 2003. Version 4.0

[4] For the data provider, validations and verifications require computing resources and employee time. For the data user, validations and verifications require a high standard of coding needed for replicability and a lag between finishing an analysis and receiving the results.

[5] Furthermore, as more results are released based on an internal data set, there is higher disclosure risk that someone could pool all released results and re-identify someone in the data or even reconstruct a record or entire dataset (Federal Committee on Statistical Methodology, 2005).

was the first to be released to the public, with the release occurring in 2007. Version 7.0, released in 2018, is the most recent release. Version 7.0 is fully synthetic, including even the missing data pattern; the SSB was only partially synthetic in prior versions. In Version 7, four synthetic files are created from a "snapshot" of the internal Gold Standard File. [6] The snapshot serves as the data set for internal validations. Additional technical information on the development of the SSB is available in Benedetto, Stanley, and Totty (2018).

Data users can apply for access to the publicly available synthetic files. Once approved users have successfully built and run their analysis on the SSB, they can submit requests for validations of their code to internal Census employees who will run the analysis code on the internal file on behalf of the data user. The statistical output is checked by Census employees to ensure it meets all disclosure avoidance requirements and then, if approved, the output is released to the data user. The GSF and SSB have been used to study topics such as earnings gaps, disability insurance, returns to education, lifecycle earnings, retirement outcomes, minimum wage effects, and many others. Examples of articles published in peer-reviewed journals include Bertrand, Kamenica, and Pan (2015); Juhn and McCue (2016); Henriques (2018); Neumeier, Sorensen, and Webber (2018); and Kejriwal, Li, and Totty (2020).

## 3. Quantitative Analysis of Synthetic and Confidential Output

When considering synthetic data and statistical estimates based on synthetic data, the comparability of estimates to those generated using the corresponding confidential internal data is of high interest. We perform several analyses in the present paper – all of which involve earnings data. Specifically, we perform descriptive analysis, check missing data patterns, and run regression analysis. For the latter, we emphasized well-known statistical relationships for which, in most cases, there exists an expected finding (e.g., gender wage gap estimates typically should show that males earn more than females on average). We focus on earnings for multiple reasons. First, earnings is a critical outcome of interest in the field of economics as well as in other social science disciplines. There are thus many use cases that we can test. Second, earnings is a continuous

[6] Results based on the SSB must be combined across all four synthetic files for proper estimation and inference. Statistics such as means and regression coefficients can simply be averaged across the four files. Correct inference must use the variance combination formulas in Benedetto, Stanley, and Totty (2018). All SSB results reported later in the present paper are based on the proper combination of results across the four files.

measure, so we can test out distributional differences while also assessing extensive margins (e.g., positive earnings or not). Third, the GSF and SSB have earnings information from two sources: the administrative records and the survey self-responses. This gives us another avenue to assess differences between the confidential and synthetic data sets.

We use several samples based on the GSF and SSB. In assessing missing data patterns, we utilize the full GSF and SSB data sets. Our primary analytical samples consist of individuals who have non-missing earnings in both the administrative records – specifically, the Detailed Earnings Record (DER) from SSA – and survey responses from the SIPP. Since the DER is annual and SIPP observations are monthly, we calculate annual earnings for the SIPP by summing monthly earnings for a full calendar year. So, individuals in our main analytical samples have non-missing SIPP earnings data for all twelve months in a given calendar year.

When not replicating select papers that used the GSF, we have a primary analytical subsample of interest consisting of positive earners – individuals from our full sample who have both DER earnings and annual SIPP earnings greater than zero for at least one calendar year. As the natural logarithm of earnings is often our regression outcome of interest, this subsample typically serves as the basis for our regression samples. Each regression analysis includes additional restrictions such as non-missing covariates and assorted age range limitations. These sample definitions are described in the respective sections pertaining to each separate regression analysis.

First, we will discuss the descriptive statistics, including summary statistics and patterns of extensive margins. Then we provide descriptions of and results from each regression analysis we perform.

## 3.1 Descriptive Analysis

### *3.1.1 Earnings Summary Statistics*

Figure 1 shows a kernel density plot for the distribution of SIPP and DER earnings in the SSB and the GSF. The results are based on our positive earners sample described above. We also drop the top and bottom five percent of annual earnings observations for each dataset and earnings source. Both SIPP and DER earnings show a hump-shaped distribution with the largest density between $10,000 and $20,000 and a long right tail. The SSB does a good job of replicating this basic shape, although the SSB has even larger density at the peak and shifts the peak to the left.

Consequently, the SSB shows less density for earnings between $20,000 and $80,000. The main takeaway is that the SSB has a larger density of individuals with low earnings levels (less than $20,000). Thus, the SSB replicates the general tendency of earnings to be right-skewed, but the densities are different in a way that could impact income inequality estimates based on, for example, decile comparisons.

Figures 2 and 3 study how well the SSB replicates differences between the SIPP and DER earnings. Figure 2 shows the average difference between SIPP and DER earnings (SIPP minus DER) across individuals by DER earnings decile. The GSF version of the figure shows evidence of over-reporting in survey earnings in the lower part of the DER distribution and under-reporting of survey earnings in the upper part of the DER distribution. This result is an important one because it suggests that measurement error in survey earnings is not classical but rather is related to the individual's true earnings. The SSB replicates this result quite well, both in terms of the over-reporting versus under-reporting pattern and in terms of the average difference in each decile. The only noticeable differences are that the SSB version shows even larger earnings differences in the outer deciles and the SSB version shifts the inflection point between a positive and negative average earnings difference from the sixth DER decile to the seventh DER decile.

Figure 3 shows a histogram for the difference in individual earnings between the SIPP and DER (SIPP minus DER). We compute the SIPP minus DER difference for each individual and then count the number of individuals in each of 21 bins. The top figure based on the GSF shows that while more individuals fall into the zero ("0") bin (SIPP minus DER earnings difference of -$1,000 to $1,000) than any other, there is a spread around the zero bin with greater density in bins closer to the zero bin. The two tails have the next most density of all bins other than the zero bin, indicating the presence of many individuals who have large (larger than -$10,000 or $10,000) differences between their SIPP and DER earnings. The bottom figure based on the SSB once again replicates the general shape seen with the GSF: a spread around the zero bin with the most common bins being the zero bin and the tails. However, there are some noticeable differences in the SSB version relative to the GSF version. First, the tails have greater density than the zero bin. Second, the spread around the zero bin is flatter. These results show that while the SSB does replicate the general tendency of SIPP and DER earnings to be similar to each other, synthesizing the data weakens the relationship.

Figures 4 through 7 show the means and medians of real earnings across our two main samples and assorted demographic groups. The four bars in each figure correspond to the sources of the earnings information – either the GSF or the SSB, and either the DER or the SIPP. Figures 4 and 5 correspond to the full sample, while Figures 6 and 7 pertain to the sample of positive earners (i.e., individuals with both positive DER earnings and positive SIPP earnings). For the means, one can see in Figures 4 and 6 that the variation across data sources is highest for the highly educated subgroups. The other demographic groups show fairly similar means within each category, and the pattern is similar comparing different categories (e.g., comparing white to Black or men to women). In assessing GSF means vs. SSB means, across groups the SSB means are slightly higher.

The median earnings are presented in Figures 5 and 7. Comparing the GSF and SSB medians shows practically no difference within each of the samples analyzed. Further, the SIPP medians are slightly higher than the DER medians; however, that patterns hold across demographic groups and in comparing GSF to SSB. The expected relative differences between demographic groups are also present regardless of the data source used for the statistical calculation.

Figures 8 and 9 shows statistics for differences between the administrative record values for earnings (the DER) and the self-reported values from the SIPP. Figure 8 presents the ratio between these earnings values while Figure 9 shows their absolute difference. In both, the bars now correspond to either the GSF or the SSB and either the mean or median difference. For Figure 8, the mean ratio (DER/SIPP) is far higher in the SSB compared to the GSF, and that pattern holds across subsamples. The medians are again nearly identical comparing the SSB to those from the GSF, which fits with the findings presented in Figures 5 and 7. Turning to the absolute differences (see Figure 9), the SSB is once again much higher than the GSF in terms of mean difference; however, the patterns across and within subsamples are consistent. For example, subsamples with larger average earnings values (e.g., men relative to women, or advanced degrees relative to high school degrees) show larger average differences between the DER and SIPP in both the GSF and the SSB. Looking at the medians, there are once again minimal differences comparing the estimated values in the SSB to the corresponding values in the GSF.

Our final descriptive check for within-person variation is comparing the standard deviations of within-person earnings in the GSF to that in the SSB. The sample for this analysis keeps individuals aged 30 through 61 who had at least three non-missing DER earnings

observations in the data. We calculate the standard deviation in real earnings for each individual in the analytical sample and then calculate several moments of interest. These results are included in Table 1. Within-person standard deviations are larger in the SSB than in the GSF. As in the other descriptive analysis, the percentiles are more similar than the means. For example, the mean standard deviation of within-person earnings in the SSB is roughly double that for the GSF. For comparison, the SSB median is roughly 42% higher than the GSF median.

In sum, the descriptive statistics analyzed here show few major differences comparing the results from the SSB to the results from the GSF. In particular, the results showing the largest differences seem to come from analysis that is more sensitive to outliers. Statistics like medians are nearly identical when comparing those generated from the SSB to those estimated with the GSF data. Static patterns seen in the GSF (e.g., the gender earning gap, education premia, Black-white earnings gap) are likewise estimated in the SSB.

*3.1.2 Positive, Non-Positive, and Missing Earnings Patterns*

This section shares results from analysis of different categorizations of earnings values – positive earnings (i.e., non-missing and greater than zero), non-positive earnings (i.e., non-missing and less than or equal to zero), and missing earnings. For missing earnings, we focus on records with a missing value for an individual in-universe. Put another way, we are interested in missing values for which there *should* be a non-missing value (e.g., person declined to respond to SIPP question when asked) rather than records for which a missing value is structural (i.e., the individual wasn't in universe). Note that for the DER earnings in our sample, missing values are present when individuals from the SIPP could not be linked to the administrative records.

Using the full samples, we calculate descriptive statistics to compare the missing data pattern for earnings in the SSB to that in the GSF. For SIPP earnings, we check each month that appears in our primary analytical samples (i.e., individuals with non-missing calendar year earnings in both the DER and the SIPP). Since we require full calendar years in our other analysis, the second, third, and fourth calendar years of a SIPP panel are potentially eligible. We calculate the percentage of missing earnings for each month in this time frame in the SSB and GSF and then take the difference between those percentages. The largest difference for a single month analyzed is 0.41 percentage points. We perform the same exercise for annual SIPP earnings (i.e., earnings is missing for at least one month in the calendar year) for each relative calendar year within the

panel – the largest difference between the SSB and GSF is 0.12 percentage points. The difference in the percentage of individuals with missing SIPP earnings in any year (i.e., flagging individuals with missing SIPP earnings for any of the second, third, or fourth calendar year of their panel) is roughly 0.20 percentage points. The difference in the percentage of individuals with missing DER earnings is roughly 0.16 percentage points. We also compare missing and non-missing values for SIPP annual earnings vs. DER earnings. The maximum percentage point difference between the GSF and SSB for records with DER missing and annual SIPP non-missing is 0.14 while the same estimate for DER non-missing and annual SIPP missing is 0.35. We interpret these statistics as evidence that the overall prevalence of missing earnings data in the SIPP is quite close to that in the GSF.[7]

We also perform descriptive analysis of positive vs. non-positive earnings in the SIPP and DER comparing the SSB to the GSF. Our aim here is to assess extensive margin differences between self-reported earnings and administrative records as well as whether any such differences are seen in both the SSB and GSF. Figure 10 shows the proportions of records with different combinations of positive and non-positive earnings in the different data sources and data sets. The figure format is like that in Figures 4 through 9. Looking within assorted samples and subsamples, we measure the proportion of records fitting the given characteristic (e.g., positive SIPP earnings and non-positive DER earnings in the GSF). Our main comparison of interest is between the GSF and SSB. So, the first and second bars for each x-axis category show how many records have positive SIPP earnings and non-positive DER earnings, while the third and fourth bars show records with non-positive SIPP earnings and positive DER earnings. The "positive DER and non-positive SIPP" proportions are usually within a percentage point or two comparing GSF to SSB within each x-axis category; however, the "positive SIPP and non-positive DER" proportion is often much higher in the SSB than in the DER. Sometimes this SIPP proportion is double or nearly triple that seen in the GSF. There are a few potential explanations for this. One is that the synthetic modeling simply falters in replicating this statistical relationship. Another is that our sampling restrictions in converting monthly SIPP earnings into annual SIPP earnings could be contributing to this. To have annual SIPP earnings less or equal to zero for the purposes of our analysis, an individual would need to have twelve non-missing months of SIPP earnings that sum to no greater than zero.

[7] The statistics in this paragraph were cleared for release: CBDRB-FY21-285.

To help assess what may be driving this finding, we generate some additional statistics concerning positive vs. non-positive earnings. For our research sample where both SIPP and DER earnings are non-missing, the GSF has positive DER earnings in 68.55 percent of such cases while the SIPP annual earnings are positive in 68.05 percent of cases. In the SSB, DER earnings are positive in 70.25 percent of cases, and SIPP annual earnings are positive in 75.34 percent of cases. So, the SSB has a higher rate of positive earnings in our research sample with a larger jump in frequency of positive SIPP annual earnings summed from the monthly level (68.05 percent to 75.34 percent) than in the frequency of positive DER earnings (68.55 percent to 70.25 percent). Looking at monthly SIPP earnings, the absolute difference between the proportion of positive earnings in the GSF and that in the SSB ranges from 0.05 percentage points to 2.45 percentage points. The median difference is around 1 percentage point. Our inference from this preliminary analysis is that our annual SIPP calculation based on summing monthly earnings is driving the starkly higher proportion of "positive SIPP and non-positive DER" cases in the SSB relative to the GSF.[8]

The final check we perform in this vein is to assess non-positive DER observations over time. Figure 11 shows the comparison of the proportion of non-positive DER earnings in our GSF sample to that in our SSB sample for the calendar years appearing in our analysis. There are several takeaways we would like to emphasize. First, the absolute difference between the proportion of non-positive DER earnings in the GSF and the proportion of non-positive DER earnings in the SSB is never more than 2 percentage points in the years analyzed. Second, the nonmonotonic time trend in proportion of DER earnings seen in the GSF is largely mirrored by the time trend in the SSB. There is a relatively steep downward trend in the early years before a leveling off and then slight increase in the last few years of the sample. This could relate to broader trends such as increases in female labor force participation in the 1980s and 1990s as well as the Great Recession of the late aughts. In terms of the focus of our present paper, an interesting finding of the analysis presented in Figure 11 is that the GSF initially shows a higher percentage of non-positive DER with the gap decreasing over time. Then in the final years analyzed, the SSB percentage is slightly higher.

Overall, our interpretation of these analyses is that the SSB is properly modeling many (though not all) of the patterns seen when categorizing earnings values as missing vs. non-missing

[8] The statistics in this paragraph were cleared for release: CBDRB-FY21-285.

and positive vs. non-positive. We do see closer replication of the GSF patterns for static individual estimates (e.g., overall frequencies of missing values, proportions of non-positive DER earnings by year) than for estimates involving within-person dynamics (e.g., frequencies of individuals with full calendar years of non-positive monthly earnings). We feel more research in this area is warranted but dedicate the remainder of this paper to comparing GSF and SSB results from regression analysis of social science questions.

### 3.2 Modeled Output

#### *3.2.1 Predictors of Missingness*

In addition to the descriptive analysis of missing data patterns, we perform regression analysis to assess what factors are related to the likelihood of missing earnings and if such relationships are estimated in both the SSB and the GSF. Table 2 shows predictors of having missing SIPP or DER earnings. We regress binary indicators for missing SIPP or DER earnings on demographic characteristics and Census regions. Columns (1)-(2) show the results based on the GSF data. Columns (3)-(4) show results based on the SSB data. The GSF results show many variables that have statistically significant correlations with missing earnings data. For example, non-White individuals are more likely to having missing SIPP and DER earnings than White individuals, while individuals who are married or have children are less likely to have missing SIPP and DER earnings than individuals who are single or childless.

Comparing column (1) to (3) and column (2) to (4), we see that most of the statistically significant predictors of missing earnings are replicated in the SSB. Column (3) has the same sign and significance as column (1) for 14 out of the 16 predictors. Column (4) has the same sign and significance as column (2) for 11 out of the 16 predictors. Of the seven total predictors that did not match sign and significance, four lost significance in the SSB while three others flipped signs (and were statistically significant). Overall, the SSB replicates many of the demographic correlates with missing earnings present in the GSF data.

#### *3.2.2 Mincer Regressions*

As one regression-based test, we perform analysis in the style of the seminal Mincer (1974) model and many subsequent labor economics analyses. Table 3 shows the results of these Mincer-style regressions to estimate age (as a proxy for experience) and education premia. We regress the

natural logarithm of real earnings on age, age-squared, categorical indicators for educational attainment, and control variables for demographic characteristics. The analytical sample used here is our main positive earners sample limited to individuals who are at least 25 years old but under age 65 and additionally have non-missing values for all covariates.

In Table 3, the first two columns show the results when using the GSF and the right-most columns show the results when using the SSB. The odd-numbered columns use SIPP earnings; the even-numbered columns use DER earnings. The estimates from the SSB all have the same sign and statistical significance as their counterparts from the GSF. The magnitudes are similar in size to varying degrees. For example, from the GSF, the earnings premium over the "less than high school" baseline is roughly 38 percent for high school degree, 67 percent for some college, 141 percent for college degree, and 328 percent for advanced degree when using DER earnings as the outcome. The corresponding estimates from the SSB are 49 percent, 92 percent, 206 percent, and 453 percent.

So, the SSB shows larger estimates across education categories with the larger differences occurring as education increases. For example, the expected earnings premium from an advanced degree is roughly 38 percent higher in the SSB relative to the GSF benchmark. The ranges of estimated coefficient magnitudes are smaller for the age, sex, and race variables. For example, the age coefficient estimates range from roughly 0.074 to 0.103, and the age-squared coefficient estimates range from -0.0008 to -0.0011. The difference (or lack thereof) between estimates when using the DER vs. the SIPP for the earnings outcome is similar comparing SSB to GSF. In sum: while the estimate magnitudes differ to varying degrees, the overall statistical takeaways from these Mincer-style regression analyses are the same comparing SSB to GSF.

*3.2.3 Time Series Evidence: Wage Gaps Over Time and Lifecycle Earnings*

Figures 12-15 show time series results for several constructed variables or wage gaps. The results are constructed by calculating the given statistic of interest separately in each SIPP panel, then plotting the results across SIPP panels. Each figure has two graphs: one for the GSF and one for the SSB. Each graph has two time series plots: one for SIPP-based earnings and one for DER-based earnings. All four figures are based on the positive earners sample described previously.

First, we study the college wage premium. The college wage premium is the average wage gap between individuals with versus without a college degree. Variation in the college wage gap

over time is informative about changes in the relative supply of versus demand for college-educated workers and thereby provides evidence on macroeconomic forces in the labor market. The college wage premium has been rising since the 1980s, although it has been rising at a decreasing rate since around the mid-1990s (e.g., Ashworth and Ransom, 2019; Card and Lemieux, 2001).[9]

Figure 12 shows the college wage premium over time. Our premium estimates are regression-adjusted with controls for highest education level, sex, race, a quartic in age, and Hispanic status.[10] We limit the sample to ages 25-54 (i.e., prime working age). We convert the SIPP and DER earnings into a wage by dividing by the individual's self-reported hours of work in the SIPP. The figure plots the coefficient estimate for a binary variable indicating whether the individual has at least a bachelor's degree.

The GSF plot matches the expected pattern of a rising college wage premium that is flattening over time. The time series is very similar for the SIPP and DER wages. The SSB plot shows similar patterns. The wage premiums are similar in magnitude to the GSF and are rising over time. The SIPP-based premiums also show the flatting of the wage premium growth over time. The SSB plot shows a larger difference between the SIPP and DER premiums than the GSF does. This is mostly due to the early 1990s and 2008 SIPP panels in which the DER wage premiums are noticeably larger than the SIPP wage premiums.

Next, we study age-earnings lifecycle profiles. The age-earnings lifecycle profile shows average earnings by age. The profile illustrates not only amounts of earnings, but lifecycle dynamics related to when earnings growth is largest and how earnings evolve as individuals approach retirement. Age-earnings profiles generally show a hump shape where earnings growth is largest during ages 25-35, earnings peak in the mid-to-late 40s, and then earnings decline beginning in the 50s as individuals approach retirement and reduce their attachment to the labor market (Murphy and Welch, 1990).

[9] The term "skill-biased technological change" is a commonly accepted explanation for a large portion of the rise in the college wage premium in recent decades (e.g., Card and DiNardo, 2002). This term refers to the rise in the use of technology that generally compliments college or even more advanced degrees as well as the replacement by technology of many manual labor jobs that were often held by individuals with fewer years of education.

[10] Our college wage premium estimates are not intended to be estimates of the causal effect of college on earnings. Individuals with versus without a college degree differ on many characteristics besides just their education level and basic demographics. Rather, college wage premium estimates are only intended to provide a description of how wages differ for those with versus without a college degree and how that has changed over time.

Figure 13 shows the age-earnings lifecycle profiles. The figures plot the coefficient estimates from a regression of earnings on age indicators without any covariates. All SIPP panels are pooled together. The profile for the GSF plot shows the expected hump shape with larger earnings growth during ages 25-35, flattened growth and peaked earnings during ages 40-50, and declining earnings beginning in the early 50s. The SIPP and DER earnings profiles are very similar. The SSB plots show the expected hump shape, although earnings growth is flatter over ages 35-50 in the SSB than the GSF and the earnings level itself is lower in the SSB across the full age range. The SIPP and DER earnings profiles track each other closely in the SSB during ages 25-35, but then begin diverging with the DER-based profile being noticeably larger for older ages. Thus, the GSF and SSB figures are visually similar and provide the same general conclusions about lifecycle earnings dynamics, although there are some noticeable differences in the figures.

Finally, we study the gender wage gap and the Black-White wage gap. These gaps show how wages differ on average between males and females or Black and White individuals. Understanding how average wages differ by gender and race, and how those differences have changed over time, provides importance information related to policy, inequality, and discrimination. The gender wage gap shrunk rapidly between 1980 and 2000 but has been relatively stable since then (Beaudry and Lewis, 2012; Blau and Kahn, 1997; Mulligan and Rubinstein, 2008). The Black-White wage gap has been widening since 1980 (Daly, Hobijn, and Pedtke, 2017).

Our wage gap estimates adjust for basic demographic characteristics including highest education level, gender (for the Black-White wage gap), race (for the gender wage gap), a quadratic in age, Hispanic status, and state of residence.[11] We convert the SIPP and DER earnings into a wage by dividing by the individual's self-reported hours of work in the SIPP. We also limit the sample to ages 25-54.

Figure 14 shows the gender wage gap results. The GSF results show the expected pattern, with the gender wag gap shrinking during the 1980s and 1990s before stabilizing since 2000. The SIPP and DER gaps are nearly identical. The SSB results show similar magnitudes and an overall declining wage gap, although the SSB results do not show as clear of a delineation between a

[11] Our estimates are not intended to represent causal effects of gender or race on earnings, nor are they intended to measure discrimination. Rather, they just report a conditional wage gap for one particular set of covariates that is commonly used to adjust the wage gap for basic demographic information.

shrinking wage gap from 1980-2000 and a stable one since 2000. Figure 15 shows the Black-White wage gap results. The GSF plots show the expected pattern of an overall widening wage gap, although there are some periods of shrinking gaps in our results. The SSB plots are flatter but show similar magnitudes and provide some visual evidence of a widening gap.

Overall, we find these results very encouraging. The SSB is able to replicate many important socioeconomic patterns related to wage gaps and lifecycle earnings dynamics, including how those statistics evolve over many decades.

### *3.2.4 Returns to Schooling*

The prior sections presented results on the average earnings differences across education levels and the college wage premium over time. In this section, we study a similar but different topic: the causal effect of schooling on earnings. This topic has long been of interest to social scientists as it is relevant for individual-level schooling decisions and for policymakers when determining education policy.

Early studies on the returns to schooling were based on ordinary least squares (OLS) estimates of regressions that attempt to explain wages or earnings as a function of schooling and experience (Mincer, 1974).[12] Decades of work since then has focused on omitted ability bias in the "Mincer equation." OLS estimates of the Mincer equation are assumed to overstate the returns to schooling due to a positive association between earnings and ability as well as ability and schooling (Griliches, 1977; Heckman, Lochner, and Todd, 2006). As a result, a large body of work has emerged using a variety of econometric techniques in attempt to provide a more reliable estimate of the return to schooling. One well-known approach was the use of quarter of birth as an instrumental variable (IV) for years of schooling using two-stage least squares (2SLS) instead of OLS (Angrist and Krueger, 1991). Identification for this IV approach stems from the idea that quarter of birth is related to earnings only through completed years of schooling. Historically, individuals born earlier in the calendar year start school at an older age and thus also reach the legal school dropout age after having attended school for a shorter period of time. This first-stage

[12] Variables commonly used to account for experience are age and "potential experience." Potential experience is typically measured as *age – years of school – 6* and is intended to proxy for actual years of work experience which is usually not observed. Age or potential experience are typically included in "Mincer regressions" in either a quadratic or quartic functional form (Murphy and Welch, 1990).

relationship between years of schooling and quarter of birth provides plausibly exogenous variation in years of schooling for the second-stage regression between earnings and schooling.

A common result in the returns to schooling literature is that IV estimates of the returns to schooling are larger than OLS estimates, despite the aforementioned assumption that omitted ability causes upward bias in the OLS estimate (Card, 2003). We test this result in the GSF using the "positive earners" sample which we further limit to non-Hispanic White males ages 25-54 with at least 30 weeks worked in calendar year and individuals without missing covariates. The results are shown in Table 4. The GSF replicates this result for both SIPP earnings (Panel A) and DER earnings (Panel B). The OLS estimate of the Mincer equation shows an 11-13 percent return to an additional year of schooling, depending on whether we use SIPP or DER earnings as the outcome variable. The 2SLS estimates show a 22 percent return to an additional year of schooling. The SSB results do not replicate this known pattern of the IV estimate for the return to schooling being larger than the OLS estimate. The OLS estimate is similar between the SSB and GSF (11-13 percent return in the GSF versus 13-16 percent in the SSB, both statistically significant), but the IV estimates in the SSB are smaller than OLS (11-13 percent return for 2SLS versus 13-16 percent for OLS) and not statistically significant.

Panel C of Table 4 illustrates why the SSB fails to replicate this pattern. The table shows the first-stage regression results from the 2SLS IV model in the GSF vs SSB.[13] This is a regression of years of schooling on indicators for quarter of birth (with quarter one as the excluded category) plus all the same covariates from the second stage (age, age squared, state fixed effects, and year fixed effects). In the GSF, later quarters of birth and age are all positively and significantly related to more completed years of schooling. The SSB replicates the relationship between years of schooling and age but fails to replicate the relationship with quarter of birth. The fact that the SSB fails to replicate the relationship that underpins the quarter of birth IV method likely explains why the SSB fails to reproduce the second stage result.

Next, we replicate the main regressions in Kejriwal et al. (2020). The authors used the GSF to estimate the returns to schooling and the interactive fixed effects estimators from Bai (2009) and Pesaran (2006) to account for omitted ability bias. The Panel A portion of Tables 5-7 reproduce several of the main results in their paper, while the Panel B portion attempts to replicate their

[13] Note that the first-stage regression model is the same between the DER earnings and SIPP earnings analyses because only the outcome variable in the second stage changes.

results when we run their code on the SSB.[14] The analysis in Table 5 is similar to the analysis described above from Table 4: the authors estimated Mincer equations using OLS and IV (based on quarter of birth) using both cross-section and panel data samples. Table 6 shows results for pooled panel data models that include interactive fixed effects (i.e., person fixed effects, time period fixed effects, and the interaction of person and time fixed effects). Table 7 shows results for heterogeneous coefficient panel data models that include interactive fixed effects.

The GSF results in Panel A of Table 5 show positive and statistically significant effects of schooling on earnings with the expected pattern that the IV estimate is larger than the OLS estimate. The SSB results in Panel B again fail to replicate this pattern. The OLS estimates are similar between the GSF and SSB – they are all positive and statistically significant – but the SSB 2SLS estimates are much smaller than the GSF 2SLS estimates and not statistically significant.

The SSB results in Tables 6 and 7, which are based on models with interactive fixed effects and/or heterogeneous coefficients, almost all fail to replicate their GSF counterpart. The various GSF results across the two tables all show a positive and statistically significant return to an additional year of schooling in the range of 2-8 percent. The SSB results are always small and close to zero. Only one of the ten total coefficient estimates remains positive and statistically significant in the SSB, while another coefficient estimate becomes negative and statistically significant.

It is noticeable across Tables 5-7 that results which rely on within-person earnings variation as a key source of identifying variation are less replicable in the SSB than results that do not rely on such variation. For example, in Table 5 there are three OLS estimates of the returns to schooling, one of which uses person fixed effects in a panel regression. The two specifications that do not use person fixed effects (column 1 and column 4) are quite close between the GSF and SSB [9.2 percent GSF return versus 7.0 percent SSB return for column 1 (24 percent reduction); 10.5 percent GSF return versus 9.3 percent SSB return for column 4 (11 percent reduction)]. The specification that does rely on person fixed effects shows a much larger attenuation of the return to schooling [7.7 percent GSF return versus 4.2 percent SSB return in column 3 (45 percent reduction)]. All the results in Tables 6 and 7 rely on within-person earnings variation. The Table 6 results include both individual fixed effects and interactive individual and time fixed effects which capture time-varying returns to unobserved individual characteristics. The Table 7 results include interactive

14 Tables 5-7 in our paper replicate the main results in Tables 3-5 of Kejriwal et al. (2020), respectively.

fixed effects and also heterogeneous coefficients that are based on individual-specific time series regressions. Unlike most of the findings we have presented so far, essentially none of the results from these models are replicated in the SSB.

In summary, the replicability of GSF returns to schooling results in the SSB is fairly encouraging for simple OLS estimates of the Mincer equation. However, the SSB failed to replicate the common result that IV estimates are larger than OLS estimates across multiple samples and specifications. One possible explanation for this is prior work showing that quarter of birth is a weak instrument which can lead to inconsistent and biased estimates (Bound et al., 1993); these types of weak relationships may be exactly the types of relationships that synthetic data models have a difficult time replicating. The SSB also noticeably failed to replicate results based on models that rely on within-person variation in earnings.

### *3.2.5 Vietnam War Draft Lottery and Civilian Earnings*

In this section, we investigate another well-known birthdate-related instrumental variable. Angrist (1990) studied the effect of military service on civilian earnings. In order to avoid the potential omitted ability bias due to non-random selection into the military, Angrist (1990) used the Vietnam draft lottery results as an instrument for military veteran status. There were three rounds of military draft lotteries during the Vietnam War period: 1970, 1971, and 1972.[15] For each lottery, individuals who turned 20 years old in that year were assigned a "random sequence number" (RSN), 1-366, based on their birthdate.[16] Later, a number from 1-366 was chosen and all individuals whose RSN was below that "draft eligible ceiling" were selected as draft-eligible.[17] The ceiling was 195 in 1970, 125 in 1971, and 95 in 1972.

We cannot attempt to exactly replicate the IV models from Angrist (1990) because military veteran status is not available in the SSB. However, Angrist (1990) showed that the effect of the draft lottery on military service and, ultimately, earnings was so strong that the draft lottery results themselves were strongly associated with civilian earnings without even accounting for actual military service. This is the result that we test in the GSF and SSB. We use the SSA's Summary

[15] There were also lotteries in 1973-1975, but nobody was actually drafted after 1972.
[16] The first draft, in 1970, also included individuals who turned 21-26 in that year.
[17] Individuals who were selected as draft-eligible by the lottery still had to pass a screening process that included physical examination and a mental aptitude test, meaning that the final selection into the military was not random. However, the fact that the initial induction into the draft-eligible population was random means that the draft lottery results still provide a plausible instrumental variable.

Earnings Record (SER) to build a panel of individual-level annual earnings from 1960-1979 for white males born from 1944 through 1952. We then estimate a difference-in-differences regression using two-way fixed effects and OLS: we regress the log value of annual earnings (adjusted for inflation) on individual fixed effects, year fixed effects, and an interaction between a categorical variable indicating whether an individual's RSN was below the draft-eligible ceiling and a categorical variable indicating the years after an individual's draft lottery year (along with birth year fixed effects). The coefficient on the interaction term is the difference-in-differences estimate for the effect of random selection into the "draft-eligible" population on subsequent civilian earnings. It represents the difference in average annual earnings between draft-eligible and non-draft-eligible individuals in years after the draft lottery relative to their difference in average earnings in years before the draft lottery.

Table 8 reports the results. The GSF result in column (1) shows an 11.43% reduction in annual earnings after the draft lottery for drafted-eligible individuals that is statistically significant. The SSB result in column (2) replicates the GSF result almost exactly: it produces an 11.65% reduction, also statistically significant. We view this as a particularly encouraging result with regards to accuracy in the SSB. Selection into the draft-eligible pool for the Vietnam War only impacted fewer than ten birth cohorts and was randomly assigned based on an individual's exact date of birth, yet the SSB managed to reproduce the lower average earnings in post-lottery years for individuals with these sets of birth dates.

*3.2.6 Social Security Disability Insurance and Positive Earnings Over Time*

The economic effects of Social Security Disability Insurance (SSDI) have long been of interest to researchers and policy makers. The GSF presents an opportunity to analyze the long-term earnings effects of SSDI application and receipt. The methodology is based on Charles (2003) and its response paper, Mok et al. (2008). An event study framework is used for a linear probability model with person-level fixed effects. As in many of our other analyses, we use the GSF and SSB to construct panel data. For this analysis, we utilize the DER for earnings information and the Master Beneficiary Record (MBR) for SSDI details. The analytical sample consists of individuals aged 30 through 61. For SSDI applicants, we exclude observations more than 10 years after an individual's reported onset. The baseline time frame is six years or more before disability onset,

and then relative year dummy variable are created for years between five years prior to disability onset and ten years post disability onset.

The outcome of interest is an indicator for positive DER earnings in year t, and the independent variables of interest are SSDI status indicators interacted with year-relative-to-disability-onset dummy variables. The SSDI status categories are one-time applicants who never receive benefits, individuals who never receive benefits but applied multiple times, individuals who received benefits on the first application, and individuals who received benefits at some point but not on the first application. The baseline sample consists of non-disabled individuals, defined as individuals who indicate no work-limiting disability in both the SIPP and the administrative records (including SSDI as well as Supplemental Security Income). The expectation is that the likelihood of positive earnings declines after disability onset with differential effects based on SSDI status.

Figure 16 shows the statistically significant effects for each of the SSDI categories for the GSF and the SSB. For comparison, a simpler difference-in-differences analysis was also performed using the same analytical sample less the non-disabled baseline used in the event study analysis. The treatment variable is SSDI benefits receipt, and the post variable is based on the year being greater than the disability onset indicated on the first SSDI application. The coefficient estimate for the treatment-post interaction term is included in Table 9.

In both the event study and the difference-in-differences analysis, the signs and significances for most of the estimates are the same. In the post onset years, there is a negative and statistically significant effect of SSDI benefits receipt on the likelihood of positive earnings. The effects are much smaller in the SSB analysis than when using the GSF. In the event study, each SSDI category shows little or no effect on positive earnings likelihood in the pre-onset period followed a steep drop post-onset. The largest decline is for SSDI applicants who receive benefits on the first application, and in the later years there is some separation between the other categories (see solid lines in Figure 16). When using the SSB, the results are all statistically significant and negative, with a linear decline over time. We see slightly larger effects in the post period for the "SSDI benefits on first application" group with little difference between the estimated effects for the other three groups (see dotted lines in Figure 16). Similarly, the treatment-post interaction term coefficient estimate in the difference-in-differences model is over 10 times as large for the GSF analysis than in the SSB analysis (see Table 9). Further, the event study sample has a larger GSF

sample compared to the SSB sample while the difference-in-differences SSB sample (which limits to the SSDI applicant pool) is larger than the GSF sample. So, it seems that there are more SSDI applicants in the SSB and more non-disabled individuals (or individuals who could be identified as lacking work-limiting disabilities) in the GSF sample.

These findings make sense in the context of our other analysis. The SSB does capture the strong and expected negative relationship between SSDI and likelihood of positive earnings, but it doesn't find the same differential effects by category. The SSB can falter with replicating GSF empirical results when extensive margin considerations are critical. That is particularly relevant to this SSDI analysis where the outcome and independent variables of interest all rely on categorical identification. For something like positive earnings, the SSB could be highly accurate with the continuous measure of earnings but miss the mark on the zero vs. positive distinction (e.g., zero earnings in the GSF vs. $5 in the SSB would be quite close in the continuous sense but a complete miss for the binary variable). Similarly, the SSB could correctly identify someone as a SSDI beneficiary, but if the model flags the successful application as the second application instead of the first, that will change the treatment category in the event study framework.

### *3.2.7 Minimum Wages and Labor Market Outcomes*

We now present regression results for the effect of minimum wage increases on a variety of labor market outcomes. There is consensus in the minimum wage literature that minimum wage increases raise wages and average earnings for the lower part of the wage distribution, but there has been debate about whether, and to what extent, this comes at the cost of reduced employment for some groups (Allegretto, Dube and Reich, 2011; Allegretto et al., 2017; Dube, Lester and Reich, 2010; Neumark, Salas and Wascher, 2014b, a). Much of this debate has centered around the appropriate specification to use in panel regressions for estimating the relationship between employment and minimum wages.

We begin by estimating the effect of minimum wage increases on earnings and employment. The sample is the "full sample" described earlier, further limited to teens ages 16-19 without any missing covariates.[18] We use panel data samples and specifications that include two-way fixed effects for state and year as well as state-specific linear time trends and Census Division-

[18] Teenagers and restaurant workers are often the focus of minimum wage studies on employment (Allegretto, Dube and Reich, 2011; Allegretto et al., 2017; Dube, Lester and Reich, 2010; Neumark, Salas and Wascher, 2014b, a).

by-year fixed effects. Specifications with these controls have the most support from the literature as they can account for regional heterogeneity in employment trends that happens to be correlated with minimum wage levels (Cengiz et al., 2019; Totty, 2017). The results are in Table 10. Columns (1)-(2) show results from the GSF and columns (3)-(4) show results from the SSB. Panel A shows results based on SIPP earnings and Panel B show results based on DER earnings.[19]

The results in columns (1)-(2) are consistent with the minimum wage literature – minimum wage increases raise wages for teenage workers with little-to-no evidence of employment loss once regional heterogeneity is accounted for. Results from the SIPP and DER are very similar. The SSB does not replicate these results very well. The SSB shows no relationship between minimum wage increases and wages for teenage workers: the coefficient estimate for the log minimum wage variable is negative and not statistically significant for both SIPP and DER wages. The SSB results for employment are similar to the GSF in that neither shows a statistically significant relationship between minimum wages and employment, but the coefficient estimates still move closer zero in a way that suggests attenuation of what little correlation there was in the GSF.

Next, we attempt to replicate the main regressions in Hampton and Totty (2021). The authors used the GSF to study the effect of minimum wage increases on employment, permanent labor force exit, and Social Security retirement benefit claiming for low-wage workers during retirement ages (ages 62-70). The Panel A portion in Tables 11-13 reproduce their results from the GSF for employment, permanent exit, and benefit claiming, respectively.[20] The Panel B portion shows the results from the SSB.

The employment results in Table 11 are broken up into three different outcomes: any employment (indicating positive DER earnings in a given year) and full-time versus part-time employment (based on the amount of a person's earnings in a given year relative to their lifetime highest amount).[21] Each of the regressions are balanced person-year panel regressions over ages 62-70. We show three different specifications for each outcome.

[19] We convert the SIPP and DER earnings into a wage by dividing by self-reported hours worked in the SIPP. For employment, the outcome variable is equal to 1 if the individual had positive SIPP/DER earnings and 0 if their earnings were equal to $0.

[20] The employment results we are reproducing from Hampton and Totty (2021) come from Tables 2-3 in their paper. The permanent exit and claiming results come from Table 4 and Table 5 of their paper, respectively.

[21] Full-time employment in a given year is defined as an individual earning at least 50% of their lifetime highest annual earnings amount (in inflation-adjusted dollars) observed in the data. Part-time employment in a given year is defined as earning less than 50% of their highest observed earnings year but still at least $5,000 in inflation-adjusted dollars, which equates to working approximately 20 hours per week at the minimum wage for a full year or working 40 hours per week at the minimum wage for six months.

The GSF results show that minimum wage increases lead to more employment for older workers and that the increased employment is made up of increases in both full-time and part-time work. The SSB replicates these results for some specifications but not for others. The results are qualitatively similar in columns (1)-(2), (4)-(5), and (7)-(8): the SSB results are similar in magnitude to the GSF and generally appear suggestive of positive effects on employment, full-time employment, and part-time employment despite several of the estimates narrowly missing statistical significance at the ten-percent level. The SSB results in columns (3), (6), and (9) that include person fixed effects, on the other hand, are noticeably different in magnitude from the GSF. They are all much closer to zero, not close to statistical significance at conventional levels, and one coefficient even changes signs.

The permanent exit from employment results in Table 12 are broken up into two different outcomes: partial and full exit, based on the amount of a person's earnings in a given year relative to their lifetime maximum amount.[22] We estimate two different regression specifications using OLS for each outcome. The regressions are effectively person-year hazard models in which the person drops out of the sample after their earnings permanently fall below a given individual-specific threshold. Like the person fixed effects regressions for employment in Table 8, the permanent exit hazards rely on within-person earnings dynamics: permanent exit is measured as the point in time at which a person's earnings permanently fall below a person-specific threshold based on their earnings history. The GSF results show that minimum wage increases lead to delayed full permanent exit from employment. The SSB does not replicate this result, as minimum wage increases show no evidence of a relationship with either partial or full permanent exit from employment.

The retirement benefit claiming results are shown in Table 13. The outcome is an indicator for whether the individual first received retirement benefits in a given month. We estimate person-month hazard regressions in which the person drops out of the sample after the first month they received retirement benefits. The GSF results show that minimum wage increases lead to delayed claiming of retirement benefits. Once again, the SSB does not replicate this result, as there is no relationship between minimum wages and claiming conditional on the other covariates in the

[22] If an individual's annual earnings permanently fall to less than 50% of their lifetime inflation-adjusted maximum but still at least $5,000 in inflation-adjusted dollars, then that is classified as permanent partial employment exit. If an individual's annual earnings permanently fall to less than 50% of their lifetime inflation-adjusted maximum and less than $5,000 in inflation-adjusted dollars, then that is classified as permanent full employment exit.

specifications. The SSB does, however, replicate a positive and statistically significant relationship between the month during which an individual first reaches their Full Retirement Age (FRA) and claiming, although the magnitude is much smaller than in the GSF.[23]

In summary, while the SSB did replicate some evidence of a relationship between minimum wages and increased employment (column 1 of Table 11) as well as reaching FRA and claiming retirement benefits (column 2 of Table 13) for older workers, the majority of statistically significant relationships between minimum wages and labor market outcomes in the GSF became insignificant in the SSB. Many of the relationships that did not show up in the SSB relied on within-person earnings dynamics, either as a key source of identifying variation or in the measurement of the outcome variable.

### *3.2.8 Relative Income Within Households*

Bertrand et al. (2015) is a well-known study on the causes and consequences of relative income within households. One of their findings is that the distribution of the share of household income earned by the wife exhibits a sharp discontinuity at 0.5, indicating that individuals are less likely to match and form a couple if the female's income is less than the male's. The authors used the GSF to show that this pattern exists in administrative earnings data from the U.S. and is thus not just a result of measurement error in self-reported survey earnings data. However, the authors did not find the discontinuity at 0.5 when first using the SSB; only when they received the validated results did it show up.

Figure 17 shows four different versions of the Bertrand et al. (2015) finding. The top left reproduces their exact results using the administrative DER earnings in the GSF. The bottom left replicates the same figure based on the SSB. The two figures on the right replicate the result using SIPP earnings rather than DER earnings. The two top two figures, based on the GSF, both show the result that there is more density to the left of 0.5 and a stark discontinuity in the density at 0.5. The SSB replicates the more general result that there is greater density to the left of 0.5 than to the right, but it is unable to replicate the discontinuity that occurs at 0.5.

[23] One other result of note from Tables 11-13 is that the sample sizes are much larger in the SSB than the GSF. The sample in Hampton and Totty (2021) is individuals whose average wage in the SIPP was less than or equal to the minimum wage plus two dollars. The larger low-wage sample in the SSB than the GSF is consistent with Figure 1 which shows a larger density of individuals with low earnings in the SSB than in the GSF.

The Bertrand et al. (2015) use of the GSF is a great example of the benefits of synthetic data with validation. Because of confidentiality concerns, the Census Bureau determined that it was necessary to synthesize the data before dissemination. Without the SSB, the authors would have either not been able to use administrative data on earnings in the U.S. or would have had to access those sensitive data through a Federal Statistical Research Data Center (FSRDC) which can be both expensive and time consuming. At the same time, the validation step was also crucial. While we have shown that synthetic data can successfully replicate a lot of socioeconomic relationships, social dynamics such as gender norms and relative income within households that generate such stark discontinuities may be tough to replicate well unless they are explicitly modeled.

## 4. Discussion

Results from the assorted analyses performed in this paper provide several implications related to our goal of assessing the comparability between output derived from the SIPP Synthetic Beta and those generated from the same analyses using the confidential SIPP Gold Standard File. In order to give an overview of the entire analysis, Figure 18 and Table 14 summarize many of the descriptive and model-based results presented previously in the paper. Figure 18 shows a scatter plot of the GSF versus SSB results. The figure shows descriptive and model-based results separately. While we have seen that the SSB does not replicate all the results from the GSF, it is clear from the figure that there is a strong association between results derived from the GSF versus the SSB. Table 14 summarizes differences in the 95% confidence interval coverage and statistical conclusions in terms of sign and statistical significance for the model-based results. Confidence intervals in the SSB, adjusted for synthesis uncertainty via multiple implicate variance, are approximately twice as long as the GSF confidence intervals on average. The synthetic confidence interval overlaps with 33% of the original confidence interval on average and covers the original coefficient estimate 35% of the time. The SSB produces the same sign as the GSF 79% of the time, the same statistical conclusion 63% of the time, and the opposite statistical conclusion only 2% of the time. Thus, while there are often meaningful differences between the exact magnitude of GSF versus SSB results, the SSB does a better job maintaining the presence of statistically significant relationships.

Where we see differences between results, many of the findings are consistent with interpretable and expected patterns. Statistics that are sensitive to outliers (e.g., means in Figures 4 and 6) may be less likely to be replicated in synthetic data than statistics that are not sensitive to outliers (e.g., medians in Figures 5 and 7) because synthetic data inherently attempt to mask sensitive values such as outliers. Additionally, regressions that rely solely on variables already in the data (e.g., the Mincer-style regressions in Table 3) may yield more replicable results than regressions that merge external data onto the synthetic data (e.g., the effect of minimum wages on earnings for teenagers in Table 10). For the former, all variables used in the regression model are included in the underlying synthetic data models; thus, relationships among variables are more likely to be retained after synthesis.

Modeling decisions when creating the synthetic data can also explain some of the differences. For example, in Table 4 we saw that the SSB does a good job of replicating the OLS estimate of the return to schooling but not the IV estimate, which is due to the SSB failing to replicate the first-stage relationship between quarter of birth and years of schooling. This can be explained by the synthetic data model for the SSB education variable not including the administrative date of birth variable used in the returns to schooling application.[24]

The tendency of results that rely on within-person earnings dynamics to hold up less well (e.g., having a full calendar year of non-positive monthly earnings discussed in section 3.1.2; person fixed effects and hazard panel regressions in Tables 5, 11, and 12; interactive fixed effects regressions in Tables 6 and 7; and the SSDI event study analysis in Table 9) can also be tied back to a modeling decision. The synthetic data models for the SSB were primary based on modeling variable *levels*. However, a similar data product from the Census Bureau known as the Synthetic Longitudinal Business Database (SynLBD) chose to model within-establishment *changes over time* in key variables rather than model variable levels (Kinney et al., 2014). If the synthetic models for the SSB had been adjusted to explicitly model within-person changes in earnings over time, then the SSB may have performed better on these types of analyses in our paper.[25]

[24] The model did include the SIPP-reported date of birth, which would be correlated with administrative date of birth, but it included date of birth as a continuous variable rather than modeling calendar effects (such as quarter of birth).

[25] For example, the annual SIPP earnings used in our present paper relied on having non-missing monthly earnings values for all twelve months of the calendar year, and having a non-positive annual value (i.e., summing all twelve non-missing months) in the SSB was directly affected by how each monthly value was synthesized. One month's synthesized earnings value could change the annual value calculation at the extensive margin (e.g., making someone who had all zeros in the GSF have one positive earnings month in the SSB). Because modeling was done in levels, it is possible that the SSB correctly matches the overall frequency of missing or zero earnings but does not correctly

Finally, we want to stress a caveat that relates to how we discuss and think about accuracy in this paper. Throughout the paper we infer accuracy based on the tendency of results from the SSB to replicate results from the GSF. Implicit in this characterization is the assumption that results derived from the confidential data are the "truth" (or full/maximum accuracy). This assumption is correct if the confidential data have no error. But we know that both survey and administrative data already contain errors, including coverage error, item non-response error, and measurement error (Abowd and Stinson, 2013; Meyer et al., 2015; Meyer and Mittag, 2020). These errors can impact important statistics (Bee and Mitchell, 2017; Meyer and Mittag, 2019; Meyer et al., 2020). Because survey and administrative data already contain error, differences between the GSF and SSB do not necessarily correspond directly to accuracy loss (or improvements in privacy). We are fine with this flawed characterization of accuracy for now, because very little is known about what synthetic data are capable of in terms of replicating a wide range of socioeconomic relationships. Our findings could therefore be seen as under-estimating accuracy in the SSB in the sense that we implicitly attribute any and all differences between the GSF and SSB to deviations from the "truth" even though we know the GSF already suffers from inaccuracies due to other sources of error.

## 5. Conclusion

It is increasingly difficult for data providers to protect the privacy of survey respondents due to the growing availability of public datasets, computing resources, and advanced statistical methods that collectively lead to rising risks of reconstruction and reidentification attacks. Synthetic data provide external researchers a chance to conduct a wide variety of analyses on microdata while still satisfying the legal objective of protecting privacy of survey respondents. Synthetic data can be used in conjunction with a validation option so that researchers can receive results based on confidential data without ever accessing the confidential data themselves.

Validation is costly in terms of resources, time, and privacy leakage; these costs affect the data provider, data user, and the individuals who appear in the data. It is therefore important to understand how well the synthetic data replicate results estimated using the confidential data. Little is known about the strengths and weaknesses of synthetic data in terms of what types of analyses

match the tendency of zeros to persist over time for particular individuals who are unemployed or out of the labor force. This explanation is consistent with the missing data pattern, where annual SIPP earnings was missing if *any* monthly earnings value was missing. In that case, the SSB and GSF rates were nearly identical.

or statistical methods are most likely to produce similar results to those based on the confidential data. We begin to fill in this gap by studying how the results from socioeconomic empirical analyses differ between an internal, confidential product of the U.S. Census Bureau (the SIPP GSF) and its synthetic equivalent (the SSB).

We find that the SSB does a good job replicating many results estimated using the GSF – including descriptive statistics, time trends in national statistics, and coefficient estimates from regression analyses. The SSB performs best in terms of replicating results from the GSF when our analysis involves only variables modeled from the GSF and when using methods that are less sensitive to outliers. The relative performance of the SSB noticeably declines when our analysis relies on merged external data or within-person variation in earnings. Overall, there is a strong association between the GSF and SSB results. Even when the SSB magnitudes and confidence intervals differ meaningfully from the GSF, the SSB still often delivers the same statistical conclusion in terms of sign and significance.

Two big picture considerations for synthetic data are that there is no universal standard for the concept of usefulness, and a synthetic model or process cannot cover or address every potential use case. With the caveat that the accuracy findings of the present paper are limited to the similarity between the SSB and the GSF, we can still make some general takeaways from our SSB experience that also relate to the aforementioned broad considerations. First, modeling decisions inherently prioritize particular use cases, and feedback from data users is mutually beneficial. What is "useful" or "accurate" to one use case or research team will not be considered so for a different use case or research team. Having a feedback loop can help determine which cases work best with a given synthetic model or process and possibly which use cases are worth prioritizing. Second, validation and/or verification are important complements to synthetic data. For a given level of privacy protection, synthetic data can only be so accurate, and it will be difficult if not impossible to provide accurate results for all use cases. Finally, the science for generating and evaluating synthetic data has advanced in the years since the SSB was first developed and is still evolving. Synthetic data are only as good as the models used to create it. Lots of decisions go into producing synthetic data. While regression-based synthetic models using sequential regression multiple imputation (SRMI) were at the frontier of the science for creating synthetic data when the SSB was first created in 2003, newer synthetic data methods such as non-parametric classification and regression trees (CART) and machine learning are easier to implement and can generate more

accurate synthetic data (Drechsler and Reiter, 2011; Reiter, 2005; Reiter and Kinney, 2012). Given the improvements in synthetic data modeling since the creation of the SSB and our strict approach to characterizing accuracy, we essentially see our findings as a lower bound for the accuracy level that future synthetic data releases can attain.

In future work we aim to build on this paper by continuing to study the usability of synthetic data for different applications and estimation methods common in empirical research. We plan to perform similar analyses on synthetic data generated using classification and regression trees. We also plan to study the amount and effect of synthesis error relative to other sources of error in survey and administrative data.

## References


Abowd, J. A., & Stinson, M. H. (2013). Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data. *The Review of Economics and Statistics, 95*(5), 1451–1467. doi:10.1162/REST_a_00352

Abowd, J. M., Benedetto, G., Garfinkel, S. L., Dahl, S. A., Dajani, A. N., Graham, M., . . . Villhuber, L. (2020). *The modernization of statistical disclosure limitation at the U.S. Census Bureau.* U.S. Census Bureau Working Paper. Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/2020/adrm/The%20modernization%20of%20statistical%20disclosure%20limitation%20at%20the%20U.S.%20Census%20Bureau.pdf

Abowd, J. M., Schmutte, I. M., Sexton, W. N., & Vilhuber, L. (2019). Why the Economics Profession Must Actively Participate in the Privacy Protection Debate. *AEA Papers and Proceedings, 109*, 397-402. doi:10.1257/pandp.20191106

Allegretto, S. A., Dube, A., & Reich, M. (2011). Do Minimum Wages Really Reduce Teen Employment? Accounting for Heterogeneity and Selectivity in State Panel Data. *Industrial Relations: A Journal of Economy and Society, 50*(2), 205-240. doi:10.1111/j.1468-232X.2011.00634.x

Allegretto, S., Dube, A., Reich, M., & Zipperer, B. (2017). Credible Research Designs for Minimum Wage Studies: A Response to Neumark, Salas, and Wascher. *ILR Review, 70*(3), 559-592. doi:10.1177/0019793917692788

Angrist, J. D. (1990, June). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security. *American Economic Review, 80*(3), 313-336. Retrieved from https://www.jstor.org/stable/2006669

Angrist, J. D., & Kreuger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics, 106*(4), 979-1014. doi:10.2307/2937954

Ashworth, J., & Ransom, T. (2019). Has the college wage premium continued to rise? Evidence from multiple U.S. surveys. *Economics of Education Review, 69*, 149-154. doi:10.1016/j.econedurev.2019.02.003

Bai, J. (2009). Panel Data Models With Interactive Fixed Effects. *Econometrica, 77*(4), 1229-1279. doi:10.3982/ECTA6135

Beaudry, P., & Lewis, E. (2014). Do Male-Female Wage Differentials Reflect Differences in the Return to Skill? Cross-City Evidence from 1980-2000. *American Economic Journal: Applied Economics, 6*(2), 178-194. doi:10.1257/app.6.2.178

Bee, A., & Mitchell, J. (2017). Do Older Americans Have More Income Than We Think? *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association, 110*, 1-85. Retrieved from https://www.jstor.org/stable/26794437

Benedetto, G., Gathright, G., & Stinson, M. (2010). *The Earnings Impact of Graduating from College during a Recession.* U.S. Census Bureau Working Paper. Retrieved from

https://www2.vrdc.cornell.edu/news/wp-content/papercite-data/pdf/benedettogathrightstinson-11301.pdf

Benedetto, G., Stanley, J. C., & Totty, E. (2018). *The creation and use of the sipp synthetic beta v7.0.* CES Technical Notes Series 18-03, U.S. Census Bureau, Center for Economic Studies. Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Creation_SSBv7.pdf

Benedetto, G., Stinson, M. H., & Abowd, J. M. (2013). *The Creation and Use of the SIPP Synthetic Beta.* U.S. Census Bureau Working Paper. Retrieved from https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf

Bertrand, M., Kamenica, E., & Pan, J. (2015). Gender Identity and Relative Income within Households. *The Quarterly Journal of Economics, 130*(2), 571–614. doi:10.1093/qje/qjv001

Blau, F. D., & Kahn, L. M. (1997). Swimming Upstream: Trends in the Gender Wage Differential in the 1980s. *Journal of Labor Economics, 15*(1), 1-42. doi:10.1086/209845

Bound, J., Jaeger, D. A., & Baker, R. M. (1993). Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association, 90*(430), 443-450. doi:10.1080/01621459.1995.10476536

Bowen, C. M., Bryant, V., Burman, L., Khitatrakun, S., McLelland, R., Stallworth, P., . . . Williams, A. R. (2020). A Synthetic Supplemental Public Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications. *International Conference on Privacy in Statistical Databases, 12276*, 257-270. doi:10.1007/978-3-030-57521-2_18

Card, D. (2003). Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica, 69*(5), 1127-1160. doi:10.1111/1468-0262.00237

Card, D., & DiNardo, J. E. (2002). Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles. *Journal of Labor Economics, 20*(4), 733-783. doi:10.1086/342055

Card, D., & Lemieux, T. (2001). Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-Based Analysis. *The Quarterly Journal of Economics, 116*(2), 705–746. doi:10.1162/00335530151144140

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). The Effect of Minimum Wages on Low-Wage Jobs. *The Quarterly Journal of Economics, 134*(3), 1405–1454. doi:10.1093/qje/qjz014

Charles, K. K. (2003). The Longitudinal Structure of Earnings Losses among Work-Limited Disabled Workers. *Journal of Human Resources, 383*(3), 618-646. doi:10.3368/jhr.43.3.721

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2008). Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation. *Journal of Official Statistics, 24*(2), 255-275. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3096944/

Daly, M. C., Hobijn, B., & Pedtke, J. H. (2017). *Disappointing Facts about the Black-White Wage Gap.* Federal Reserve Bank of San Francisco. Retrieved from https://www.frbsf.org/economic-research/publications/economic-letter/2017/september/disappointing-facts-about-black-white-wage-gap/

Drechsler, J., & Haensch, A.-C. (2023). *30 years of synthetic data.* arXiv. Retrieved from https://arxiv.org/abs/2304.02107v1

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis, 55*, 3232-3243. doi:10.1016/j.csda.2011.06.006

Dube, A., Lester, T. W., & Reich, M. (2010). Minimum Wage Effects Across State Borders: Estimates Using Contiguous Counties. *The Review of Economics and Statistics, 92*(4), 945–964. doi:10.1162/REST_a_00039

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. *Proceedings of the Third Conference on Theory of Cryptography* (pp. 265-284). Heidelberg: Springer.

Federal Committee on Statistical Methodology. (2005). *Report on Statistical Disclosure Limitation Methodology.* Washington, DC: US Office of Management and Budget. Retrieved from https://nces.ed.gov/FCSM/pdf/SPWP22_rev.pdf

Griliches, Z. (1977). Estimating the Returns to Schooling: Some Econometric Problems. *Econometrica, 45*(1), 1-22. doi:10.2307/1913285

Gruzd, A., & Hernandez-Garcia, A. (2018). Privacy Concerns and Self-Disclosure in Private and Public Uses of Social Media. *Cyberpsychology, Behavior, and Social Networking, 21*(7), 418-428. doi:10.1089/cyber.2017.0709

Hampton, M., & Totty, E. (2021). *Minimum Wages, Retirement Timing, and Labor Supply.* Working Paper. doi:10.13140/RG.2.2.32952.67840

Heckman, J. J., Lochner, L. J., & Todd, P. E. (2006). Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond. In E. Hanushek, & F. Welch (Eds.), *Hankbook of Economics of Education* (Vol. 1, pp. 307-458). New York: Elsview. doi:10.1016/S1574-0692(06)01007-5

Henriques, A. (2018). How does social security claiming respond to incentives? Considering husbands' and wives' benefits separately. *Journal of Human Resources, 53*(2), 382-413. doi:10.3368/jhr.53.2.1212-5371R2

Juhn, C., & McCue, K. (2016). Evolution of the marriage earnings gap for women. *American Economic Review: Papers and Proceedings, 106*(5), 252-256. doi:10.1257/aer.p20161120

Kejriwal, M., Li, X., & Totty, E. (2020). Multidimensional skills and the returns to schooling: Evidence from an interactive fixed-effects approach and a linked survey-administrative data set. *Journal of Applied Econometrics, 35*(5), 548– 566. doi:10.1002/jae.2759

Kinney, S. K., Reiter, J. P., & Miranda, J. (2014). SynLBD 2.0: Improving the synthetic Longitudinal Business Database. *tatistical Journal of the IAOS, 30*(2), 129-135. Retrieved from https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji00808

Kinney, S., Reiter, J. P., Reznek, A. P., Mirando, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Science Review, 79*(3), 362-384. doi:10.1111/j.1751-5823.2011.00153.x

Little, R. J. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics, 9*(2), 407-426. Retrieved from https://www.proquest.com/scholarly-journals/statistical-analysis-masked-data/docview/1266808565/se-2?accountid=36218

Meyer, B. D., & Mittag, N. (2019). Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net. *American Economic Journal: Applied Economics, 11*(2), 176-204. doi:10.1257/app.20170478

Meyer, B. D., & Mittag, N. (2020). An empirical total survey error decomposition using data combination. *Journal of Econometrics*, forthcoming. doi:10.1016/j.jeconom.2020.03.026

Meyer, B. D., Mittag, N., & George, R. M. (2020). Errors in Survey Reporting and Imputation and their Effects on Estimates of Food Stamp Program Participation. *Journal of Human Resources*, forthcoming. doi:10.3368/jhr.58.1.0818-9704R2

Meyer, B. D., Wallace, K. C., & Sullivan, J. X. (2015). Household Surveys in Crisis. *Journal of Economic Perspectives, 29*(4), 199-226. doi:10.1257/jep.29.4.199

Mincer, J. (1974). *Schooling, Experience, and Earnings* (Vol. No. 2). Human Behavior & Social Institutions. Retrieved from https://eric.ed.gov/?id=ED103621

Mok, W. K., Meyer, B. D., CHarles, K. K., & Achen, S. (2008). A Note on The Longitudinal Structure of Earnings Losses among Work-Limited Disabled Workers. *Journal of Human Resources, 43*(3), 721-728. doi:10.3368/jhr.43.3.721

Mulligan, C. B., & Rubinstein, Y. (2008). Selection, Investment, and Women's Relative Wages over Time. *The Quarterly Journal of Economics, 123*(3), 1061-1110. doi:10.1162/qjec.2008.123.3.1061

Murphy, K. M., & Welch, F. (1990). Empirical Age-Earnings Profiles. *Journal of Labor Economics, 8*(2), 202-229. doi:https://doi.org/10.1086/298220

Neumark, D., Salas, J. M., & Wascher, W. (2014). More on recent evidence on the effects of minimum wages in the United States. *IZA Journal of Labor Policy, 3*(24). doi:10.1186/2193-9004-3-24

Neumark, D., Salas, J. M., & Wascher, W. (2014). Revisiting the Minimum Wage—Employment Debate: Throwing Out the Baby with the Bathwater? *ILR Review, 67*(3), 608-648. doi:10.1177/00197939140670S307

Neumeier, C., Sorensen, T., & Webber, D. (2018). The Implicit Costs of Motherhood over the Lifecycle: Cross-Cohort Evidence from Administrative Longitudinal Data. *Southern Economic Journal, 84*(3), 716-733. doi:10.1002/soej.12239

Pesaran, M. H. (2006). Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica, 74*(4), 967-1012. doi:10.1111/j.1468-0262.2006.00692.x

Reiter, J. P. (2011). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics, 21*, 441-462.

Reiter, J. P., & Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics, 28*, 583-590.

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics, 9*(2), 461-468. Retrieved from https://www.proquest.com/scholarly-journals/discussion-statistical-disclosure-limitation/docview/1266818482/se-2?accountid=36218

Totty, E. (2017). THE EFFECT OF MINIMUM WAGES ON EMPLOYMENT: A FACTOR MODEL APPROACH. *Economic Inquiry, 55*(4), 1712-1737. doi:10.1111/ecin.12472

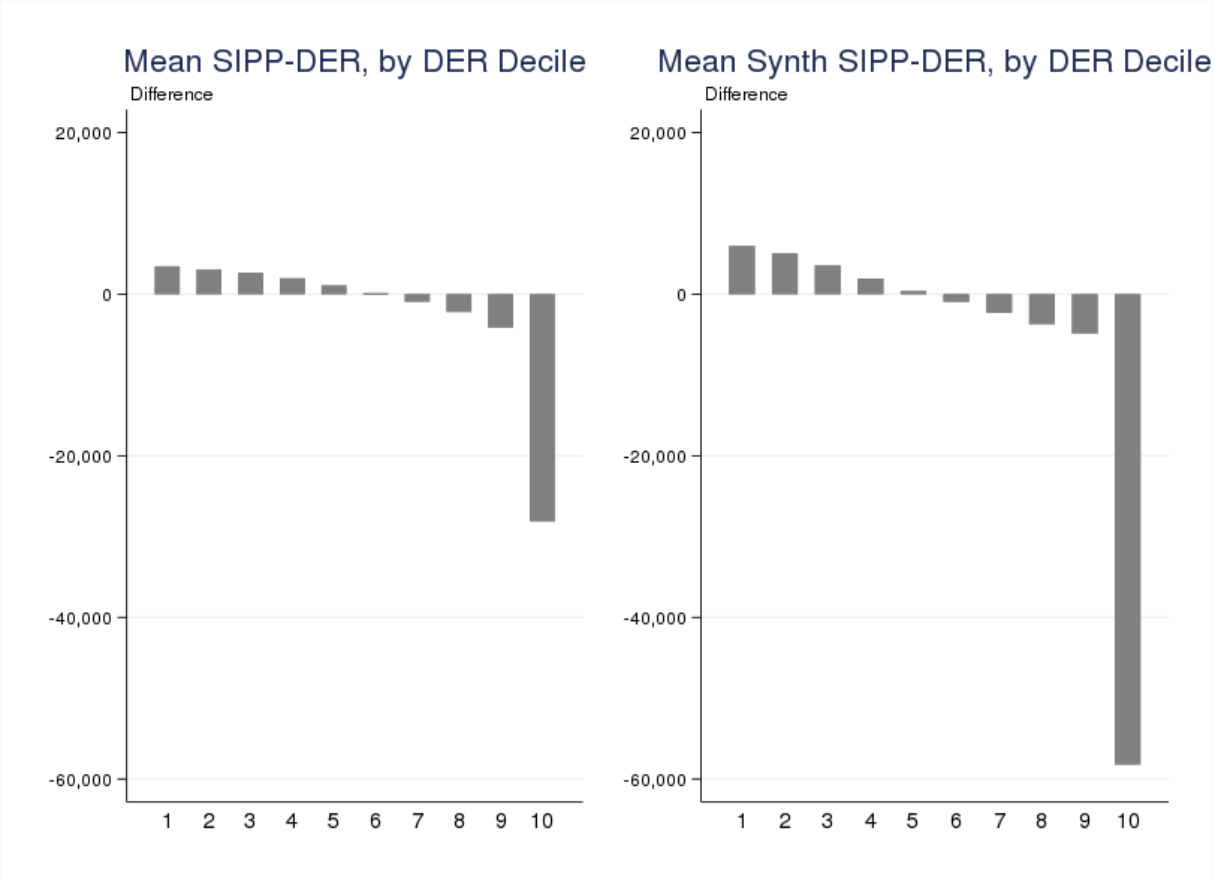
## Figures and Tables

### Figure 1: Self-Reported and Administrative Earnings Distributions


Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure shows a kernel density estimation (KDE) plot of the density for SIPP and DER earnings in the original (GSF) and synthetic (SSB) data. The sample consists of all person-year observations during which both DER and SIPP earnings are observed and are positive. Monthly SIPP earnings are summed the annual level and only counted as non-missing if all 12 months were non-missing. The top and bottom five percent of earnings observations were trimmed for each earnings measure (DER or SIPP) and data source (GSF or SSB). Additional details in Section 3.1.1.

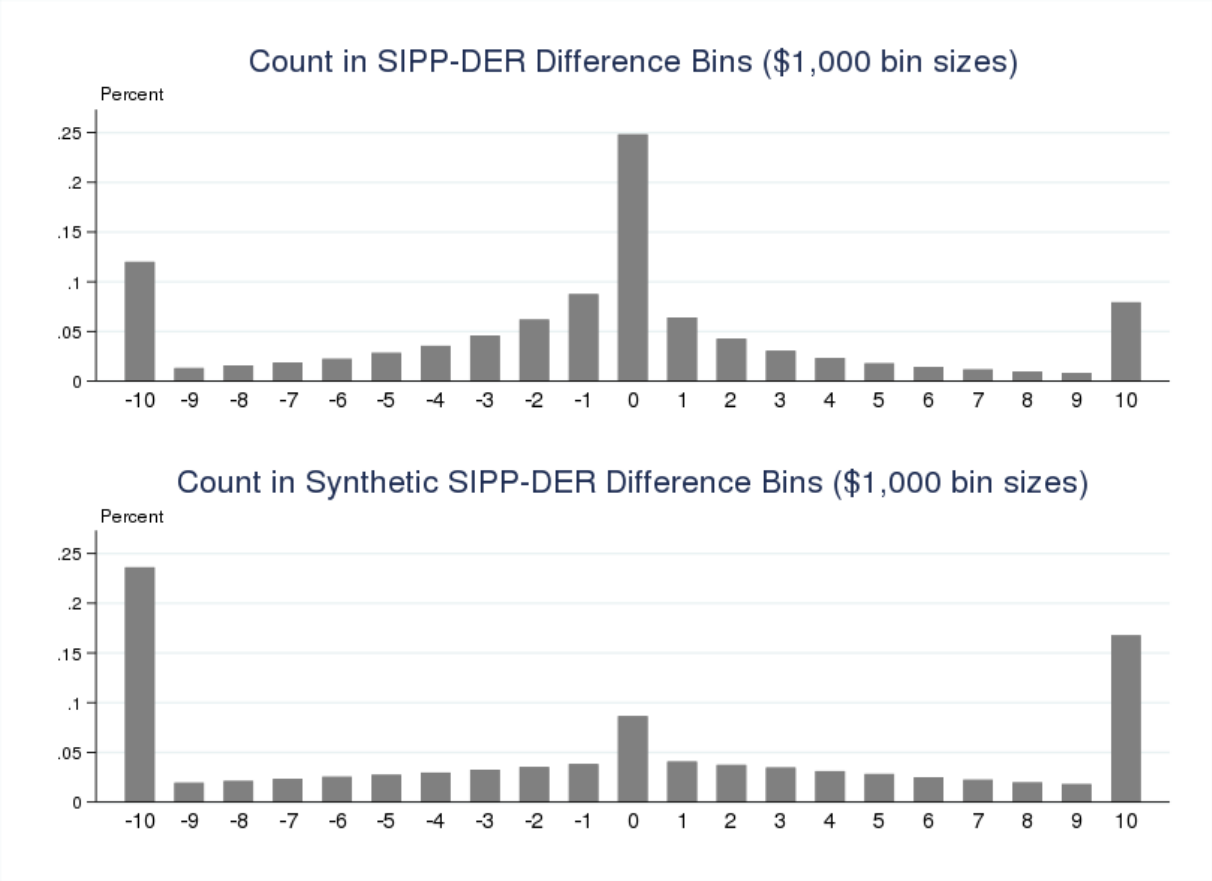
**Figure 2: Earnings Differences Across Data Sets and Sources**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure shows the average person-level difference between SIPP and DER earnings (SIPP minus DER) in each decile of DER earnings for the original (GSF) and synthetic (SSB) data. The sample consists of all person-year observations during which both DER and SIPP earnings are observed and are positive. Monthly SIPP earnings are summed the annual level and only counted as non-missing if all 12 months were non-missing. Additional details in Section 3.1.1.

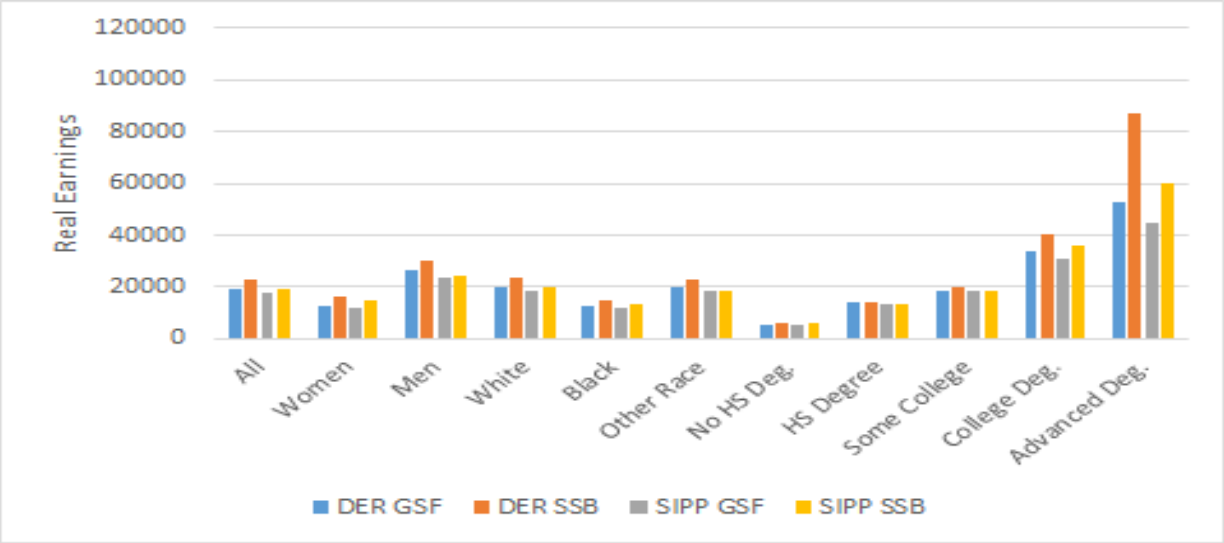
**Figure 3: Distribution of Earnings Differences Across Data Sets and Sources**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure shows a histogram of the difference between SIPP and DER earnings (SIPP minus DER) for the original (GSF) and synthetic (SSB) data. The differences are binned into bin sizes of $1,000, except the "0" bin which ranges from -$1,000 to $1,000. The tail bins, "-10" and "10", correspond to -$10,000+ and $10,000+. The sample consists of all person-year observations during which both DER and SIPP earnings are observed and are positive. Monthly SIPP earnings are summed the annual level and only counted as non-missing if all 12 months were non-missing. Additional details in Section 3.1.1.

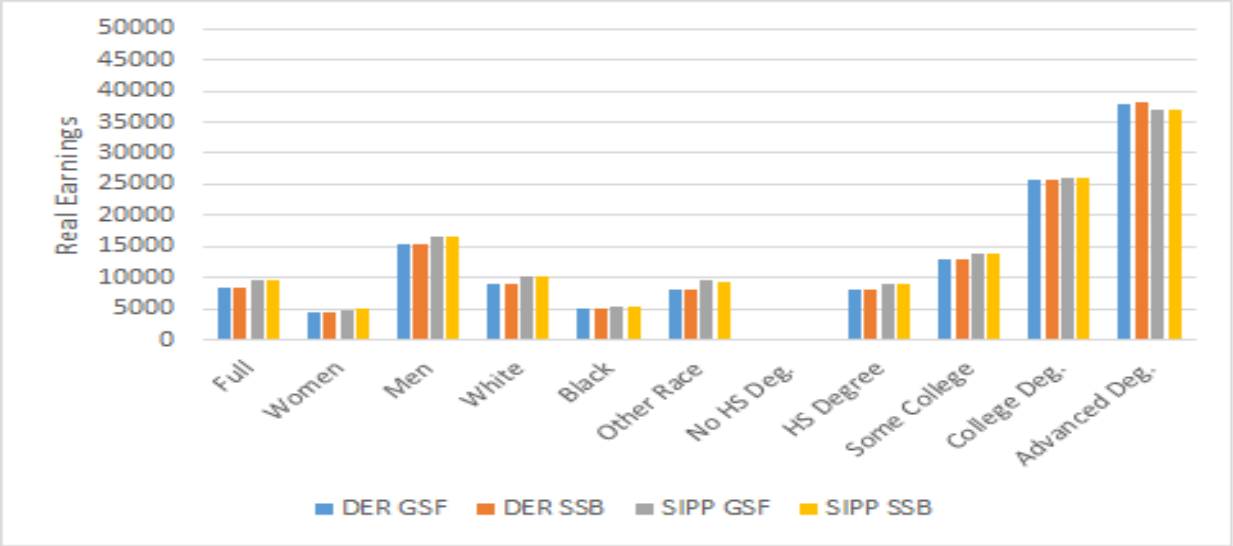
**Figure 4: Mean Earnings by Data Source and Data Set – Full Sample**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure shows mean real earnings for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earnings data sources (SIPP or DER). The "full" sample consists of individuals in the respective data set who have non-missing earnings values in both the DER and SIPP. Additional details in Section 3.1.1.

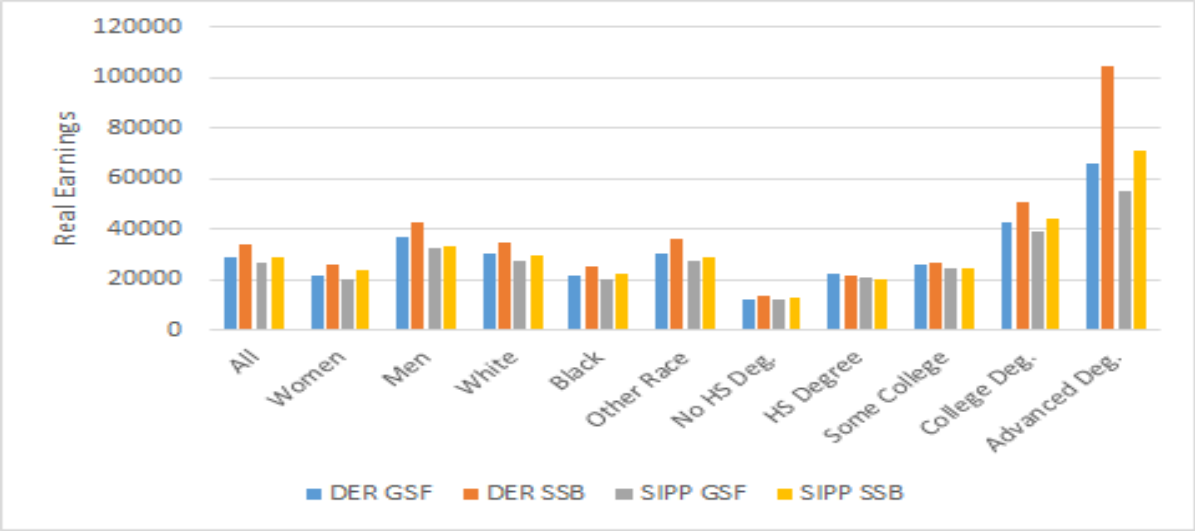
**Figure 5: Median Earnings by Data Source and Data Set – Full Sample**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure shows median real earnings for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earnings data sources (SIPP or DER). The "full" sample consists of individuals in the respective data set who have non-missing earnings values in both the DER and SIPP. Additional details in Section 3.1.1.

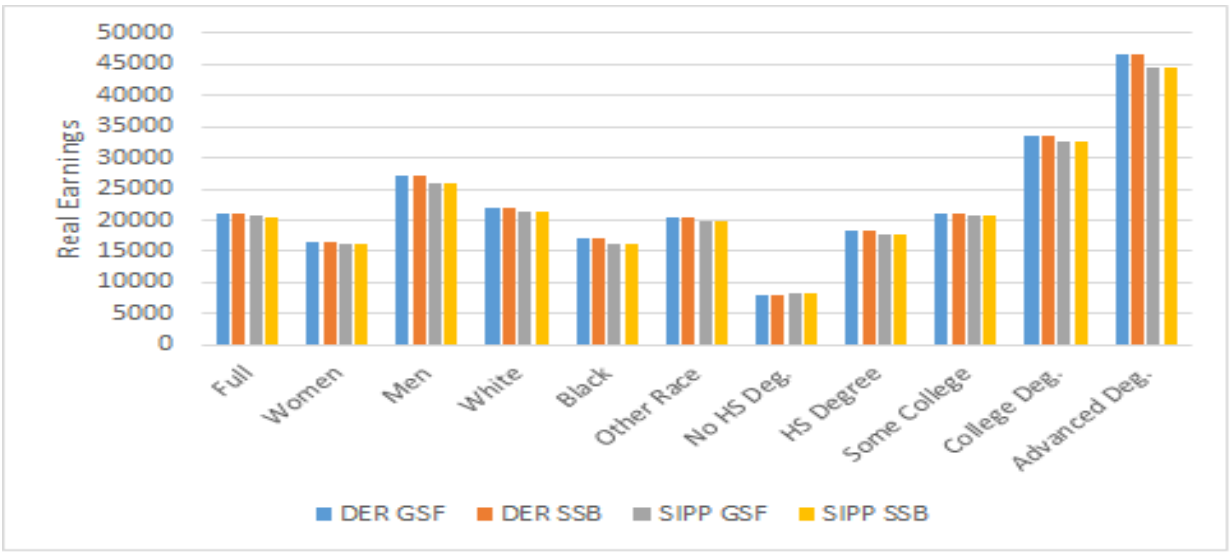
**Figure 6: Mean Earnings by Data Source and Data Set– Positive Earners**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure shows mean real earnings for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earnings data sources (SIPP or DER). The "full" sample consists of individuals in the respective data set who have positive (i.e., greater than zero) earnings values in both the DER and SIPP. Additional details in Section 3.1.1.

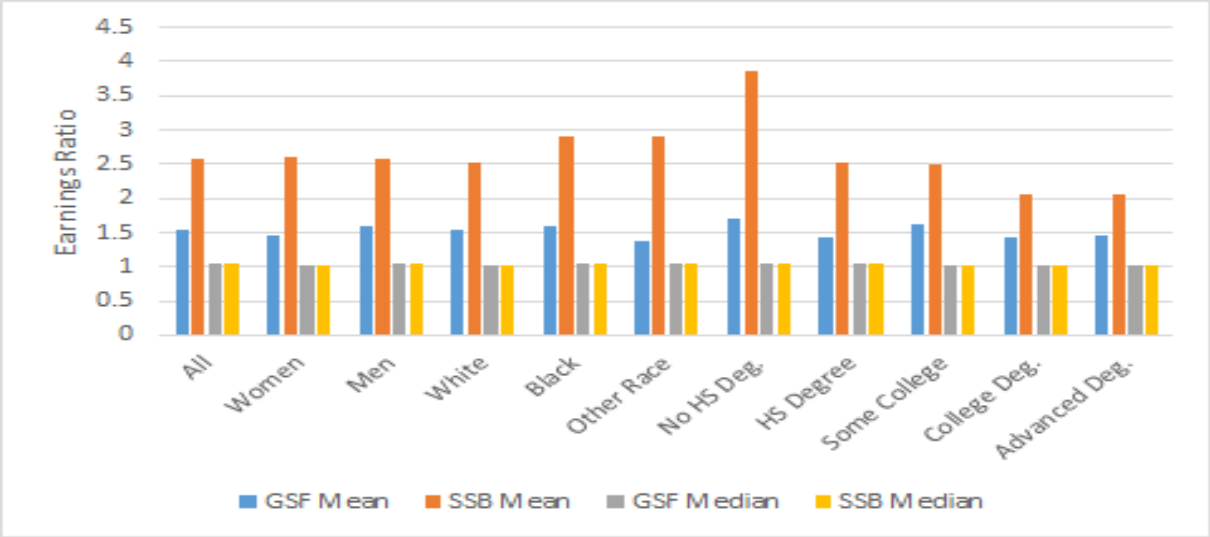
**Figure 7: Median Earnings by Data Source and Data Set– Positive Earners**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure shows median real earnings for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earnings data sources (SIPP or DER). The "full" sample consists of individuals in the respective data set who have positive (i.e., greater than zero) earnings values in both the DER and SIPP. Additional details in Section 3.1.1.

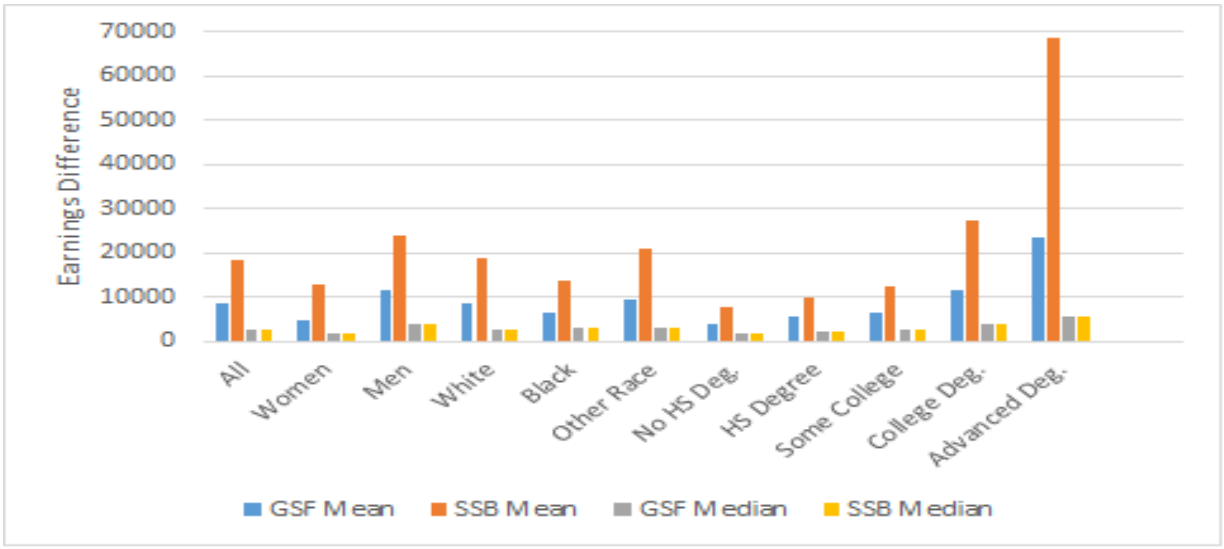
**Figure 8: Earnings Ratios Between Administrative Records and Survey Responses**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure shows average earnings ratios (i.e., quotients) between DER earnings values and SIPP earnings values for assorted demographic groups. The bars represent different data sets (SSB or GSF) and statistics (mean or median). The "full" sample consists of individuals in the respective data set who have positive (i.e., greater than zero) earnings values in both the DER and SIPP. Additional details in Section 3.1.1.

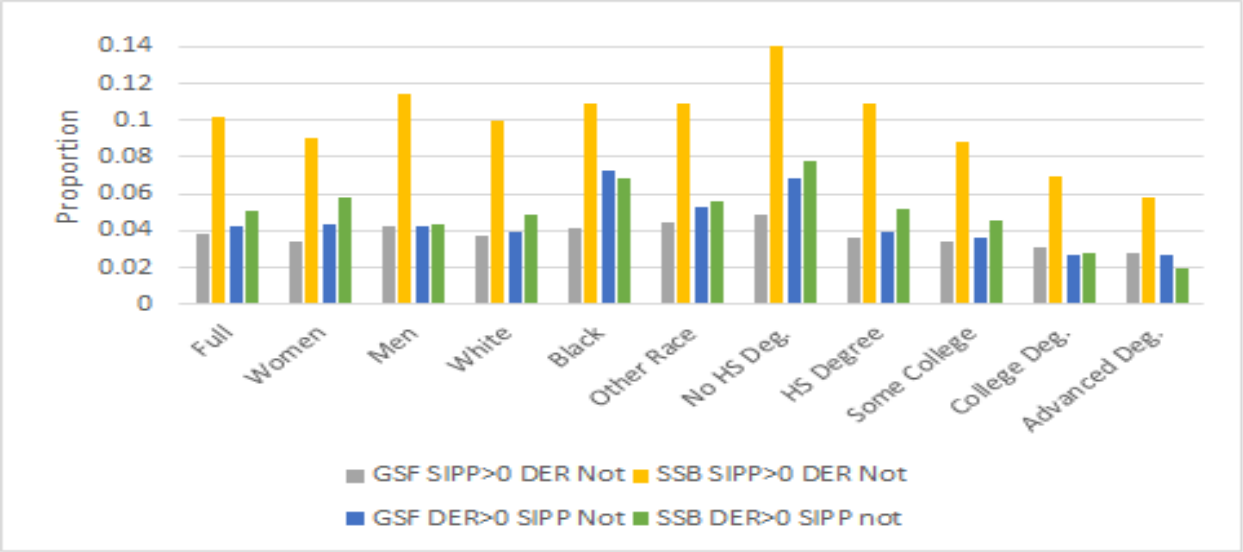
**Figure 9: Earnings Differences Between Administrative Records and Survey Responses**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure shows average or median absolute earnings differences between DER earnings values and SIPP earnings values for assorted demographic groups. The bars represent different data sets (SSB or GSF) and statistics (mean or median). The "full" sample consists of individuals in the respective data set who have positive (i.e., greater than zero) earnings values in both the DER and SIPP. Additional details in Section 3.1.1.

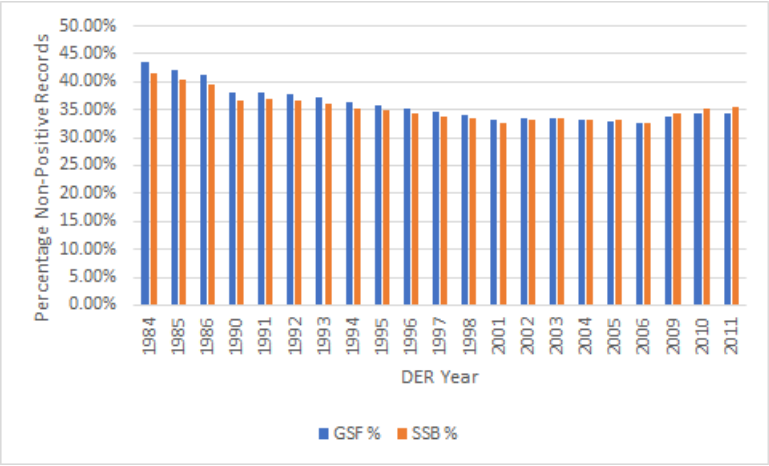
**Figure 10: Comparison of Positive Earnings Values Across Data Sets and Sources**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195

Notes: The figure shows comparisons of proportions of records with positive or non-positive and non-missing earnings values for assorted demographic groups. The bars represent different data sets (SSB or GSF) and earning value combinations. The "full" sample here consists of individuals in the respective data set who have non-missing earnings values in both the DER and SIPP. Additional details in Section 3.1.2.

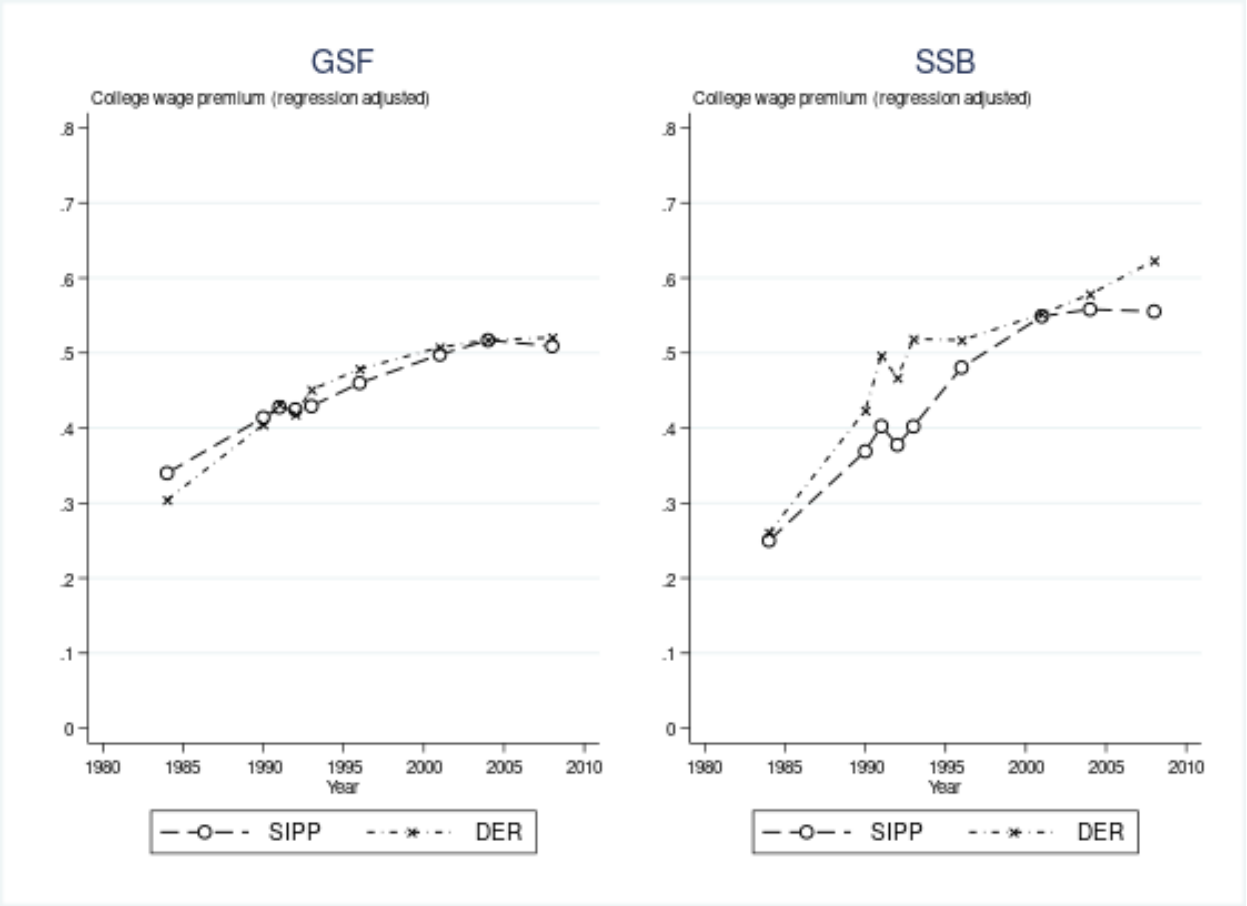
**Figure 11: Comparison of Records with Non-Positive Earnings Across Data Sources**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-285.

Notes: The figure shows the proportion of records with non-positive and non-missing DER earnings in assorted years. The "full" sample here consists of all individuals in the respective data set. Additional details in Section 3.1.2.

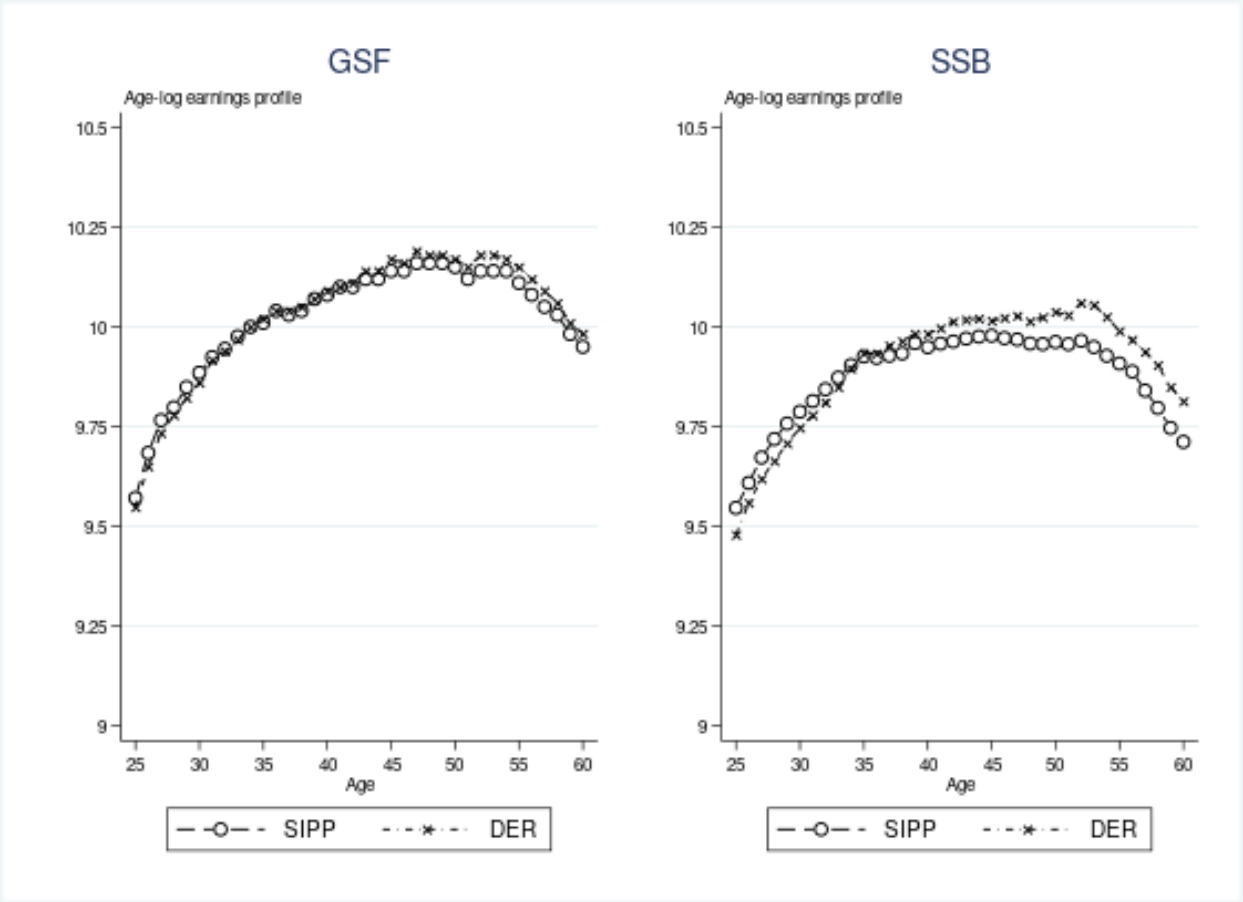
**Figure 12: College Wage Premium Estimates Comparing Data Sets**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure plots the regression-adjusted college wage premium across SIPP panels. Four different versions of the regression are estimated based on the different earnings measure (SIPP/DER) and data source (GSF/SSB). The regression adjusts for highest education level, sex, race, age-quartic, Hispanic status, and limits to ages 25-54. The figure plots the 'college' coefficient (reference group = 'high school') by SIPP panel. The sample is the "positive earners sample" described in Section 3. Additional details in Section 3.2.3.

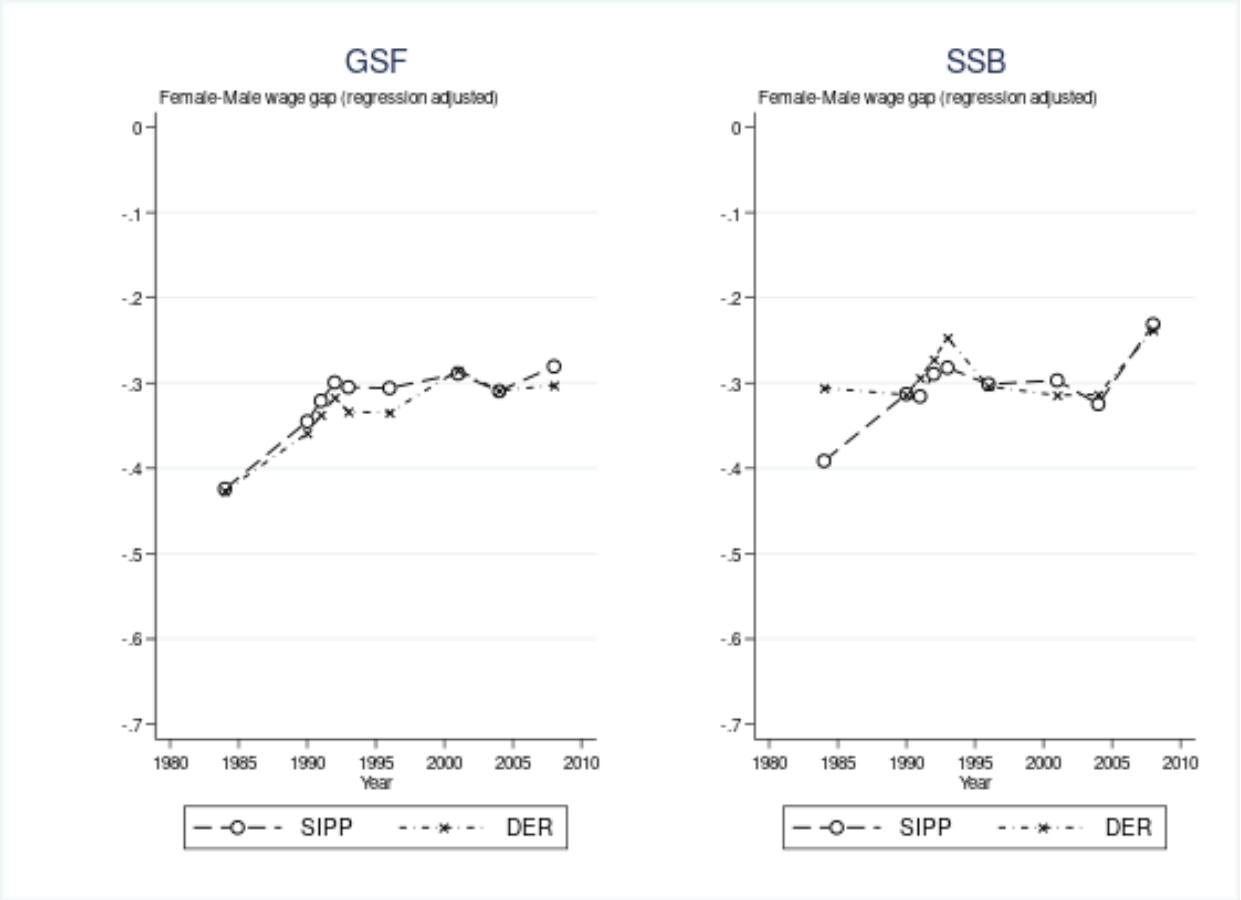
**Figure 13: Lifecycle Earnings Trends Comparing Data Sets**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure plots the lifecycle-earnings profile. Four different versions of the regression are estimated based on the different earnings measure (SIPP/DER) and data source (GSF/SSB). The figure plots the coefficient for age indicators without any covariates. All SIPP panels are pooled together. The sample is the "positive earners sample" described in Section 3. Additional details in Section 3.2.3.

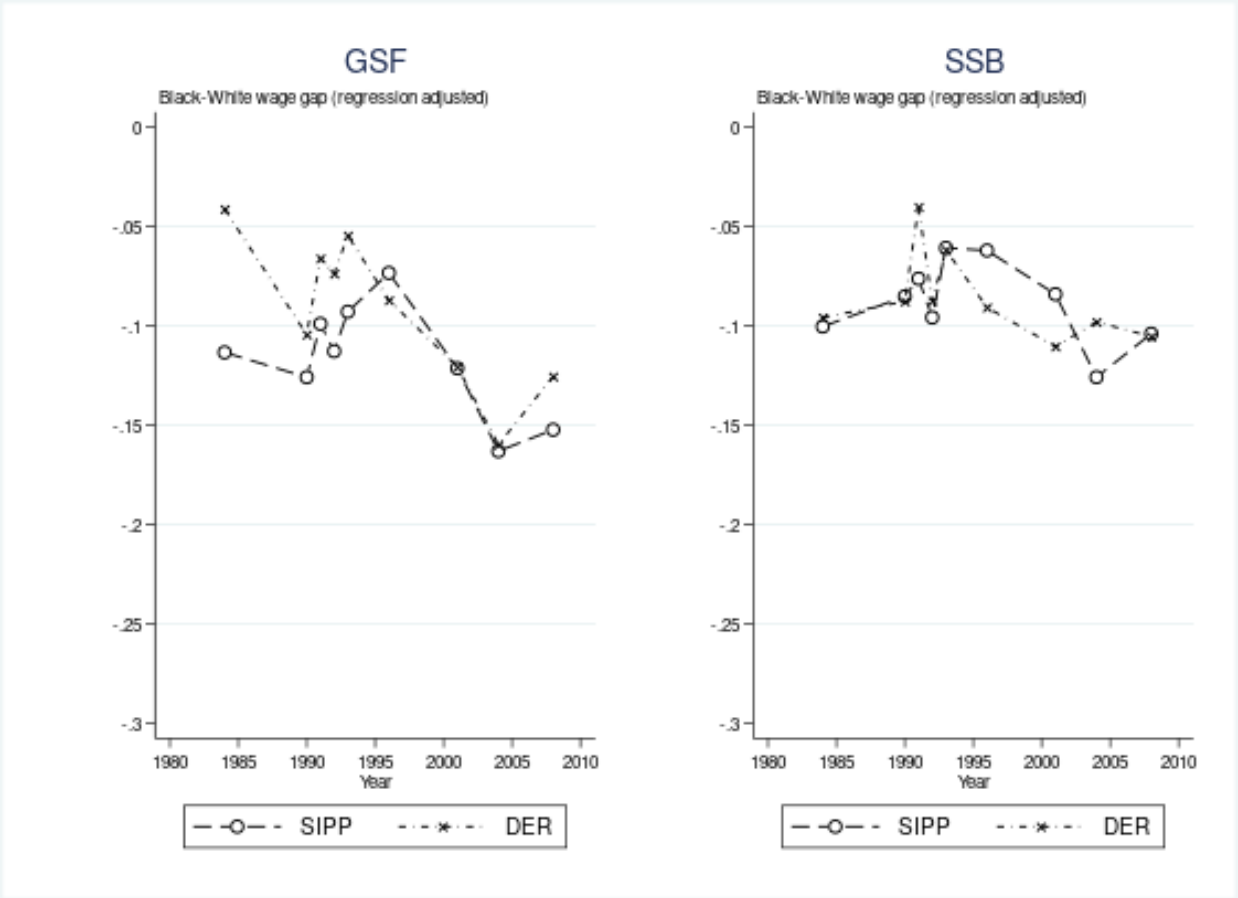
**Figure 14: Gender Wage Gap Trends Comparing Data Sets**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure plots the regression-adjusted gender wage gap across SIPP panels. Four different versions of the regression are estimated based on the different earnings measure (SIPP/DER) and data source (GSF/SSB). The regression adjusts for highest education level, sex, race, age-quadratic, Hispanic status, and limits to ages 25-54. The figure plots the 'Female' coefficient by SIPP panel. The sample is the "positive earners sample" described in Section 3. Additional details in Section 3.2.3.

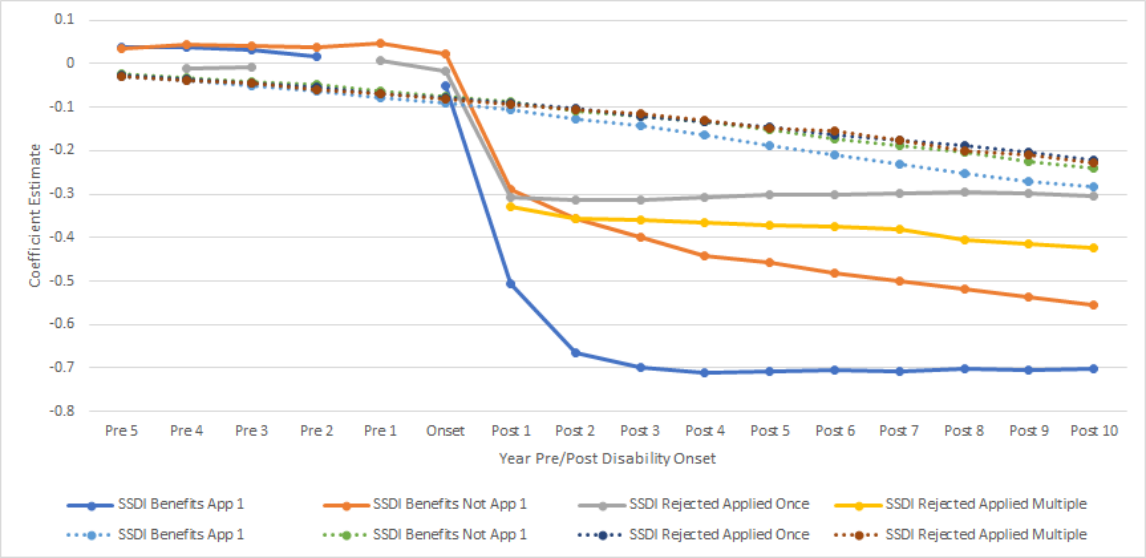
**Figure 15: Black-White Wage Gap Trends Comparing Data Sets**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure plots the regression-adjusted Black-White wage gap across SIPP panels. Four different versions of the regression are estimated based on the different earnings measure (SIPP/DER) and data source (GSF/SSB). The regression adjusts for highest education level, sex, race, age-quadratic, Hispanic status, state, and limits to ages 25-54. The figure plots the 'Black' coefficient by SIPP panel. The sample is the "positive earners sample" described in Section 3. Additional details in Section 3.2.3.

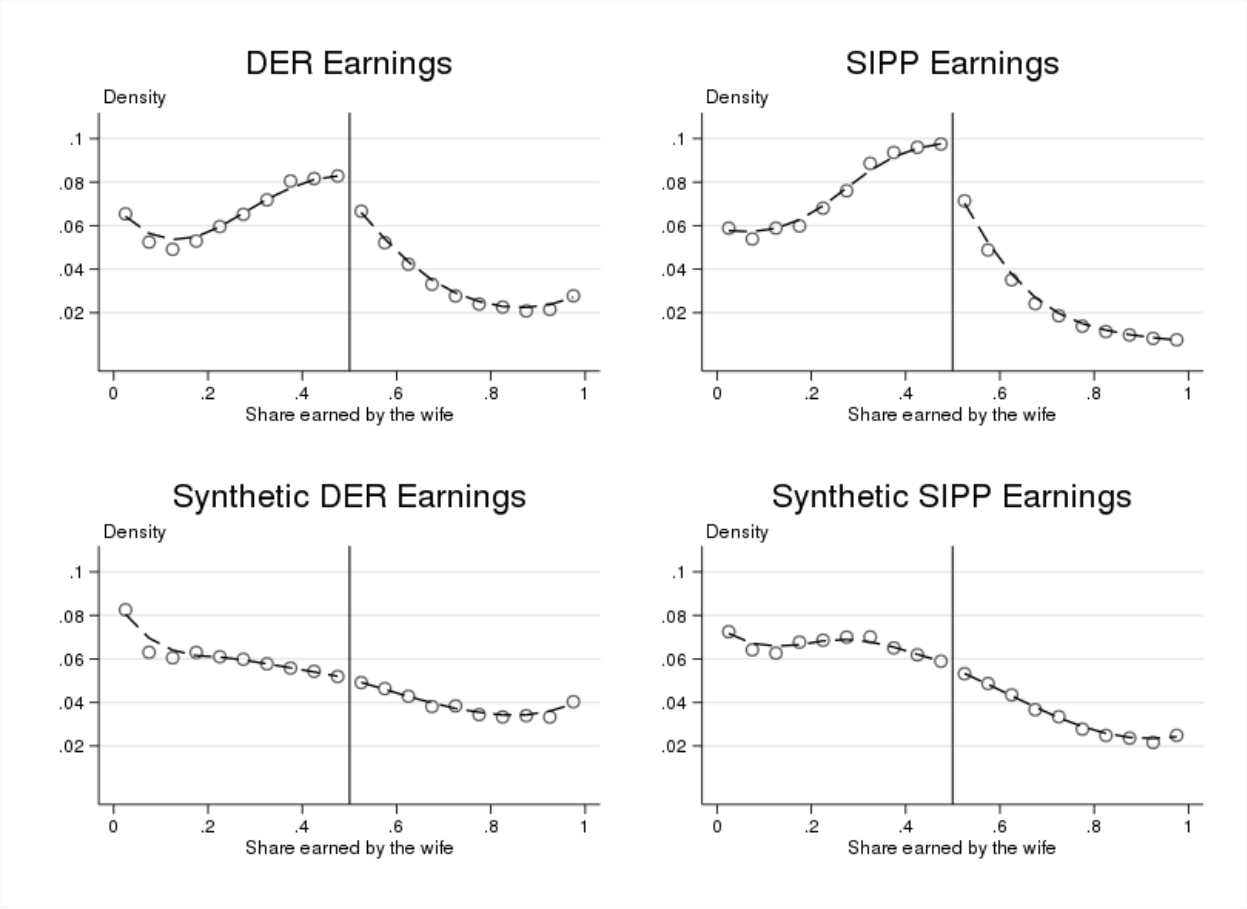
**Figure 16: Estimated Effects of SSDI Application Status on Likelihood of Positive Earnings**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number CBDRB-FY23-CED009-0001

Notes: The solid lines represent GSF estimates, and the dotted lines represent SSB estimates. N for SSB is 4,167,000 and N for GSF is 4,740,000. Analytical sample includes individuals aged 30 through 61 in the SIPP GSF who applied for SSDI benefits or never applied for SSDI benefits and had non-missing disability information and no indication of a work-limiting or work-preventing health condition. Further, the 1984 panel was dropped in the interest of having sufficient pre-SIPP DER observations. The dependent variable is a binary indicator for positive DER earnings. The independent variables of interest are interactions between categorical indicators for receiving SSDI application history and relative year indicators. The SSDI categories are "SSDI Benefits App 1" (received SSDI benefits on first application), "SSDI Benefits Not App 1" (received SSDI benefits but not on the first application), "SSDI Rejected Applied Once" (applied for SSDI once and did not receive benefits), and "SSDI Rejected Applied Multiple" (applied for SSDI multiple times but never received benefits). The baseline group are non-disabled individuals defined as persons in the SIPP GSF who never applied for SSDI and have non-missing work disability information in both the SIPP data and administrative records with no indication of a work-limiting or work-preventing health condition. The relative year dummies are based on the disability onset year indicated on the first SSDI application. Individual fixed effects and calendar year dummy variables were included in the model as were variables for age, age squared, and a binary time-variant indicator for married. Standard errors are clustered are the person level. All reported estimates are statistically significant at the 5% level; statistically insignificant estimates are depicted as blanks in the figure. See section 3.2.6 for additional details.

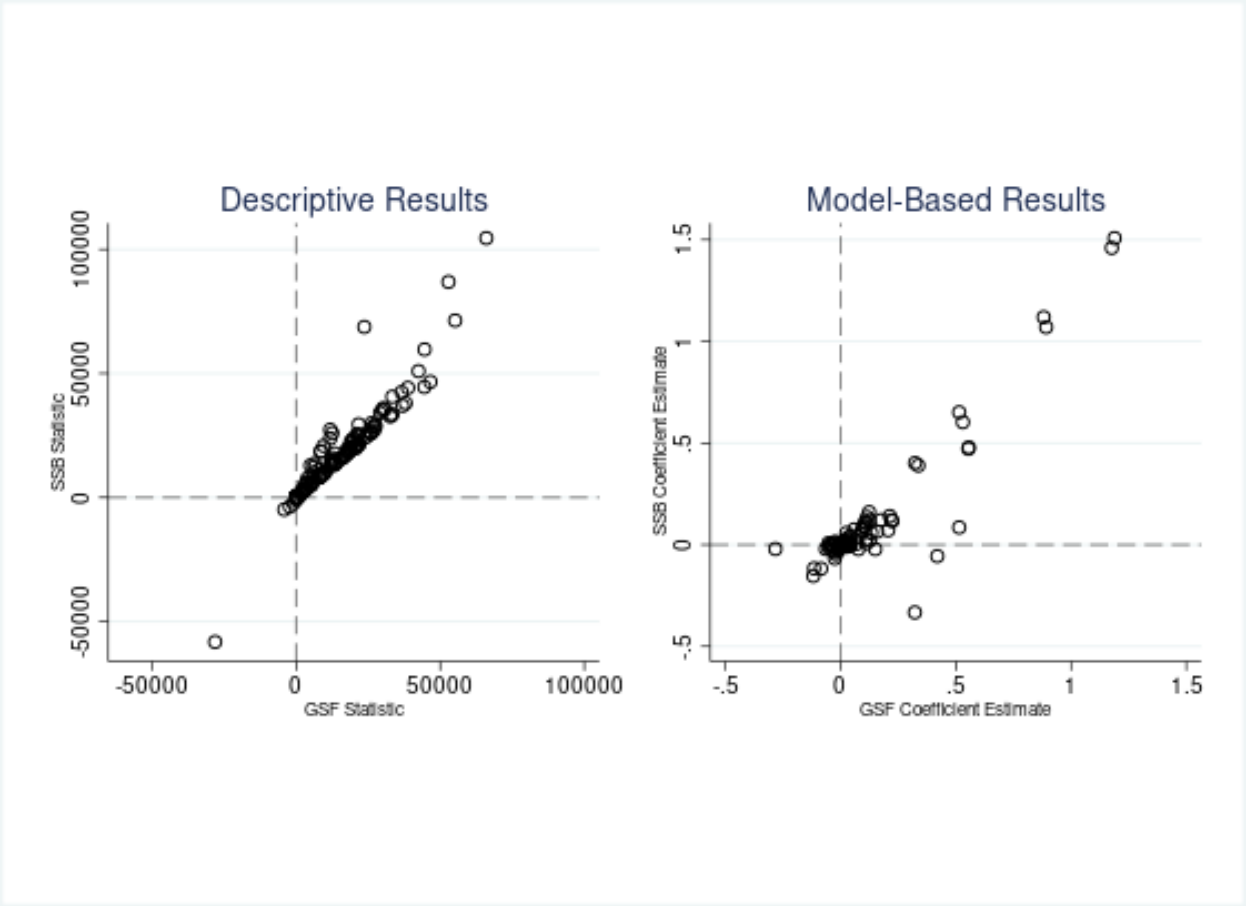
**Figure 17: Distributions of Wife's Earnings Share Across Data Sets and Sources**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The figure plots the share of household income that is earned by the wife. Four different versions of the regression are estimated based on the different earnings measure (SIPP/DER) and data source (GSF/SSB). The figure also plots a discontinuous density estimator with the discontinuity at 0.5. The sample includes all couples from the 2008 SIPP panel in which both spouses have positive earnings in 2009 for the given measure. Additional details in Section 3.2.8.

**Figure 18: Scatter Plot of GSF and SSB Results**




Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025, CBDRB-FY20-CED001-B0003, CBDRB-FY21-CED002-B0003, CBDRB-FY21-195, CBDRB-FY21-285, and CBDRB-FY23-CED009-0001.

Notes: The right figure plots the GSF versus SSB results for the regression-based results in the paper. The left figure plots the remaining statistics in the paper (e.g., means, medians, ratios, and counts). The earnings distributions (Figure 1) and relative earnings within couples (Figure 17) are excluded because we only released the figures themselves and not the underlying statistics. We also excluded the model-based results in the paper that do not also report a standard error or confidence interval (Figures 12-16) so that the model-based scatterplot is directly comparable to the model-based inference summary we performed in Table 14.

**Table 1: Moments for the Standard Deviation of Within-Person Earnings**

| | (1) Mean | (2) P10 | (3) P25 | (4) P50 | (5) P75 | (6) P90 |
|---|---|---|---|---|---|---|
| GSF | 12,540 | 1,760 | 4,160 | 7,820 | 13,450 | 21,690 |
| SSB | 25,930 | 2,590 | 6,510 | 11,110 | 17,780 | 29,480 |

Source; CBDRB-FY23-CED009-0001

Notes: This table shows moments for the distribution of the standard deviation of within-person earnings using the Detailed Earnings Record (DER). To be specific, standard deviations in real earnings were calculated for each person in the sample, and then one observation is kept per person. The samples consist of individuals aged 30 through 61 who had at least three earnings observations in the data set. The 1984 SIPP panel was excluded. There were 390,000 individuals in the SSB sample and 378,000 individuals in the GSF sample.

## Table 2: Predictors of Missing Earnings

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | GSF | | SSB | |
| | Missing SIPP | Missing DER | Missing SIPP | Missing DER |
| Age | -0.002409*** | -0.0005*** | -0.0036*** | -0.0001 |
| | (0.0001578) | (0.0001) | (0.0002) | (0.0001) |
| Age Squared | 0.00002458*** | 0.000001 | 0.00003*** | -0.000004*** |
| | (0.000002) | (0.0000001) | (0.000001) | (0.000001) |
| Male | 0.01018*** | -0.0037*** | 0.0101*** | -0.0023 |
| | (0.000865) | (0.0013) | (0.0011) | (0.0020) |
| Black | 0.0133*** | 0.0432*** | 0.0170*** | 0.0405*** |
| | (0.001597) | (0.001286) | (0.0019) | (0.0036) |
| Other Race | 0.03713*** | 0.01821*** | 0.0380*** | 0.0238*** |
| | (0.002405) | (0.003812) | (0.0030) | (0.0055) |
| Less than HS | -0.03625*** | -0.007418*** | -0.0282*** | -0.0141** |
| | (0.001327) | (0.002012) | (0.0020) | (0.0045) |
| Hispanic | 0.008784*** | 0.02552*** | 0.0143*** | 0.0586*** |
| | (0.001907) | (0.003079) | (0.0023) | (0.0046) |
| Foreign Born | -0.006737*** | 0.06034*** | -0.0119*** | 0.0755*** |
| | (0.001831) | (0.003161) | (0.0028) | (0.0070) |
| Some College | 0.0007907 | -0.0202*** | -0.0100*** | -0.0354*** |
| | (0.001139) | (0.001653) | (0.0017) | (0.0027) |
| Bachelor's | -0.006153*** | -0.01988*** | -0.0230*** | -0.0475*** |
| | (0.001425) | (0.001997) | (0.0026) | (0.0049) |
| Graduate | -0.0100*** | -0.02444*** | -0.0284*** | -0.0675*** |
| | (0.001819) | (0.002461) | (0.0024) | (0.0032) |
| Married | -0.05152*** | -0.01184*** | 0.0081*** | -0.0096*** |
| | (0.001072) | (0.001343) | (0.0012) | (0.0014) |
| Any Children | -0.01363*** | -0.03092*** | -0.0193*** | -0.0203*** |
| | (0.001142) | (0.00143) | (0.0015) | (0.0026) |
| Midwest | 0.01284*** | -0.007935*** | 0.0114*** | 0.0025 |
| | (0.001428) | (0.001923) | (0.0018) | (0.0023) |
| South | 0.01144*** | 0.001291 | 0.0122*** | 0.0019 |
| | (0.001356) | (0.00188) | (0.0017) | (0.0027) |
| West | 0.01463*** | 0.003777* | 0.0059*** | -0.0002 |
| | (0.001447) | (0.002157) | (0.0020) | (0.0026) |
| Observations | 12,140,000 | 8,655,000 | 9,514,000 | 6,733,000 |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-285.

Notes: The outcome variable is equal to 1 if earnings are missing and 0 otherwise. The independent variables are listed in the table. The sample for the SIPP columns is all person-month observations for individuals age 15 and older. The sample for the DER columns is all person-year observations for individuals age 15 and older. Standard errors, shown in parentheses, are clustered at the person level. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.1.

## Table 3: Mincer Model Regression Results

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | GSF | | SSB | |
| | Log SIPP | Log DER | Log SIPP | Log DER |
| Age | 0.095*** | 0.1025*** | 0.0737*** | 0.0881*** |
| | (0.0015) | (0.0017) | (0.0044) | (0.0083) |
| Age Squared | -0.0009*** | -0.0011*** | -0.0008*** | -0.001*** |
| | (0.00009) | (0.00002) | (0.00005) | (0.00002) |
| High School Degree | 0.3379*** | 0.3237*** | 0.3895*** | 0.4015*** |
| | (0.0071) | (0.008) | (0.0107) | (0.0134) |
| Some College | 0.5309*** | 0.515*** | 0.6028*** | 0.6514*** |
| | (0.0072) | (0.0081) | (0.0133) | (0.0201) |
| College Degree | 0.8922*** | 0.8816*** | 1.070*** | 1.119*** |
| | (0.0077) | (0.0087) | (0.0228) | (0.0247) |
| Advanced Degree | 1.177*** | 1.189*** | 1.460*** | 1.508*** |
| | (0.0087) | (0.00099) | (0.0125) | (0.0238) |
| Male | 0.5566*** | 0.5549*** | 0.4786*** | 0.472*** |
| | (0.0037) | (0.0042) | (0.0269) | (0.0408) |
| Non-White | -0.1176*** | -0.084*** | -0.1537*** | -0.1176*** |
| | (0.0045) | (0.0051) | (0.0075) | (0.0045) |
| Observations | 382,000 | 382,000 | 372,000 | 372,000 |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195.

Notes: The table shows coefficient estimates and standard errors from a Mincer-style regression analysis. The samples consist of individuals with positive earnings who are under age 65 but at least 25 years old. The model covariates are age, age squared, and binary/categorical indicators for sex, white non-Hispanic, education, and year. Standard errors are clustered by individual. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.2.

## Table 4: Returns to Schooling Regression Results

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | GSF | | SSB | |
| | OLS | 2SLS | OLS | 2SLS |
| | Panel A: SIPP Earnings | | | |
| Years of School | 0.1156*** | 0.2208*** | 0.1360*** | 0.1211 |
| | (0.0008) | (0.0751) | (0.0012) | (0.5458) |
| Observations | 126,000 | 126,000 | 118,000 | 118,000 |
| | Panel B: DER Earnings | | | |
| Years of School | 0.1246*** | 0.2251** | 0.1600*** | 0.1162 |
| | (0.0010) | (0.0916) | (0.0017) | (0.6691) |
| Observations | 126,000 | 126,000 | 118,000 | 118,000 |
| | Panel C: 2SLS First Stage | | | |
| Birth Quarter = 2 | | 0.0697*** | | 0.0046 |
| | | (0.0178) | | (0.0193) |
| Birth Quarter = 3 | | 0.0217 | | 0.0013 |
| | | (0.0173) | | (0.0183) |
| Birth Quarter = 4 | | 0.0413** | | -0.0047 |
| | | (0.0175) | | (0.0194) |
| Age | | 0.1263*** | | 0.1206*** |
| | | (0.0075) | | (0.0164) |
| Age Squared | | -0.0015*** | | -0.0014*** |
| | | (0.0001) | | (0.0002) |
| Observations | | 126,000 | | 118,000 |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY21-195, CBDRB-FY23-CED009-0001.

Notes: The table reports regression results for log earnings on years of schooling and other covariates. Two-stage least squares (2SLS) estimates are based on quarter of birth as an instrumental variable (Angrist and Krueger, 1991). The covariates include age, age-squared, state fixed effects, and year fixed effects. The sample is the "positive earners" sample described in the main text, further limited to non-Hispanic White males age 25-54 with at least 30 weeks worked in calendar year and individuals without missing covariates. Robust standard errors are reported in parentheses. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.4.

## Table 5: Kejriwal, Li, Totty (2020) Table 3 Replication

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Cross-Section | | | Panel | |
| | OLS | 2SLS | OLS | OLS | 2SLS |
| | | | Panel A: GSF | | |
| Years of School | 0.092*** | 0.134*** | 0.077*** | 0.105*** | 0.127*** |
| | (0.004) | (0.025) | (0.005) | (0.003) | (0.016) |
| Observations | 3,600 | 3,600 | 123,000 | 123,000 | 123,000 |
| | | | Panel B: SSB | | |
| Years of School | 0.070*** | 0.058 | 0.042*** | 0.093*** | 0.021 |
| | (0.006) | (0.037) | (0.005) | (0.004) | (0.026) |
| Observations | 3,700 | 3,700 | 125,000 | 125,000 | 125,000 |
| Age & Age-Squared | Yes | Yes | Yes | Yes | Yes |
| Person Fixed Effects | | | Yes | No | No |
| Year Fixed Effects | | | No | Yes | Yes |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval numbers: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025.

Notes: Panel A reproduces the results in Table 3 of Kejriwal et al. (2020). Panel B replicates their analysis on the SSB instead of the GSF. The columns show cross-section and panel data evidence on the returns to schooling. 2SLS results are based on quarter-of-birth interacted with year-of-birth as the instrumental variable. Standard errors, shown in parentheses, are robust for the cross-section results and clustered at the person level for the panel data results. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.4.

## Table 6: Kejriwal, Li, Totty (2020) Table 4 Replication

| | (1) IFE | (2) IFE | (3) CCEP | (4) CCEP | (5) CCEP-2 | (6) CCEP-2 |
|---|---|---|---|---|---|---|
| | | | Panel A: GSF | | | |
| Years of School | 0.020*** | 0.026** | 0.038*** | 0.037*** | 0.023*** | 0.024*** |
| | (0.003) | (0.003) | (0.004) | (0.006) | (0.004) | (0.004) |
| Observations | 123,000 | 123,000 | 123,000 | 123,000 | 123,000 | 123,000 |
| | | | Panel B: SSB | | | |
| Years of School | 0.001 | -0.0004 | -0.0003 | 0.024** | 0.003 | -0.002 |
| | (0.003) | (0.003) | (0.006) | (0.011) | (0.004) | (0.004) |
| Observations | 125,000 | 125,000 | 125,000 | 125,000 | 125,000 | 125,000 |
| Age & Age-Squared | Yes | Yes | Yes | Yes | Yes | Yes |
| Person Fixed Effects | Yes | No | Yes | No | Yes | No |
| Year Fixed Effects | No | Yes | No | Yes | No | Yes |
| Interactive Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval numbers: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025.

Notes: Panel A reproduces the results in Table 4 of Kejriwal et al. (2020). Panel B replicates their analysis on the SSB instead of the GSF. IFE corresponds to the Interactive Fixed Effects estimator from Bai (2009), CCE to the Common Correlated Effects estimator in Pesaran (2006), and CCEP-2 to a combination of the two described in Kejriwal et al. (2020). Standard errors, shown in parentheses, are clustered at the person level. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.4.

## Table 7: Kejriwal, Li, Totty (2020) Table 5 Replication

| | (1) OLSMG | (2) IFEMG | (3) CCEMG | (4) CCEMG-2 |
|---|---|---|---|---|
| | Panel A: GSF | | | |
| Years of School | 0.078*** | 0.028*** | 0.044*** | 0.041*** |
| | (0.006) | (0.003) | (0.006) | (0.006) |
| Observations | 123,000 | 123,000 | 123,000 | 123,000 |
| | Panel B: SSB | | | |
| Years of School | -0.022*** | -0.010 | -0.004 | 0.009 |
| | (0.006) | (0.006) | (0.008) | (0.007) |
| Observations | 125,000 | 125,000 | 125,000 | 125,000 |
| Age & Age-Squared | Yes | Yes | Yes | Yes |
| Person Fixed Effects | Yes | Yes | Yes | Yes |
| Interactive Fixed Effects | Yes | Yes | Yes | Yes |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval numbers: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025.

Notes: Panel A reproduces the results in Table 5 of Kejriwal et al. (2020). Panel B replicates their analysis on the SSB instead of the GSF. OLSMG, IFEMG, CCEMG, and CCEMG-2 correspond to mean group (MG) version of the OLS, IFE, CCE, and CCE-2 estimators. The MG estimators allow for heterogeneous coefficients by estimating person-level regressions. Standard errors, shown in parentheses, are clustered at the person level. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.4.

## Table 8: Vietnam War Draft Lottery and Civilian Earnings

| | (1) GSF | (2) SSB |
|---|---|---|
| Draft Eligible x Post | -0.1143*** | -0.1165*** |
| | (0.0110) | (0.0144) |
| Observations | 369,000 | 378,000 |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY23-CED009-0001.

Notes: The table reports difference-in-differences regression results for log earnings on Vietnam War draft lottery selection and other covariates. The outcome variable is individual-level annual earnings in the SSA's SER from 1960-1979. The sample is limited to white males born during 1944-1952. "Draft eligible" is an indicator for individuals who were randomly selected as "draft eligible" based on their birth date during the Vietnam War draft lotteries that occurred in 1970, 1971, and 1972. "Post" is an indicator for years after an individual's draft lottery year. The covariates are person fixed effects, birth year fixed effects, and year fixed effects. Robust standard errors are reported in parentheses. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.5.

## Table 9: Estimated Effect of SSDI Benefits Receipt on Likelihood of Positive Earnings

| | (1) GSF | (2) SSB |
|---|---|---|
| SSDI Benefits x Post | -0.2822*** | -0.02096*** |
| Disability Onset | (0.004223) | (0.004908) |
| Observations | 917,000 | 1,372,000 |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number CBDRB-FY23-CED009-0001.

Notes: N for SSB is 4,167,000 and N for GSF is 4,740,000. Analytical sample includes individuals aged 30 through 61 in the SIPP GSF who applied for SSDI benefits. Further, the 1984 panel was dropped in the interest of having sufficient pre-SIPP DER observations. The dependent variable is a binary indicator for positive DER earnings. The independent variable of interest is the interaction between a binary indicator for receiving SSDI benefits and a binary indicator for post disability onset where disability onset is based on the date of onset from the individual's first SSDI application. Individual fixed effects and calendar year dummy variables were included in the model as were variables for age, age squared, and a binary time-variant indicator for married. Robust standard errors are reported in parentheses. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. See section 3.2.6 for additional details.

## Table 10: Minimum Wages, Wages, and Employment for Teens

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | GSF | | SSB | |
| | Log Wage | Employed | Log Wage | Employed |
| | Panel A: SIPP Earnings | | | |
| Log Minimum Wage | 0.4201*** | -0.0406 | -0.0561 | -0.0110 |
| | (0.0853) | (0.0814) | (0.1721) | (0.0720) |
| | 31000 | 46500 | 32875 | 46000 |
| | Panel B: DER Earnings | | | |
| Log Minimum Wage | 0.3229* | -0.0462 | -0.3330 | 0.0013 |
| | (0.1815) | (0.0497) | (0.2903) | (0.0861) |
| | 29500 | 46500 | 28750 | 46000 |
| State Fixed Effects | Yes | Yes | Yes | Yes |
| Year Fixed Effects | Yes | Yes | Yes | Yes |
| State Linear Trends | Yes | Yes | Yes | Yes |
| Census Division x Year Effects | Yes | Yes | Yes | Yes |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval numbers: CBDRB-FY21-195.

Notes: The table reports regression results for employment on the log value of the minimum wage and other covariates. Employment is measured as an indicator for having positive earnings. The covariates include state-year unemployment rate, state-year population, sex, race, Hispanic status, highest education level, and age indicators. The sample is the "full sample" described in the text, further limited to teens ages 16-19 without any missing covariates. Standard errors, are shown in parentheses, are clustered at the state level. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.7.

## Table 11: Hampton and Totty (2021) Tables 2-3 Replication

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| | Employed (Has DER Earnings) | | | Part-Time Employed | | | Full-Time Employed | | |
| | | | | Panel A: GSF | | | | | |
| Log Minimum Wage | 0.214*** | 0.175*** | 0.151*** | 0.207** | 0.160** | 0.112* | 0.106 | 0.113* | 0.111* |
| | (0.065) | (0.065) | (0.043) | (0.085) | (0.076) | (0.066) | (0.090) | (0.068) | (0.064) |
| Observations | 27,000 | 27,000 | 27,000 | 27,000 | 27,000 | 27,000 | 27,000 | 27,000 | 27,000 |
| | | | | Panel B: SSB | | | | | |
| Log Minimum Wage | 0.141* | 0.121 | -0.022 | 0.071 | 0.065 | 0.007 | 0.111 | 0.108 | 0.028 |
| | (0.080) | (0.075) | (0.055) | (0.054) | (0.053) | (0.042) | (0.067) | (0.063) | (0.050) |
| Observations | 51,000 | 51,000 | 51,000 | 51,000 | 51,000 | 51,000 | 51,000 | 51,000 | 51,000 |
| State, Age, and Year Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Covariates | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes |
| Person Fixed Effects | No | No | Yes | No | No | Yes | No | No | Yes |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY20-CED001-B0003.

Notes: Panel A reproduces the results in Table 2 and Table 3 of Hampton and Totty (2021). Panel B replicates their analysis on the SSB instead of the GSF. The dependent variable in columns (1)-(3) is an indicator for positive DER earnings. The dependent variables in columns (4)-(9) are indicators for part-time or full-time employment based on the amount of the person's DER earnings relative to their lifetime highest earning year. See Hampton and Totty (2021) for more details. Standard errors, shown in parentheses, are clustered at the state level. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.7.

## Table 12: Hampton and Totty (2021) Table 4 Replication

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Partial Exit Hazard | | Full Exit Hazard | |
| | Panel A: GSF | | | |
| Log Minimum Wage | -0.0233 | -0.0247 | -0.0641** | -0.0512* |
| | (0.0247) | (0.0225) | (0.0274) | (0.0303) |
| Observations | 14,500 | 14,500 | 14,500 | 14,500 |
| | Panel B: SSB | | | |
| Log Minimum Wage | 0.014 | 0.011 | -0.021 | -0.016 |
| | (0.023) | (0.023) | (0.026) | (0.027) |
| Observations | 99,000 | 99,000 | 99,000 | 99,000 |
| State, Age, and Year Fixed Effects | Yes | Yes | Yes | Yes |
| Covariates | No | Yes | No | Yes |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY201-CED002-B0003.

Notes: Panel A reproduces the results in Table 4 of Hampton and Totty (2021). Panel B replicates their analysis on the SSB instead of the GSF. The dependent variable is an indicator that permanently changes from 0 to 1 when a person's earnings permanently fall below a person-specific threshold. See Hampton and Totty (2021) for more details. Standard errors, shown in parentheses, are clustered at the state level. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.7.

## Table 13: Hampton and Totty (2021) Tables 5 Replication

| | (1) | (2) |
|---|---|---|
| | Claimed Hazard | |
| | Panel A: GSF | |
| Log Minimum Wage | -0.0351** | -0.0380*** |
| | (0.0151) | (0.0139) |
| Month of FRA | | 0.515*** |
| | | (0.0240) |
| Observations | 68,500 | 68,500 |
| | Panel B: SSB | |
| Log Minimum Wage | 0.001 | 0.001 |
| | (0.007) | (0.007) |
| Month of FRA | | 0.0851*** |
| | | (0.009) |
| Observations | 206,000 | 206,000 |
| State, Age, and Year Fixed Effects | Yes | Yes |
| Covariates | No | Yes |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY20-CED001-B0003

Notes: Panel A reproduces the results in Table 4 of Hampton and Totty (2021). Panel B replicates their analysis on the SSB instead of the GSF. The dependent variable is an indicator that permanently changes from 0 to 1 when an individual first receives Social Security retirement benefits. See Hampton and Totty (2021) for more details. Standard errors, shown in parentheses, are clustered at the state level. Statistical significance is as follows: one-percent=***, five-percent=**, and ten-percent=*. Additional details in Section 3.2.7.

## Table 14: SSB versus GSF Inference Comparison

| | (1) | (2) |
|---|---|---|
| Panel A: Confidence Interval Comparison | | |
| GSF CI average length | | 0.069 |
| SSB CI average length | | 0.129 |
| Average proportion of GSF CI overlapped by SSB CI | | 0.331 |
| Proportion of GSF estimates that lie within SSB CI | | 0.351 |
| | | |
| Panel B: Statistical Conclusion Comparison | Count | Percent |
| (1) Same sign and significance | 56 | 59.57% |
| (2) Same sign, change significance | 18 | 19.15% |
| (3) Change sign, neither significant | 3 | 3.19% |
| (4) Change sign and significance | 15 | 15.96% |
| (5) Change sign, both significant | 2 | 2.13% |
| Total | 94 | 100% |
| | | |
| Same sign [(1) + (2)]: | | 78.72% |
| Same statistical conclusion [(1) + (3)]: | | 62.67% |
| Opposite statistical conclusion (5) | | 2.13% |

Source: U.S. Census Bureau Gold Standard File (GSF) and SIPP Synthetic Beta (SSB). U.S. Census Bureau Disclosure Review Board approval number: CBDRB-FY19-CED001-B0014, CBDRB-FY19-CED001-B0025, CBDRB-FY20-CED001-B0003, CBDRB-FY21-CED002-B0003, CBDRB-FY21-195, CBDRB-FY21-285, and CBDRB-FY23-CED009-0001.

Notes: The comparison includes all regression-based results in the paper except for those that do not report a standard error or confidence interval (Figures 12-16).